UnOS: Unified Unsupervised Optical-flow and Stereo-depth Estimation by Watching Videos

Yang Wang[1] Peng Wang[1] Zhenheng Yang[2] Chenxu Luo[3] Yi Yang[1] Wei Xu[1]
[1]Baidu Research [2] University of Southern California [3] Johns Hopkins University
{wangyang59, wangpeng54, yangyi05, wei.xu}@baidu.com zhenheny@usc.edu chenxuluo@jhu.edu

Abstract

In this paper, we propose UnOS, an unified system for unsupervised optical flow and stereo depth estimation using convolutional neural network (CNN) by taking advantages of their inherent geometrical consistency based on the rigid-scene assumption [31]. UnOS significantly outperforms other state-of-the-art (SOTA) unsupervised approaches that treated the two tasks independently. Specifically, given two consecutive stereo image pairs from a video, UnOS estimates per-pixel stereo depth images, camera ego-motion and optical flow with three parallel CNNs. Based on these quantities, UnOS computes rigid optical flow and compares it against the optical flow estimated from the FlowNet, yielding pixels satisfying the rigid-scene assumption. Then, we encourage geometrical consistency between the two estimated flows within rigid regions, from which we derive a rigid-aware direct visual odometry (RDVO) module. We also propose rigid and occlusion-aware flow-consistency losses for the learning of UnOS. We evaluated our results on the popular KITTI dataset over 4 related tasks, i.e. *stereo depth, optical flow, visual odometry and motion segmentation.*

1. Introduction

Estimating stereo depth [23] and optical flow [19] are two fundamental problems in computer vision. Jointly considering the two provides dense 3D scene flow [34], which enables numerous applications such as autonomous driving [34], robot navigation [7, 12] and video analysis [42, 25].

Current state-of-the-art (SOTA) strategies for both tasks rely on the advance of CNNs with supervised learning, *e.g.* PSMNet [5] and PWCNet [40], which depend heavily on the availability of training data[11, 32]. However, videos are from various scenes when considering open-world problems [58], so it is not practical to collect dense ground truths for these tasks at every place. Therefore, lots of efforts and progresses have been made recently in unsupervised learning of stereo depth/matching [57], monocular depth estimations [60] and optical flow [38] with CNNs by just providing stereo pairs or videos. These methods vastly improve the generalization ability of the learned models. Nevertheless, in those works, the two tasks were mostly treated independently in their pipeline, although it has shown to be very useful to consider both of the tasks as a whole in the aspect of 3D scene flow in traditional methods [43, 35, 41].

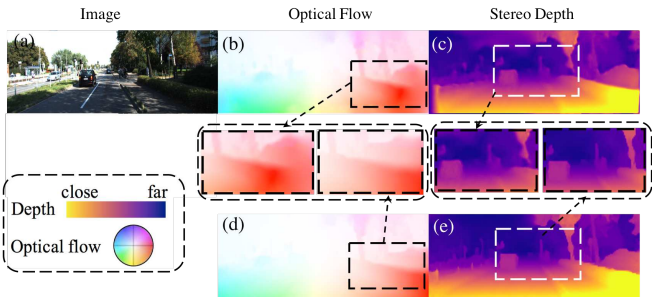

Figure 1. Comparison between UnOS and other unsupervised methods. (a) left image, (b) optical flow from [33], (d) UnOS optical flow, (c) stereo depth from [14], (e) UnOS stereo, It can be seen for both optical flow and stereo depth, UnOS generated results are more regularized and have sharper boundaries following scene structures as shown in zoomed regions (best view in color).

This paper completes this missing piece of unsupervised learning by proposing joint learning of the two tasks, which explores their geometrical relationship during training, and boosts the performance on both sides as illustrated in Fig. 1. We provide an overview of our system in Fig. 2. During training, given two consecutive stereo image pairs, *i.e.* (L_t, R_t) and (L_s, R_s), UnOS jointly outputs stereo depth estimation ($\mathbf{D}_t$), camera ego-motion ($\mathbf{T}_{t\to s}$) and optical flow ($\mathbf{F}_{t\to s}$) from StereoNet, MotionNet and FlowNet respectively. Then, a rigid-aware direct visual odometry (RDVO) module is applied after the MotionNet, which refines and updates the camera motion ($\mathbf{T}^u_{t\to s}$). Next, we use $\mathbf{D}_t$ and $\mathbf{T}^u_{t\to s}$ to compute rigid flow $\mathbf{F}^r_{t\to s}$ representing the motion induced solely by the camera, which is compared against $\mathbf{F}_{t\to s}$ and yields a rigid mask $\mathbf{M}$. In addition to the individual loss for each network, $\mathbf{F}^r_{t\to s}$ and $\mathbf{F}_{t\to s}$ are encouraged to be consistent within rigid regions $\mathbf{M}$, yielding more robust estimation for both tasks.

Our contributions are summarized as below,

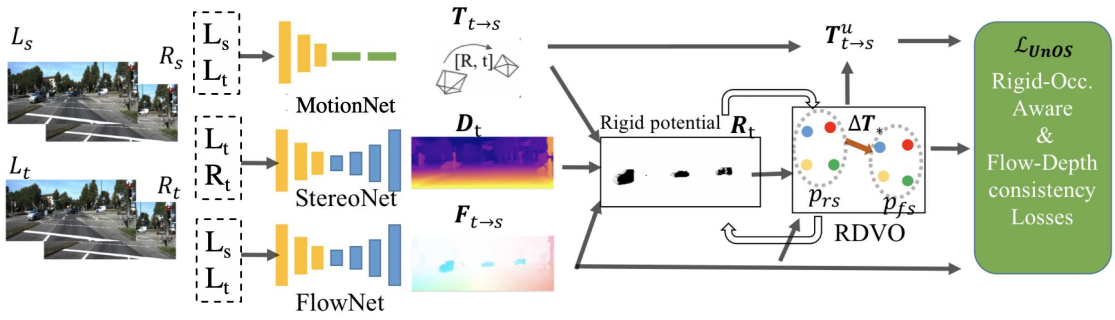

Figure 2. UnOS system. Given two consecutive stereo pairs, optical flow $\mathbf{F}_{t\to s}$, stereo depth $\mathbf{D}_t$, camera motion $\mathbf{T}_{t\to s}$ are predicted from three networks. Potentially rigid pixels are then discovered, and a rigid-aware direct visual odometry (RDVO) module is designed to refine the camera motion. All of the information is sent to our full set of losses $\mathcal{L}_{UnOS}$ with rigid awareness, occlusion (Occ.) awareness and flow-depth consistency (Details in Sec. 4.1). Please note that here the term FlowNet refers to the network for estimating optical flow which we used PWCNet [40] in our work.

