future internet

MDPI

Article

Occlusion-Aware Unsupervised Learning of Monocular Depth, Optical Flow and Camera Pose with Geometric Constraints

Qianru Teng [1], Yimin Chen [1,2,*] and Chen Huang [1]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; tqr0416@shu.edu.cn or ghost416@msn.cn (Q.T.); channinghuang@shu.edu.cn or huangchenemail@163.com (C.H.)
[2] Shanghai Institute for Advanced Communication and Data Science, Shanghai 200444, China
* Correspondence: ymchen@mail.shu.edu.cn

Received: 13 August 2018; Accepted: 13 September 2018; Published: 21 September 2018

Abstract: We present an occlusion-aware unsupervised neural network for jointly learning three low-level vision tasks from monocular videos: depth, optical flow, and camera motion. The system consists of three different predicting sub-networks simultaneously coupled by combined loss terms and is capable of computing each task independently on test samples. Geometric constraints extracted from scene geometry which have traditionally been used in bundle adjustment or pose-graph optimization are formed as various self-supervisory signals during our end-to-end learning approach. Different from prior works, our image reconstruction loss also takes account of optical flow. Moreover, we impose novel 3D flow consistency constraints over the predictions of all the three tasks. By explicitly modeling occlusion and taking utilization of both 2D and 3D geometry relationships, abundant geometric constraints are formed over estimated outputs, enabling the system to capture both low-level representations and high-level cues to infer thinner scene structures. Empirical evaluation on the KITTI dataset demonstrates the effectiveness and improvement of our approach: (1) monocular depth estimation outperforms state-of-the-art unsupervised methods and is comparable to stereo supervised ones; (2) optical flow prediction ranks top among prior works and even beats supervised and traditional ones especially in non-occluded regions; (3) pose estimation outperforms established SLAM systems under comparable input settings with a reasonable margin.

Keywords: monocular depth; camera pose; optical flow; joint learning; occlusion-aware; scene geometry

1. Introduction

The ability to perceive visual environment of one agent is more required than ever due to the advancing expand in industries such as robotics [1], autonomous driving [2] and augmented reality [3], among which has many crucial tasks to fulfill including reasoning one's ego-motion and the scene structure. In this paper, we propose a jointly learning network in an utterly unsupervised manner to predict depth, camera pose and optical flow from monocular video sequences with no labeling data or ground truth.

Years of research in scene understanding has already been studied. Structure from motion (SfM) [4–6] is a long-standing task in computer vision which aims at reconstructing camera motion and scene structure, but often hard to integrate reasonable priors for small camera translation or outliers from low-level sparse correspondences. Visual Simultaneous localization and mapping (VSLAM) [7–9] use handcrafted sparse or dense abstractions to represent geometry but also limited to a particular kind of scenes. They also suffer from absolute scale regarding monocular approach and noisy data

regarding depth sensor. In this paper, we employ deep neural networks for better representation of high-level cues instead of being restricted to a certain scenario.

Given the advantage of numerous data and learning effectiveness of deep network, recent works have emerged to formulate these tasks through deep models and achieve considerable results compared to traditional ones. Most previous works [10–14] target at one specific task only, they neglect the inherent relationships in between relative tasks. Ref. [15,16] combine depth prediction with camera motion by image reconstruction loss based on photometric quality. Ref. [17] comprise optical flow in their work in an implicit way, predicting depth and camera motion meantime. Ref. [18] introduce 3D geometric constraint during propagation. However they all did not explicitly handle occlusions. Ref. [19] form these three tasks through a stage-wise structure which enforce geometric consistency with non-rigidity filtered in the second stage, but they cannot fully exploit on geometric relations due to the cascade network structure. Therefore, we focus on avoiding these shortcomings and elaborate on them to achieve more accurate results.