1. We design an unified framework for unsupervised learning of optical flow and stereo depth, named as "UnOS", by explicitly encouraging their geometrical consistency with automatically found rigid regions, yielding SOTA performance on both tasks.
2. We design a rigid-aware direct visual odomotry (RDVO) module that carefully handles rigid regions using optical flow matching, yielding more accurate camera motion estimations.
3. We jointly include the properties of rigidity and occlusion in our training schema, which is effective for learning the CNNs.

UnOS significantly improves over other unsupervised stereo depth and optical flow methods. For example, on KITTI 2012 benchmark, UnOS reduces the optical flow error from a previous unsupervised method [33] by 50%. For stereo depth, it also outperforms the SOTA method taken stereo video as an input [57]. UnOS also achieves better performance for unsupervised moving object segmentation when comparing against [52]. The code and models of our method can be found at `https://github.com/baidu-research/UnDepthflow`.

2. Related Work

Stereo matching and optical flow estimation have long been important problems for computer vision. Here we summarize the closely related works using deep CNNs due to space limitation. We refer readers to survey papers on both tasks [9, 16] for broader understanding.

Supervised optical flow and stereo depth. In general, the two tasks share the same methodology for finding dense pixel correspondences, where stereo depth is a more constrained problem. Therefore, here we review them as a whole since an optical flow method can be easily extended to stereo matching by limiting the search within a disparity line. Based on CNNs, early works [55, 8, 28, 39, 15] started to learn matching for stereo using various losses with image patches as input, which might be time consuming during both training and inference. Recently, works like SPyNet [36] was designed to find 2D optical flow by explicitly using image warping in the architecture to enable efficient learning. PWCNet [40] built a 3D cost volume calculated within a local region. Despite the limited range in matching, PWCNet achieved SOTA optical flow results thanks to the coarse-to-fine scheme used.

To fully exploit the limited dimension and matching range in stereo depth, researchers built more specific architectures and losses. GCNet [23] proposed to generate a 3D cost volume by densely comparing the feature at a pixel from the reference image to all possible matching pixels at the target image. The network finds the best matching through a $soft\text{-}argmin$ operation. PSMNet [5] adopted pyramid spatial pooling and hourglass networks for exploiting image context. Later work [6] appended a post processing module, yielding better recovered details. These network architectures provide strong foundation for developing unsupervised learning methods.

Unsupervised optical flow. To reduce the requirements for large amount of training data, unsupervised optical flow learning was introduced recently in [38] and [22], where the basic idea was to use the spatial transform network [20] to backprop the photometric matching error from comparing the original and warped target images. Later works [46, 33] improved their results by explicitly handling the occlusions. In our case, we introduce geometrical regularization by jointly considering stereo depth, which produces further improvements.

Unsupervised monocular and stereo depth. Unsupervised learning of depth was first introduced for monocular images based on the supervision via stereo image pairs. Specifically, recent works [50, 10] adopted a CNN to take a single image as input and predict its disparity, where the supervision came from the photometric comparison. It was later improved by using inherent geometrical regularizations [14, 29]. Zhou *et al.* [60] incorporated camera egomotion into the training pipeline using structure from mo-

tion (SfM) [48], which made depth learning possible from monocular videos. Later works improved the performance by regularizing scene structures [53], refining camera ego-motion [45, 30], and jointly using stereo and monocular video for learning [26, 13]. Here, our RDVO is motivated by Differentiable Direct Visual Odometry (DDVO) [45]. However, rather than using photometric distance in solving relative camera pose, RDVO relies on estimated optical flow within rigid regions for pixel matching.

Due to the success of monocular depth estimation, researchers extended the corresponding losses to the problem of stereo depth estimation [14, 59], where the corresponding network architectures were borrowed from those shown to be effective with supervised learning such as GCNet [23]. These methods showed significant performance boost over the traditional unsupervised stereo algorithms based on local patch-wise matching and smoothing [18, 4, 27]. In our case, we adopt more light-weight PWCNet for stereo depth estimation by limiting the matching space.

Most recently, leveraging video for unsupervised stereo depth estimation was also proposed by Zhong *et al.* [58], where a RNN was used to implicitly aggregate the information from previous frames. Therefore, a video sequence is necessary for testing. In our case, we explicitly account for the depth transformation between consecutive frames with camera motion and optical flow in training, and only need a stereo pair for testing.

Joint unsupervised learning of depth and flow. Understanding depth and flow jointly from a video is commonly known as 3D scene flow estimation [43, 44], where 2D optical flow is explained with 3D scene structures and camera geometry. Recently traditional methods for scene flow estimation using stereo videos rely on bottom-up superpixel piece-wise planar matching [34], or top-down recognition [2, 49]. Taniai *et al.* [41] accelerated these algorithms with per-pixel scene flow understanding by jointly enforcing consistency among stereo depth, camera motion and optical flow. However, there were no learning components introduced in their systems.

Within the scope of unsupervised deep learning, joint depth and optical flow learning was studied based on monocular videos. GeoNet [54] used a residual FlowNet to refine the rigid flow from depth and ego-motion to the full optical flow, but no explicit geometry consistency was considered and it did not explicitly distinguish between static and moving regions. EPC [52] discovered rigid regions and encourage consistency between depth and flow estimations, but it did not do joint learning. Recent work [37] pieced the optical flow and rigid flow together and did iterative learning for refinement. DF-net [61] also proposed a consistency loss between rigid flow and optical flow. However, neither of them [37, 61] showed much improvements on the flow task due to the intrinsic limitation of the monocular depth accuracy. As mentioned in Sec. 1, including stereo depth estimations in our system fundamentally facilitates the learning of the two tasks.

3. Learning with self-supervision

In order to make the paper self-contained, we first introduce the preliminaries for unsupervised stereo depth [14], monocular depth [60] and optical flow [38] estimation, which share similar underlying idea of supervision by synthesis.

Finding corresponding pixels. As introduced in Sec. 1, we take consecutive stereo image pairs, (L_t, R_t) and (L_s, R_s) as inputs, where L, R indicate the left and right image respectively, and t, s indicate the target and source image. The networks estimate a stereo depth map $\mathbf{D}_t$ using (L_t, R_t), a relative camera pose $\mathbf{T}_{t\to s} \in \mathbf{SE}(3)$, and an optical flow map $\mathbf{F}_{t\to s}$ using (L_t, L_s). For each pixel p_t in a target image L_t, we can find the corresponding source pixels by,

$$\begin{aligned} p_{rs} &= \pi(\mathbf{K}[\mathbf{T}_{t\to s}\phi(p_t|\mathbf{K}, \mathbf{D}_t)]), \\ p_{fs} &= p_t + \mathbf{F}_{t\to s}(p_t), \\ p^x_{ss} &= p^x_t - f \cdot B/\mathbf{D}_t(p_t) \end{aligned} \tag{1}$$

where p_{rs} represents the pixel found at L_s based on the rigid scene assumption and camera motion, p_{fs} represents the pixel found at L_s through optical flow, and p_{ss} represents the pixel found at R_t via stereo disparity (the superscript x specifies the horizontal component). Here, $\phi(p_t|\mathbf{K}, \mathbf{D}_t) = \mathbf{D}_t(p_t)\mathbf{K}^{-1}h(p_t)$ is a back-projection function mapping a 2D pixel to a 3D point. $h(p_t)$ is the homogeneous coordinate of p_t. $\pi([x, y, d]) = [x/d, y/d]^T$ returns 2D non-homogeneous coordinates. $\mathbf{K}$ is the camera intrinsic matrix, and f, B are the focal length and baseline of the stereo image pair.

Supervise with view synthesis. Given the corresponding pixel pairs p_t and p_{*s} ($*$ could be r, f or s in Eq. (1)), we may generate synthesized target images $\hat{L}_{*t}$ from various source images using a differentiable bilinear interpolation [20], and the system can be trained by minimizing photometric error. The corresponding loss function term is defined as,

$$\mathcal{L}_{*v}(\mathbf{O}) = \sum\nolimits_{p_t} \mathbf{V}_*(p_t, \mathbf{O})|L_t(p_t) - \hat{L}_{*t}(p_t, \mathbf{O})|. \tag{2}$$

where $\mathbf{V}_*(p_t, \mathbf{O})$ is a visibility mask, indicating whether p_t can find a valid matching pixel given certain information $\mathbf{O}$ and a source image. $\mathbf{O}$ could be depth $\mathbf{D}_t$ or optical flow $\mathbf{F}_{t\to s}$. Here, the visibility mask $\mathbf{V}_*$ is computed by forward warping of the reverse optical flow as proposed in [46]. Thus, adopting different matching pairs triggers different unsupervised learning pipelines, *e.g.* using p_{rs}, p_{ss} or p_{fs}

Figure 3. An example of our rigid potential. (a) Image. (b) Flow consistency map. (c) Visibility mask $\mathbf{V}_f$. (d) Rigid potential. (e) Ground truth rigid mask. We can see flow consistency falsely indicates rigid potential in occluded regions.

induces mono-depth [60], stereo-depth [14], and optical-flow [38] respectively. Using both p_{rs} and p_{ss} leads to deep visual odometry with stereo video [26].

Regularization with edge-aware smoothness. Pixel color matching alone is unstable and ambiguous. Therefore, an edge-aware smoothness term is often applied for each prediction. Specifically,

$$\mathcal{L}_s(\mathbf{O}, \mathbf{W}, o) = \sum_{p_t} \sum_{d \in x,y} \mathbf{W}(p_t) |\nabla_d^o \mathbf{O}(p_t)| e^{-\beta |\nabla_d^2 L_t(p_t)|} \quad (3)$$

where $\mathbf{O}$ represents the type of the input, $\mathbf{W}$ is a weight map, and o is the order of smoothness gradient. For example, $\mathcal{L}_s(\mathbf{D}_t, \mathbf{1}, 2)$ is a spatial smoothness term penalizing the L1 norm of second-order gradients of depth $\mathbf{D}_t$ along both x and y directions over all images, as proposed by [14].

4. Unifying optical flow and stereo depth

One possible approach for unifying the learning of depth and flow is to use matching pixels of p_{rs}, p_{ss} and p_{fs} together during training. However, it may not work well, as also mentioned in prior works [26, 30], since errors from one task may negatively impact the other. This is mainly because there are moving things from t to s, and pixels belonging to those regions fail the one rigid motion assumption handling only the ego-motion. Thus, the discovered p_{rs} will be different from pixels found by optical flow p_{fs}. This systematic error will affect the learning of the whole model. Therefore, one key to successfully unify learning of both tasks is to find pixels having high potential satisfying the rigid assumption.

Locating rigid regions with soft potential. Here, rather than using a hard binary rigid mask as in [37], we consider using a soft rigid region mask [60], where each pixel has a potential of satisfying the rigid assumption. This will be useful in our RDVO module and losses later. In particular, the rigid potential at a pixel p_t is computed as,

$$\mathbf{R}_t(p_t) = \max\{1 - \mathbf{V}_f(p_t), \exp\{-\gamma(|p_{fs} - p_{rs}|)\}\} \quad (4)$$

where γ is a hyper parameter. Here, We first check the consistency between p_{fs} and p_{rs}, and also consider the regions that are occluded $1 - \mathbf{V}_f(p_t)$ as rigid. For example, regions that become occluded at image boundaries or road occluded by moving cars should be considered as rigid. Fig. 3 visualizes an example of our soft rigid mask, where more complete rigid regions are discovered using the two criteria. One possible error here could be having mutual occlusions between moving objects, which will be considered in our future work.