Inspired by works that impose geometric constraints on learning procedures [18,19], we also make our effort in exploiting scene structure constraints in a deep learning system. The main idea of this paper is by taking account of optical flow simultaneously in addition to depth and camera motion estimation, jointly learning a network with both 2D and 3D geometry transformations during training to create additional constraints for better exploitation of nature geometry structure. Since the depth and camera motion as well as optical flow are calculable, we are able to make following contributions:

1. Jointly learn an unsupervised deep neural network from monocular videos that predicts depth, optical flow, and camera motion simultaneously at training time coupled by combined loss term.
2. Modeling occlusion all through entire process explicitly through bidirectional flow from consecutive frames to make the model occlusion-aware and non-occluded region better constrained.
3. Mutually supervise each component of the network in use of both 2D and 3D geometric constraints combined with occlusion module. The 2D image reconstruction loss takes optical flow into consideration. The 3D constraint contains two part: 3D point alignment loss and a novel 3D flow loss.

2. Related Work

2.1. Deep Learning vs. Geometry for Scene Understanding

The study of estimating scene structure and camera motion from sequences of images or simultaneously mapping and localization in computer vision was considered a purely geometric problem for decades. It is usually accomplished by pipelines containing several successive processing steps, as well as simultaneously mapping and localization. However the handcrafted methods often highly rely on accurate image correspondence, and structure will be ill-posed if arbitrary deformations are allowed. Whereas deep learning is more capable of capture relatively random details and learning from them. Possessed of such qualities, more recent methods focus on how to alleviate such reliance by introducing deep learning into geometric problems. However, rather than apply deep learning models naively, imposing geometry in deep learning allow us to learn a geometric problem without massive amount of labeled data, extracting enforcement from nature structure. This is an exciting breakthrough and has proven to be effective in many researches.

2.2. Deep Learning with Geometry for Scene Understanding

Deep learning from videos has made significant progress since it first appeared and has a promising future. Many works have explored tasks including depth prediction, optical flow, and pose estimation. These approaches are mainly divided into two categories, supervised methods and unsupervised ones.

2.2.1. Supervised Videos Learning

The work of Eigen et al. [10] manifested deep models' capability for single view depth estimation with a coarse-to-fine strategy. Kendall et al. [12] proposed a stereo regression architecture for sub-pixel disparity from a rectified pair of stereo images by leveraging knowledge of the problem's geometry. Brahmbhatt et al. [13] formulated loss terms from geometric constraints expressed by sensory inputs which usually used in traditional ways were exploited to bring up camera localization accuracy. Similar spirits have been shared in learning optical flow. Ref. [14,20] proposed FlowNet to compute dense flow prediction through fully connected convolutional neural networks in an end-to-end fashion supervised by synthetic datasets. Ummenhofer et al. [16] engineered settings that alternate optical flow estimation with camera motion and depth estimation, which required various forms of supervision including an abundant amount of scanned depth data. To eliminate heavy reliance on large labeled data, an unsupervised setting is exploited in our method.

2.2.2. Unsupervised Videos Learning

Garg et al. [21] leveraged well-understood ideas in visual geometry by proposing a coarse-to-fine stereopsis-based auto-encoder to predict single view depth using projection objective. Godard et al. [11] exploited epipolar geometry constraints consistency between the disparities generated during network training from monocular videos by introducing a left-right consistency loss. While such stereo formulation has a heavier reliance on scene priors, a monocular setting is preferred by many recent methods.

Ref. [22] achieved a monocular VO system along with dense depth map with recovered absolute scale in an unsupervised manner by utilizing both temporal and spatial geometric constraints. Zhou et al. [15] formulated a view synthesis pipeline which learns monocular depth and ego-motion in a coupled way by building upon the rigid projective geometry with an explainability mask for compensation of any dynamic factor. Concurrently, Vijayanarasimhan et al. [17] mimicked the traditional problem of structure from motion through explicit modeling of scene geometry and several object masks. Ref. [18] considered the inferred 3D geometry with a proposed objective which directly penalizes inconsistencies in the estimated depth during image reconstruction process. However, they all overlook the fact that including outliers such as occlusions in training could potentially corrupt the process. Meister et al. [23] introduced an occlusion-aware bidirectional census loss in optical flow learning. Yin and Shi [19] proposed a divide-and-conquer strategy to solve depth, optical flow, and camera motion estimation in a combined way constrained by reconstruction loss formulated of geometric relationships extracted from the tasks, demonstrating that learning a non-rigid flow residual is helpful to their designed 3D geometric scene understanding. However, in their rigid flow stage, learning from overall region including occluded area may cause side effects. The optical flow was mainly to refine the rigid-flow produced in the first stage, hence they struggled in capturing more inherent geometric nature between these tasks.