From Eq. (1) by letting $p_{rs} = p_{fs}$, we can see optical flow $\mathbf{F}_{t\to s}$, depth $\mathbf{D}_t$ and camera motion $\mathbf{T}_{t\to s}$ turn out to be three conjugated quantities within rigid regions. Given $\mathbf{D}_t$ and $\mathbf{F}_{t\to s}$, we can apply the n-point algorithm [17] to solve for $\mathbf{T}_{t\to s}$ with a closed-form using SVD, based on which we later propose rigid-aware direct visual odometry (RDVO) for pose refinement. It refines the camera pose obtained from the MotionNet. Given the refined camera pose $\mathbf{T}^u_{t\to s}$, $\mathbf{D}_t(p_t)$ and $\mathbf{F}_{t\to s}$, we propose to include geometrical consistency in our loss design. These two components will be elaborated below.

4.1. Rigid-aware direct visual odometry (RDVO)

In this module, given estimated $\mathbf{D}_t$, $\mathbf{F}_{t\to s}$ and an initial estimation of camera pose $\mathbf{T}_{t\to s}$ from MotionNet, our target is to find a relative pose $\Delta\mathbf{T}_{t\to s}$ to refine the pose $\mathbf{T}_{t\to s}$. This is necessary because MotionNet itself lacks geometrical constraints, which was also mentioned in [45, 30]. Here we propose a simpler and more efficient solution using the discovered rigid potential.

Specifically, the target of RDVO based on the notation from Eq. (1) is,

$$\min_{\Delta\mathbf{T}_{t\to s}} \sum_{p_t \in \mathcal{S}} \|p_{rs} - p_{fs}\|^2 \quad (5)$$

By substituting corresponding items in Eq. (1) yields,

$$\begin{aligned} p_{fs} - p_{rs} &= p_{fs} - \pi(\mathbf{K}[\Delta\mathbf{T}_{t\to s}\mathbf{T}_{t\to s}\phi(p_t|\mathbf{K}, \mathbf{D}_t)]) \\ &\Leftrightarrow \phi(p_{fs}|\mathbf{K}, \mathbf{D}_s) - \Delta\mathbf{T}_{t\to s}\mathbf{T}_{t\to s}\phi(p_t|\mathbf{K}, \mathbf{D}_t) \\ &= \psi(p_{fs}|\mathbf{D}_s)\mathbf{K}^{-1}h(p_{fs}) - \Delta\mathbf{T}_{t\to s}\mathbf{T}_{t\to s}\mathbf{D}_t\mathbf{K}^{-1}h(p_t) \end{aligned} \quad (6)$$

which means we back project 2D pixels to 3D point cloud for optimization. Here, $\psi(p_{fs}|\mathbf{D}_s)$ is a bilinear interpolation operation returning depth value at float coordinate p_{fs} using the depth map $\mathbf{D}_s$ from source images. Note p_{fs} does not necessarily have discrete values, therefore an interpolation is needed. $\Leftrightarrow$ means 2D to 3D projection. Now, Eq. (5) is a standard L2 minimization problem which can be easily solved using SVD [3]. In practice, computing pose can be more accurate by selecting the most reliable matching for visual odometry [1] rather than using all pixels. Therefore, $\mathcal{S}$ in Eq. (5) is chosen with two criteria, (1) $\mathbf{V}_f(p_t) > 0.75$ since only pixels without occlusion are valid for matching. (2) the potential $\mathbf{R}_t(p_t)$ is within top 25%. Here we choose these parameters based on the corresponding validation set.

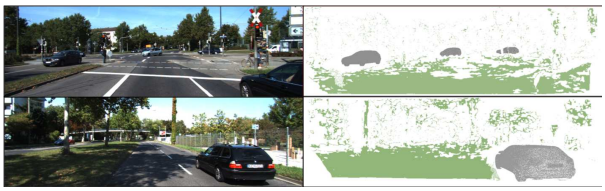

Figure 4. Left column: left target images (L_t). Right column: the regions selected in the RDVO module (i.e., region $\mathcal{S}$ described in Eq. (5)) in green overlaying on the ground truth moving object mask in grey.

Fig. 4 visualizes the selected pixels (green) in $\mathbf{S}$ for minimization. We see that the selected pixels (green) clearly separate from the moving objects (grey).

After RDVO, we obtain an updated camera motion $\mathbf{T}^u_{t\to s} = \Delta\mathbf{T}_{t\to s}\mathbf{T}_{t\to s}$, which we can feed back to calculate rigid matching of p_{rs}, yielding a better rigid potential ($Eq.$ (4)). We may iterate this process till convergence, and use the updated p^u_{rs} for generating various losses. In practice, we iterate twice for each sample and found it is already good enough in achieving SOTA results.

Finally, based on our rigid potential, we may generate a rigid segmentation mask with a threshold,

$$\mathbf{M}_t = \mathbf{R}_t(p_t) > 0.5, \tag{7}$$

which distinguishes the regions of static background and moving objects, and will be applied later for training the networks.

4.2. Learning with geometrical consistency

In this section, we discuss on how to leverage consistency in our losses and network architectures to effectively supervise UnOS.

4.2.1 Training losses

Rigid and occlusion-aware structural matching. As discussed in Sec. 3, photometric matching Eq. (2) follows Lambersian assumption based on pixel colors, which is not robust against illumination variations. To capture local structures, following [14], we add structural matching cost from SSIM [47]. Specifically, our pixel matching loss is,

$$\begin{aligned}\mathcal{L}_{*v}(\mathbf{O}) &= \sum\nolimits_{p_t} \mathbf{V}_*(p_t,\mathbf{O})\cdot s(L_t(p_t),\hat{L}_{*t}(p_t,\mathbf{O})),\\ \text{where, } s(L(p),\hat{L}(p)) &= (1-\alpha)\cdot|L(p)-\hat{L}(p)|+\\ &\alpha\cdot(1-\frac{1}{2}\text{SSIM}(L(p),\hat{L}(p))).\end{aligned} \tag{8}$$

Here, α is a balancing hyper-parameter. Same as in Eq. (2), $\mathbf{O}$ represents the type of output we need to supervise, which could be stereo depth estimations $\mathbf{D}_t$ or optical flow $\mathbf{F}_{t\to s}$. $\mathbf{V}_*$ indicates visibility mask depending on the type of source image for synthesis. Specifically, for supervising with stereo pairs, $\hat{L}_{*t}$ is from p_{ss}, $\mathbf{V}_s$ is computed using disparity. For supervising optical flow, $\hat{L}_{*t}$ is from p_{fs}, $\mathbf{V}_f$ is computed based on using backward optical flow, *i.e.* $\mathbf{F}_{s\to t}$ [46]. For supervising with consecutive images, *i.e.* $\hat{L}_{*t}$ is from p_{rs} (before RDVO), $\mathbf{V}_r$ represents rigid and non-occluded regions, which is computed as $\mathbf{V}_r = \mathbf{V}_f \odot \mathbf{M}_t$. We denote different view synthesis loss terms as $\mathcal{L}_{sv}, \mathcal{L}_{fv}, \mathcal{L}_{rv}$ respectively.

In addition, as mentioned in Sec. 4.1, we also obtain a better matching pixel p^u_{rs} after RDVO through an optimized camera motion $\mathbf{T}^u_{t\to s}$ yielding a new structural matching loss, which we denote as $\mathcal{L}^u_{rv}$ and $\mathbf{V}^u_r$ is computed accordingly.

Edge-aware local smoothness. We adopt similar smoothness loss functions as formulated in Eq. (3). Specifically, for depth, we follow [14] and use $\mathcal{L}_{ss} = \mathcal{L}_s(\mathbf{D}_t, \mathbf{1}, 2)$, which penalize the second-order gradient of depth. For optical flow, we choose to smooth over the moving regions, *i.e.* $\mathcal{L}_{fs}(\mathbf{F}_{t\to s}, 1-\mathbf{M}_t, 2)$.

Rigid-aware flow consistency Given updated camera motion $\mathbf{T}^u_{t\to s}$ after RDVO, we then further encourage consistency between rigid flow and optical flow. The consistency loss is formulated as,

$$\mathcal{L}_{fc}(\mathbf{F},\mathbf{D},\mathbf{T}^u) = \sum_{p_t}\mathbf{M}_t(p_t)|p^u_{rs}-p_{fs}| \tag{9}$$

where $\mathbf{M}_t(p_t)$ is the rigid mask computed in Eq. (7). Since our RDVO is not differentiable, this consistency loss only supervises FlowNet and StereoNet.

Left-right consistency Given stereo pairs, Godard *et al.* [14] showed that jointly predicting depth for both left and right images, and checking their consistency helps depth learning. We also include such loss for our StereoNet, which is denoted as $\mathcal{L}_{sc}$.

In summary, our loss functional for UnOS is written as,

$$\begin{aligned}\mathcal{L}_{UnOS} =& (\mathcal{L}_{fv}+\lambda_{fs}\mathcal{L}_{fs})+\lambda_{rv}(\mathcal{L}_{rv}+\mathcal{L}^u_{rv})\\ &+(\lambda_{sv}\mathcal{L}_{sv}+\lambda_{ss}\mathcal{L}_{ss}+\lambda_{sc}\mathcal{L}_{sc})+\lambda_{fc}\mathcal{L}_{fc}\end{aligned} \tag{10}$$

$\lambda = [\lambda_{fs}, \lambda_{rv}, \lambda_{sv}, \lambda_{ss}, \lambda_{sc}, \lambda_{fc}]$ is the set of hyper-parameters balancing different losses.

4.2.2 Network Architectures.

As reviewed in Sec. 2, SOTA stereo depth and optical flow algorithms are able to share similar architecture and methodology. In our work, due to joint multi-task training, we prefer more light-weighted architectures in order to fit everything into a single GPU. Therefore, for handling stereo matching, rather than using a stronger but relative heavy network, *e.g.* GCNet [23] or PSMNet [5], we choose PWCNet[40] used in the optical flow estimation, which is

Method	Train Stereo	Test Stereo	Super-vised	KITTI 2012 train Noc	KITTI 2012 train Occ	KITTI 2012 train All	KITTI 2012 test All	KITTI 2015 train move	KITTI 2015 train static	KITTI 2015 train all	KITTI 2015 test all
Flownet2			✓	–	–	**4.09**	–	–	–	**10.06**	–
Flownet2+ft			✓	–	–	(1.28)	1.8	–	–	(2.3)	11.48%
PWC-Net			✓	–	–	4.14	–	–	–	10.35	–
PWC-Net+ft			✓	–	–	(1.45)	**1.7**	–	–	(2.16)	**9.60%**
UnFlow-CSS [33]				1.26	–	3.29	–	–	–	8.10	–
Geonet [54]				–	–	–	–	–	–	10.81	–
Ranjan et al. [37]				–	–	–	–	6.35	6.16	7.76	–
Wang et al. [46]				–	–	3.55	4.2	–	–	8.88	31.2%
Janai et al. [21]				–	–	–	–	–	–	6.59	22.94%
DF-net [61]				–	–	3.54	4.4	–	–	8.98	25.70%
UnOS (FlowNet-only)				1.15	11.2	2.68	3.2	5.92	7.68	7.88	23.75%
UnOS (Ego-motion)	✓	✓		2.27	6.67	2.86	3.1	35.9	4.53	11.9	43.86%
UnOS (Ego+RDVO)	✓	✓		1.46	**4.88**	1.93	2.1	36.5	**2.99**	10.69	32.34%
UnOS (Full)	✓			**1.04**	5.18	**1.64**	**1.8**	**5.30**	5.39	**5.58**	**18.00 %**

Table 1. Quantitative evaluation on the optical flow task. The numbers reported here are all average end-point-error (EPE) except for the last column (KITTI2015 test) which is the percentage of erroneous pixels (Fl-all). A pixel is considered to be correctly estimated if the flow end-point error is <3px or <5%.

light-weighted and also achieves good performance. Differently, we modify PWCNet [40] to exploit the epipolar geometry constraints, *i.e.* p_{ss} in Eq. (1) can only be found along the horizontal axis and at the left side of p_t. Therefore, we limit the search range to the horizontal line in the cost volume, and the value of horizontal flow to be negative.

The structure of MotionNet is similar to the one used in [60] except our network only takes two consecutive images as the input instead of three or five images, and has two more convolutional layers. We use PWCNet for optical flow estimation.