Even though unsupervised-based methods have gained attention in recent times due to non-requirement of any type of ground truth and have made considerable improvement; however, there is still need to discover better-suited network structures and loss functions. Differently from Yin and Shi [19], by imposing optical flow into our system and training synchronously, we are able to form the combined losses over all predictions which contain abundant geometric information to lead the backpropagation proving to be effective after. Furthermore, occlusion masks produced from optical flow estimation are employed for all three tasks all through entire training which will ensure the filtering of non-rigid cases.

3. Method

Our approach combines depth estimation, optical flow, and camera motion in a whole, in which they can jointly benefit from each other effectively through geometric observation. This section starts by giving an overview of our system and then explain each component individually.

3.1. System Overview

The schematic network architecture is showed in Figure 1. It handles these three tasks separately with the utilization of three sub-networks, within which each one of them is capable of jointly training and supervising one another mutually, resolving the scene perception task easier. Additionally, rich occlusion-aware geometric constraints are applied within and between these tasks to supervise one task another mutually. We present a way to employ the nature of occlusion to a greater extent instead of simply filtering it out, by combining it with various geometric constraints.

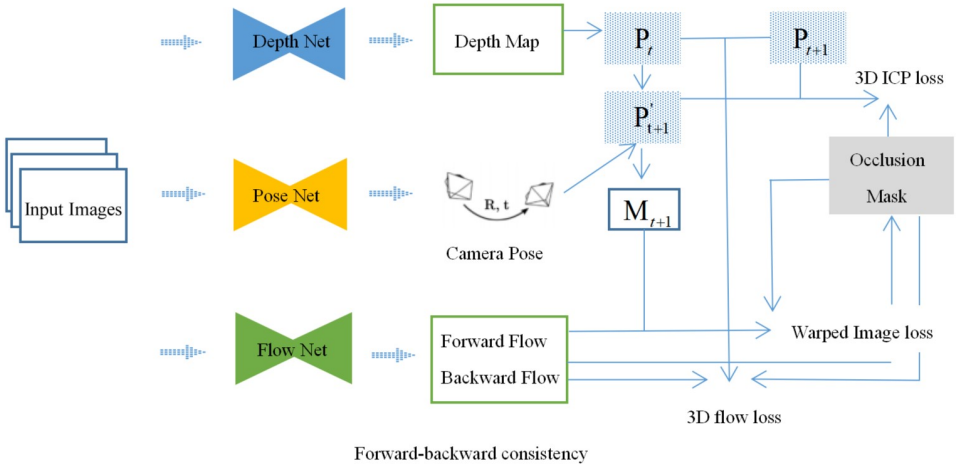


Figure 1. Overview of our system. It consists of three task-specific sub-networks targeting at estimating monocular depth, optical flow, and camera motion. Rich geometric constraints are employed extracting from the natural structure of the scene. Here M denotes input image and P denotes point cloud.

During training, we formulate this problem as: Given consecutive frames M_{t-1}, M_t with known camera intrinsic, D_t as per-pixel depth map of each frame is estimated, and T_{t-1} denotes relative camera motion from $t-1$ to t, F_f presenting forward optical flow while F_b standing for the backward optical flow. The forward and backward optical flow are used to generalize the occlusion mask to reduce the influence of occlusion when applying rigid geometric constraints. As depth and camera motion are available, we generate point cloud P_t of frame M_t which is a basic element throughout our 3D constraints of the network. The frames are preprocessed as four pyramid scales from l_1 to l_4. Details of the methods are discussed below.