The whole training process includes three stages. First, we train the FlowNet using $\mathcal{L}_{fv}, \mathcal{L}_{fs}$. At the second stage, we train the StereoNet and MotionNet jointly using $\mathcal{L}_{sv}, \mathcal{L}_{ss}, \mathcal{L}_{sc}, \mathcal{L}_{rv}$ without RDVO or flow consistency. The two pre-training stages provide us a reasonable optical flow and stereo depth estimation. At the last stage, we add our RDVO module and consistency terms, and train all networks together using the total loss $\mathcal{L}_{UnOS}$.

For inference, we obtain optical flow $\mathbf{F}_{t\to s}$ and stereo depth $\mathbf{D}_t$ directly from the corresponding networks, and obtain camera motion $\mathbf{T}^u_{t\to s}$ after RDVO. Moving object segmentation is computed by $\mathbf{1} - \mathbf{M}_t$.

5. Experiments

We evaluate UnOS on the KITTI dataset with multiple types of ground truth, and compare our results to existing supervised and unsupervised SOTA methods on the tasks of optical flow, stereo depth, visual odometry and motion segmentation.

Training Details In all of training stages, we used Adam optimizer [24] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to be 10^{-4}. The hyper-parameters $\beta = 10.0$ in Eq. (3), $\alpha = 0.85$ in Eq. (8), $\gamma = 0.17$ in Eq. (4). For the loss functional in Eq. (10), we borrow the parameters from [14] for stereo losses parameters, and $[\lambda_{fs}, \lambda_{rv}, \lambda_{sv}, \lambda_{ss}, \lambda_{sc}, \lambda_{fc}]$ are set to be $[10.0, 10.0, 1.0, 10.0, 1.0, 0.01]$ by balancing the scale of various losses without too much tuning.

During training, we use a batch size of 4. In each stage, we train for around 15 epochs and choose the model with the best validation accuracy for the start of next stage training. Images are scaled to have values between 0 and 1, and size of 832×256. The only data augmentation we perform is random left-right flipping and random time order switching.

Dataset Following previous works [14, 46, 60, 52], for the depth, optical flow and segmentation tasks, we train our networks using all of the raw data in KITTI excluding the scenes appeared in the training set of KITTI 2015 [34], which we adopt as our validation set and use to compare with other methods. We also evaluate UnOS on KITTI 2012 [11] to additionally verify our algorithm. For segmentation, we only evaluate on KITTI 2015 since there is no moving things in KITTI 2012. For the odometry task, we use the official odometry split, *i.e.* using sequences 00-08 as training and sequences 09, 10 as validation. All of our models are trained from scratch in a pure unsupervised manner.

5.1. Evaluation

Optical flow. We evaluate our method on the optical flow estimation task using both KITTI 2012 and KITTI 2015, and the quantitative results are shown in Tab. 1. UnOS (FlowNet-only) is our baseline model after training FlowNet in the first stage. We could see that it is better than one of the unsupervised optical flow method UnFlow-CSS [33] demonstrating the effectiveness of our occlusion-aware loss and PWC network structure. "UnOS (Ego-motion)" is the result of rigid flow, *i.e.* computing flow using $p_{rs} - p_t$, at the end of the second stage training. The rigid flow is shown to be better than the previous general optical flow in

Method	Train Stereo	Test Stereo	Super-vised	Lower the better					Higher the better		
				Abs Rel	Sq Rel	RMSE	RMSE log	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
EPC [52]	✓			0.109	1.004	6.232	0.203	–	0.853	0.937	0.975
Zhou et al. [59]	✓	✓		–	–	–	–	9.41%	–	–	–
SegStereo [51]	✓	✓		–	–	–	–	8.79%	–	–	–
Godard et al. [14]	✓	✓		0.068	0.835	4.392	0.146	9.194%	0.942	0.978	0.989
Zhong et al. [57]	✓	✓		0.075	1.726	4.857	0.165	6.424%	0.956	0.976	0.985
OpenWorld [58]	✓	✓		(0.056)	(0.692)	(3.176)	(0.125)	(5.140%)	(0.967)	–	–
UnOS (Stereo-only)	✓	✓		0.060	0.833	4.187	0.135	7.073%	0.955	0.981	0.990
UnOS (Ego-motion)	✓	✓		0.052	0.593	3.488	0.121	6.431%	0.964	**0.985**	0.992
UnOS (Full)	✓	✓		**0.049**	**0.515**	**3.404**	**0.121**	5.943%	**0.965**	0.984	**0.992**
PSMNet	✓	✓	✓	–	–	–	–	**1.83%**	–	–	–

Table 2. Quantitative evaluation of the stereo depth task on the KITTI2015 training set. Abs Rel, Sq Rel, RMSE, RMSE log, $\delta < 1.25, 1.25^2, 1.25^3$ are standard metrics for depth evaluation [60]. We capped the depth to be between 0-80 meters to compare with existing literature. D1-all is the error rate of the disparity. Please note that the OpenWorld results were obtained by directly training on the KITTI2015 training set by replicating the images 200 times to form a pseudo-video, therefore not directly comparable with other methods which held out that dataset.

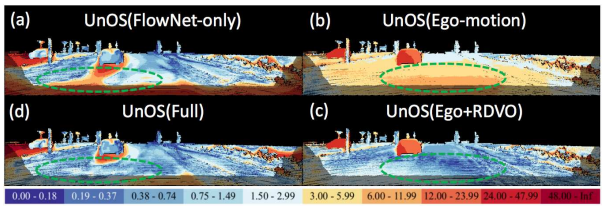

Figure 5. Visualization of the flow error map for different stages in the training. Color legend for errors is plotted at the bottom.

Method	frames	Stereo	Sequence 09	Sequence 10
ORB-SLAM(Full)	All		0.014 ± 0.008	0.012 ± 0.011
Zhou *et al.* [60]	5		0.016 ± 0.009	0.013 ± 0.009
Geonet [54]	5		0.012 ± 0.007	0.012 ± 0.009
Mahjourian*et al.* [30]	3		0.013 ± 0.010	0.012 ± 0.011
Adv. [37]	5		0.012 ± 0.007	$\mathbf{0.012 \pm 0.008}$
UnOS (MotionNet)	2	✓	0.023 ± 0.010	0.022 ± 0.016
UnOS (+RDVO)	2	✓	0.013 ± 0.006	0.015 ± 0.010
UnOS (Full)	2	✓	$\mathbf{0.012 \pm 0.006}$	0.013 ± 0.008

Table 3. Quantitative evaluation of the odometry task using the metric of the absolute trajectory error.

Method	Sequence 09		Sequence 10	
	$t_{err}\%$	$r_{err}(^\circ/100)$	$t_{err}\%$	$r_{err}(^\circ/100)$
ORB-SLAM(Full)	15.30	**0.26**	**3.68**	**0.48**
Zhan *et al.* [56]	11.92	3.60	12.62	3.43
UnOS (MotionNet)	13.98	5.36	19.67	9.13
UnOS (+RDVO)	8.15	3.02	9.54	4.80
UnOS (Full)	**5.21**	1.80	5.20	2.18

Table 4. Quantitative evaluation of the odometry task using the metric of average translational and rotational errors. Numbers of ORB-SLAM (Full) are adopted from [56].

occluded (6.67 vs. 11.2) and static (4.53 vs. 7.68) regions. This observation is consistent with our assumption about the advantage of the rigid flow in those areas, and provides motivation for our proposed flow consistency loss ($\mathcal{L}_{fc}$). The rigid flow is much worse in moving regions which is expected since it is only supposed to be accurate in static regions. "UnOS (Ego+RDVO)" is the result of refined rigid flow, *i.e.* computing flow using $p^u_{rs} - p_t$ after RDVO without the third stage training. The result shows that the rigid alignment module significantly improves the rigid flow in static regions (1.93 vs. 2.86 and 2.99 vs. 4.53). "UnOS (Full)" represents our optical flow estimation at the end of the third training stage with flow consistency. It is still worse than rigid flow after RDVO in static regions but has the best overall performance. For KITTI 2012, our method reduces the error from previous unsupervised method [33] by 50%, and reaches similar performance of the supervised methods [19], which demonstrates the benefits of our proposed method and the utilization of stereo data. For KITTI 2015, our method also outperforms previous unsupervised methods by a large margin, although it still lags behind the corresponding supervised methods [40]. Visualization of our estimated optical flow can be found in Fig. 6, and we can see that our results are more regularized with sharper boundaries.

We also show the error map of optical flow from different training stages in Fig. 5 (bluer means better while redder means worse). Initially, rigid flow from ego-motion 'UnOS(Ego-motion)' has worse performance than 'UnOS(FlowNet-only)'. After adding RDVO, we can see that the flow estimations in the static region are greatly improved (comparing green circles in (b) and (c)). After adding the moving object mask and applying the consistency loss, 'UnOS(Full)' shows even better results in both static and occluded regions compared to our baseline (comparing green circles in (a) and (d)).

Stereo-depth. We evaluate our depth estimation on the KITTI 2015 dataset, and show the resutls in Tab. 2. Here, the numbers of Zhong *et al.* [57] and OpenWorld [58] were obtained through private communications with the authors. "UnOS (StereoNet-only)" is the StereoNet trained using only stereo images, and is our baseline algorithm. It is al-

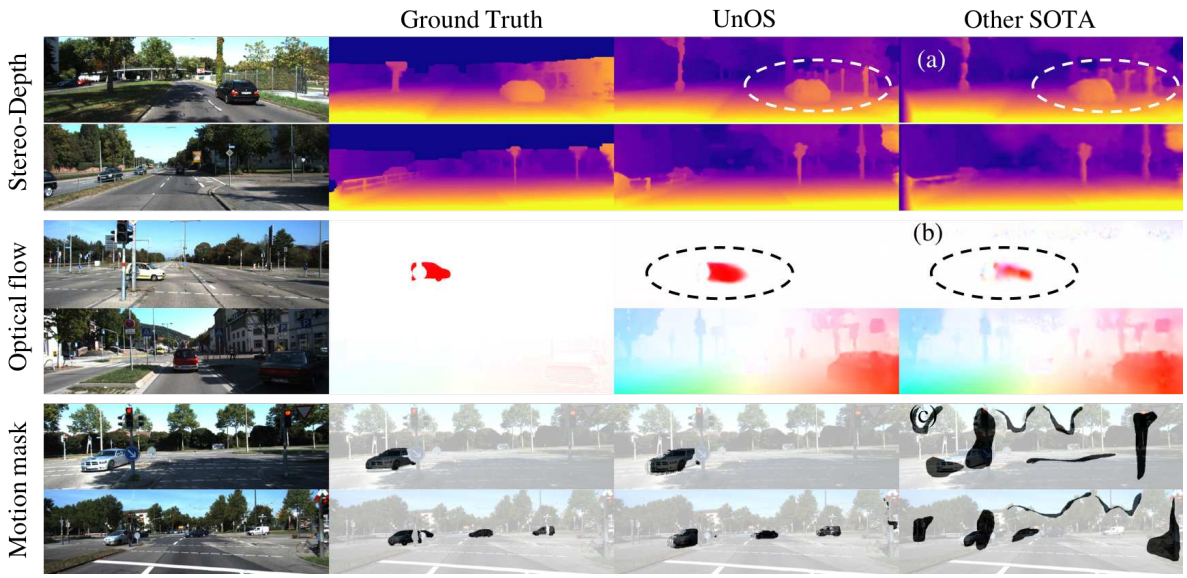

Figure 6. Qualitative results of UnOS. We compare each of our output to previous SOTA results. Specifically, (a) Godard *et al.* [14], (b) UnFlow-CSS [33], (c) EPC [52].

ready better than some of the existing unsupervised stereo depth algorithms [14, 51] demonstrating the effectiveness of our StereoNet. Our stereo depth also performs much better than the SOTA monocular depth method [52]. "UnOS (Egomotion)" shows the results at the end of our second training stage. After adding the data of time consecutive images, the depth accuracy improves especially in the large distance regions (0.593 vs. 0.833). "UnOS (Full)" shows the results after using RDVO with rigid-aware flow consistency, and gives the best performance. However, its performance is still worse than the supervised method like PSMNet [5]. We provide 3D scene flow evaluation on KITTI 2015 test set in the supplementary materials, and the qualitative results of our estimated depth are shown in Fig. 6, where UnOS figures out better scene structures with less noise.