3.2. Bidirectional Flow Loss as Occlusion Mask

To detect occlusion, we conduct the method recommended by [23] which is based on forward-backward consistency assumption to mask the occluded area. The forward flow $F_t^f(u_f, v_f)$ and backward flow $F_t^b(u_b, v_b)$ are computed by performing a second pass with the input images exchanged with shared weights and symmetric loss. Occluded pixels are marked whenever there is a significant mismatch between these two flows. The occlusion flag is set to be 1 when the following constraint

$$\|F_t^f + F_t^b|^2 < \alpha_1(|F_t^f|^2 + |F_t^b|^2) + \alpha_2 \tag{1}$$

is against, and 0 otherwise. In our experiment we set $\alpha_1 = 0.01$, $\alpha_2 = 0.5$. Then the occluded pixel mask of frame M_t is calculated as O_t.

Forward-backward consistency between adjacent flows for the non-occluded region is applied during training, along with a constant penalty $\lambda_p = 3.5$ to all occluded pixels in case the whole scene happens to be occluded:

$$L_{fb} = \sum_{X \in M} (1 - O_t)\rho(\|F_t^f(X) + F_t^b(X)\|_1) + O_t\lambda_p \tag{2}$$

with X denoting pixels of image M, ρ denoting the robust generalized Charbonnier penalty $\rho(x) = \sqrt{x^2 + 0.001^2}$.

3.3. 2D Image Reconstruction Loss as Supervision

Following the fashion of most unsupervised approaches, we conduct synthesis image as fundamental supervision of each task in our network by using *differentiable inverse warping* [24] with a bilinear sampler between nearby frames. Given frame M_t and depth D_t, the image projected into a structured three-dimensional point cloud can be presented as:

$$P_t(x, y, z) = D_t(x, y) \cdot K^{-1} \cdot [i, j, 1]^T \tag{3}$$

where (i, j) are the coordinates of the image pixel and (x, y, z) are the point coordinates, K is the homogeneous camera intrinsic matrix. This way we compute per frame point cloud $Pt, t \in (1, ...)$.

We transform the nearby point cloud P_{t-1} using camera intrinsic and known rigid scene transformation T following $P_t^{'} = KTP_{t-1}$, then back project it to the image plane to get the warped frame M_t^p. M_t^p is warped through both forward and backward directions.

The self-supervision approach for optical flow is similar, only the warped image M_t^f is obtained by the predicted flow, which is also in both directions. Let F_t^f be the forward flow from M_{t-1} to M_t, F_t^b be the backward flow from M_t to M_{t-1}. The M_t^f warped from M_{t-1} is formulated as $M_{t-1}(X + F_t^f(X))$, where X denotes image pixel of frame M_{t-1}. For the backward direction, we define in the same way with F_t^f and F_t^b exchanged.

The reconstruction consistency loss is then produced by comparing image M_t to warped image $M_t^{'}$. Note that the warping schemes make implicitly assumption that there is no occlusion in the scene so that incorrect deformations could be learned from the occluded regions. Therefore we mask occluded pixels to alleviate the negative effect of outliers and penalize photometric difference only for every non-occluded pixel. The occlusion mask is modeled as O_t in Section 3.2, the photometric consistency loss is formulated as:

$$L_{re} = \sum_{X \in M} ((1 - O_t)\|M_t(X), M_t^p(X)\|_1 + (1 - O_t)\|M_t(X), M_t^f(X)\|_1) \tag{4}$$

By performing this way we also take into account the discrepancy between warped image M_t^p and M_t^f without occlusion since they should be ideally identical if predictions of the network are perfect.

3.4. Mutually Supervised 3D Geometric Consistency Loss

3.4.1. 3D Point Alignment Loss

An additional 3D geometric constraint is exploited to the non-occluded area in a differentiable way recommended by [18] to reinforce the backpropagation for depth map and camera pose estimations. The core of this constraint is a rigid point matching algorithm, Iterative Closest Point(ICP) which calculates a transformation between two 3D points by optimizing the minimal point-to-point error, in our case, between point $P_t(1 - O_t)$ and warped point $P_t^{'}(1 - O_t)$. To be specific, we use both outputs, the best-fit transformation $T^{'}$ and a residual distance r after the best transformation has been applied to guide the regression. r is used as an approximation to the negative gradient of the loss about depth D_t for adjustment and $T^{'}$ is to adjust camera pose. The 3D constraint is formulated as:

$$L_{icp} = \|T^{'} - T\|_1 + \|r\|_1 \tag{5}$$

3.4.2. 3D Flow Loss

Figure 2 depicts how 3D flow loss is formed in our method. For the better understanding of scene geometry, we form an additional geometric enforcement to maintain the coherence between the 3D flow generated from estimated 2D flow and from point cloud, so that all the information we obtained from these three sub-tasks can be utilized as an unified constraint. As $F_t^f(u_f, v_f)$ is the predicted dense 2D flow on X-Y image plane from frame t to $t+1$, and $D_t(x,y)$ is the predicted depth map, the 3D flow can be calculated as:

$$F_t^f(z) = D_{t+1}((x,y) + F_{tf}(u,v)) - D_t(x,y) \tag{6}$$

$$F_{3d} = C(F_{tf}(u,v) + F_t^f(z)) \tag{7}$$

where C denotes the concatenation operation. For the 3D flow computed from point cloud using both depth and pose prediction, is formulated as:

$$F_{3d}^{'} = KTP_t(x,y,z) - P_t(x,y,z) \tag{8}$$

Thus, the complete mutually supervised geometric consistency loss is:

$$L_{mu} = \sum_{(x,y,z)\in P} (1 - O_t)\|F_{3d}(x,y,z), F_{3d}^{'}(x,y,z)\|_1 \tag{9}$$

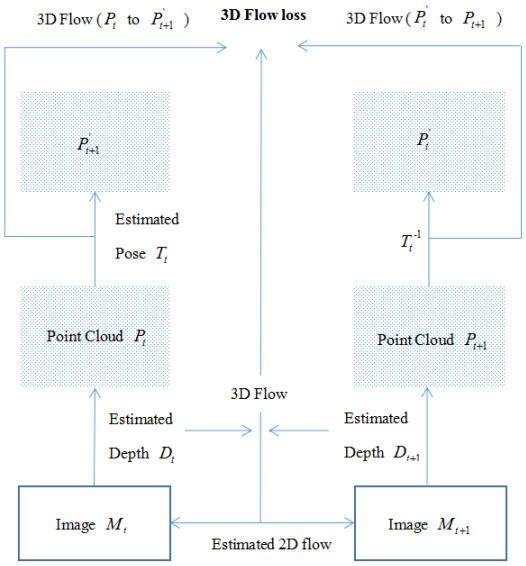


Figure 2. The 3D Flow is computed for both directions. The loss takes utilization of estimated depth, optical flow and camera pose.

3.5. Smoothness Constraint

However, depth borders have a tendency to be locally blurred. To smooth out prediction discontinuities and preserve sharp details, edge-aware weighted smoothness penalties are adapted to our system. The occluded area is solely guided by smoothness loss since it violates both photometric and geometric consistency. Depth smoothness is formulated by image gradients while second-order derivative is used for optical flow loss term:

$$L_{sm} = \sum_{X\in M} |\nabla D_t(X)| \cdot (e^{-\alpha|\nabla M_t(X)|})^T + \sum_{X\in M} |\partial_d^2 F_f(X)| \cdot e^{-\alpha|\partial_d M_t(X)|} \tag{10}$$

where α controls the weight of edges which is set to be 3 in our experiment, d indexes over partial derivative on x and y directions.

The final loss term is a weighted summation of above all in which λ denotes loss weights and l indexes over four different pyramid image scales:

$$L = \sum_{l=(1,2,3,4)} (\lambda_{fb} L_{fb} + \lambda_{re} L_{re} + \lambda_{icp} L_{icp} + \lambda_{mu} L_{mu} + \lambda_{sm} L_{sm}) \quad (11)$$

4. Experiments

In this section we first introduce our implementation specification, including network architecture and training details. Then we show quantitative and qualitative performance in each of these tasks respectively compared with prior approaches.