Visual odometry. We evaluate camera motion using two commonly adopted metrics. The first one was proposed in SfMLearner [60] which measures the absolute trajectory error averaged over all overlapping 5-frame snippets after factor rescaling with the ground truth. In our case, we only have two frames as input to the MotionNet to predict their relative pose. For evaluation, we accumulate 4 consecutive predictions to get the result for the 5-frame snippet. The other metric was proposed in [56] which measures the average translation and rotation errors for all sub-sequences of length (100, 200, ..., 800). For this metric, we accumulate all of two frames estimations together for the entire sequence without any post-processing. The results for the two metrics are shown in Tab. 3 and Tab. 4 respectively. We can see direct output from MotionNet (UnOS (MotionNet)) is not satisfying, and is much worse than other SOTA methods. However, after RDVO module, we see significant improvements (UnOS (MotionNet+RDVO)). After training with the flow consistency and RDVO, the results can be further improved, and on par with other SOTA methods despite using stereo info. In Tab. 4, UnOS is worse than traditional ORB-SLAM, we argue that it uses bundle adjustment to avoid drifting error, which is complementary to UnOS.

Method	Pixel Acc.	Mean Acc.	Mean IoU	f.w. IoU
EPC [52]	0.89	0.75	0.52	0.87
UnOS (Full)	**0.90**	**0.82**	**0.56**	**0.88**

Table 5. Motion segmentation evaluation. The metrics are pixel accuracy, mean pixel accuracy, mean IoU, and frequency weighted IoU.

Motion segmentation. The motion segmentation task is evaluated using the object map provided by the KITTI 2015 dataset [52], where the moving objects are manually segmented. We follow the metrics used in [52] including pixel accuracy and mean intersect-over-union. As shown in Tab. 5, we also outperform their method. The qualitative results are shown in Fig. 6, where UnOS discovers more compact and cleaner segments for moving objects.

6. Conclusion

In summary, our paper propose an unified system (UnOS) to learn optical flow and stereo-depth, which mutually leverages stereo and temporal information in a video. Specifically, it automatically discovers rigid regions, and substantially improves unsupervised learning of stereo-depth, optical flow, visual odometry and motion segmentation on the KITTI dataset.

References

[1] Dan Barnes, Will Maddern, Geoffrey Pascoe, and Ingmar Posner. Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments. In *ICRA*, pages 1894–1900. IEEE, 2018.

[2] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *International Conference on Computer Vision*, 2017.

[3] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–607. International Society for Optics and Photonics, 1992.

[4] Michael J Black and Paul Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996.

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[6] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *European Conference on Computer Vision*, pages 108–125. Springer, Cham, 2018.

[7] Guilherme N DeSouza and Avinash C Kak. Vision for mobile robot navigation: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 24(2):237–267, 2002.

[8] Yiliu Feng, Zhengfa Liang, and Hengzhu Liu. Efficient deep learning for stereo matching with larger image patches. In *Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017 10th International Congress on*, pages 1–5. IEEE, 2017.

[9] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

[10] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3354–3361. IEEE, 2012.

[12] Andrea Giachetti, Marco Campani, and Vincent Torre. The use of optical flow for road navigation. *IEEE transactions on robotics and automation*, 14(1):34–48, 1998.

[13] Clément Godard, Oisin Mac Aodha, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. *arXiv preprint arXiv:1806.01260*, 2018.

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.

[15] Fatma Guney and Andreas Geiger. Displets: Resolving stereo ambiguities using object knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4165–4175, 2015.

[16] Rostam Affendi Hamzah and Haidi Ibrahim. Literature survey on stereo vision disparity map algorithms. *Journal of Sensors*, 2016, 2016.

[17] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[18] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[19] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[20] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[21] Joel Janai, Fatma Güney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 690–706, 2018.

[22] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *Computer Vision–ECCV 2016 Workshops*, pages 3–10. Springer, 2016.

[23] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[24] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Junghwan Ko and Jungsuk Lee. Stereo camera-based intelligence surveillance system. *Journal of Automation and Control Engineering Vol*, 3(3), 2015.

[26] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. Undeepvo: Monocular visual odometry through unsupervised deep learning. *arXiv preprint arXiv:1709.06841*, 2017.

[27] Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011.

[28] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[29] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *CVPR*, pages 155–163, 2018.

[30] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[31] David Marr. A computational investigation into the human representation and processing of visual information. *Freeman, San Francisco, CA*, 1982.

[32] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

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

[34] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[35] Jean-Philippe Pons, Renaud Keriven, and Olivier Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

[36] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 2. IEEE, 2017.

[37] Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *arXiv preprint arXiv:1805.09806*, 2018.

[38] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, pages 1495–1501, 2017.

[39] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2017.

[40] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *arXiv preprint arXiv:1709.02371*, 2017.

[41] Tatsunori Taniai, Sudipta N Sinha, and Yoichi Sato. Fast multi-frame stereo scene flow with motion segmentation. In *CVPR*, pages 6891–6900. IEEE, 2017.

[42] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3899–3908, 2016.

[43] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 722–729. IEEE, 1999.

[44] Sundar Vedula, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):475–480, 2005.

[45] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, 2018.

[46] Yang Wang, Yi Yang, Zhenheng Yang, Peng Wang, Liang Zhao, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4884–4893, 2018.

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[48] Changchang Wu et al. Visualsfm: A visual structure from motion system. 2011.

[49] Jonas Wulff, Laura Sevilla-Lara, and Michael J. Black. Optical flow in mostly rigid scenes. In *CVPR*, July 2017.

[50] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European Conference on Computer Vision*, pages 842–857. Springer, 2016.

[51] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. *arXiv preprint arXiv:1807.11699*, 2018.

[52] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. *arXiv preprint arXiv:1806.10556*, 2018.

[53] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. *arXiv preprint arXiv:1711.03665*, 2017.

[54] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018.

[55] Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.

[56] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 340–349, 2018.

[57] Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.

[58] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. *arXiv preprint arXiv:1808.03959*, 2018.

[59] Chao Zhou, Hong Zhang, Xiaoyong Shen, and Jiaya Jia. Unsupervised learning of stereo matching. In *International Conference on Computer Vision*, 2017.

[60] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017.

[61] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision*, 2018.