4.1. Implementation Specification

4.1.1. Network Architecture

Jointly train three sub-networks could be highly computational burdened. Thus under this consideration, we choose rather generic networks to conduct and evaluate our method. For depth estimation, we adopt the network from [11] which uses skip connections between encoder and decoder at different corresponding resolutions. For optical flow, it is based on a modified structure of FlowNetS proposed by [14]. This Optical flow prediction is generated using a multi-stage refinement process. Both our depth and optical flow networks take two consecutive images as input. For camera pose regression the structure in [15] which regress 6-DoF camera pose is implemented. The input to the pose estimation network is one target view concatenated with two source views. A multi-scale image pyramid strategy is applied in image preprocessing so that all tasks can benefit from different scale information.

4.1.2. Training Details

Our unsupervised network is trained end-to-end on monocular video streams from KITTI dataset [25] using TensorFlow framework [26] and Adam optimizer with $\beta 1$ = 0:9 and $\beta 2$ = 0:999. The KITTI dataset contains sequences captured from a moving vehicle. The initial learning rate is set to be 0.0002 and decrease gradually when there is convergence in training loss. The mini-batch size is 4. During training, we train our model at a reduced resolution of 416 $\times$ 128 pixels over 400,000 iterations to make the computational burden decreased. Our experiments are performed on single NVIDIA Quadro M5000 GPU. We experimentally find that a weighting combination of $[\lambda_{fb}; \lambda_{re}; \lambda_{icp}; \lambda_{mu}; \lambda_{sm}] = [0.3; 1; 0.5; 0.5; 0.8]$ in our final loss function results in a stable training.

We evaluated our method with different test splits from training for each task on the popular KITTI dataset for the sake of fair comparison. The splits are same as Zhou [15]. Our joint training computes forwardly corresponding to each of these three tasks independently, and backpropagate through combined loss. Data augmentation is implemented as a widely used strategy to improve generalization of neural networks which is crucial to avoid over-fitting.

4.2. Experimental Evaluation

4.2.1. Depth Evaluation

We take the split provided by [10] to compare with prior related methods to evaluate the performance of our network in monocular depth estimation. We exclude the visually similar frames to the test scenes according to [15]. The predictions are resized at input image resolution by interlinear interpolation. We use ground truth obtained by a laser scanner. Both 50 m and 80 m threshold of maximum depth for evaluation are used. It takes two consecutive frames during training because of

the 3D alignment loss. We compare with both supervised methods and state-of-the-art unsupervised methods. We also compare with the results of Garg et al. [21] which takes stereo pairs to predict depth. Some qualitative results are shown in Figure 3. The metrics in Tables 1–3 show the quantitative results of all standard error measures of these methods. To verify the advantage of the network architecture and the effectiveness of our loss terms, we trained "Ours(VGG)" on KITTI which shares the same structure with [15]. The results validated our conclusion. Furthermore, "Ours(ResNet)" significantly outperforms both supervised works [10,27] and prior unsupervised baselines [15,18,19] with a reasonable margin, which is benefited from our extensive geometric constraint and mutually supervised mechanism. We also experimented with pre-training on Cityscapes first. As same as [19], our result is slightly inferior to [11] when trained on Cityscapes and KITTI due to characteristics of training data. We believe that such dataset stereo settings may provide more details, which we would like to explore in the future. Note that during our training process, robust similarity technique like SSIM [28] are not adopted with which should lead to further improvement.

We also conduct Ablation Experiments on the components of our network in Table 4. In order to study the importance of each component, we trained and evaluated a series of models with each one component missing. The study shows our approach is able to benefit from all these components, and 3D losses make a great difference. On the other hand, a visualized comparison to failed cases of Zhou [15] in Figure 4. further illustrates our method's capability of reasoning scene geometry. However, there are still failure cases remained when large dynamic objects close to camera or intense lighting changes happened caused by gradient locality. The search for a solution to the gradient locality of warping-based loss is still in need.

Table 1. Monocular depth results on KITTI by the split of Eigen et al. [10] capped 80 m. Results of other methods are taken from [15,18,19].

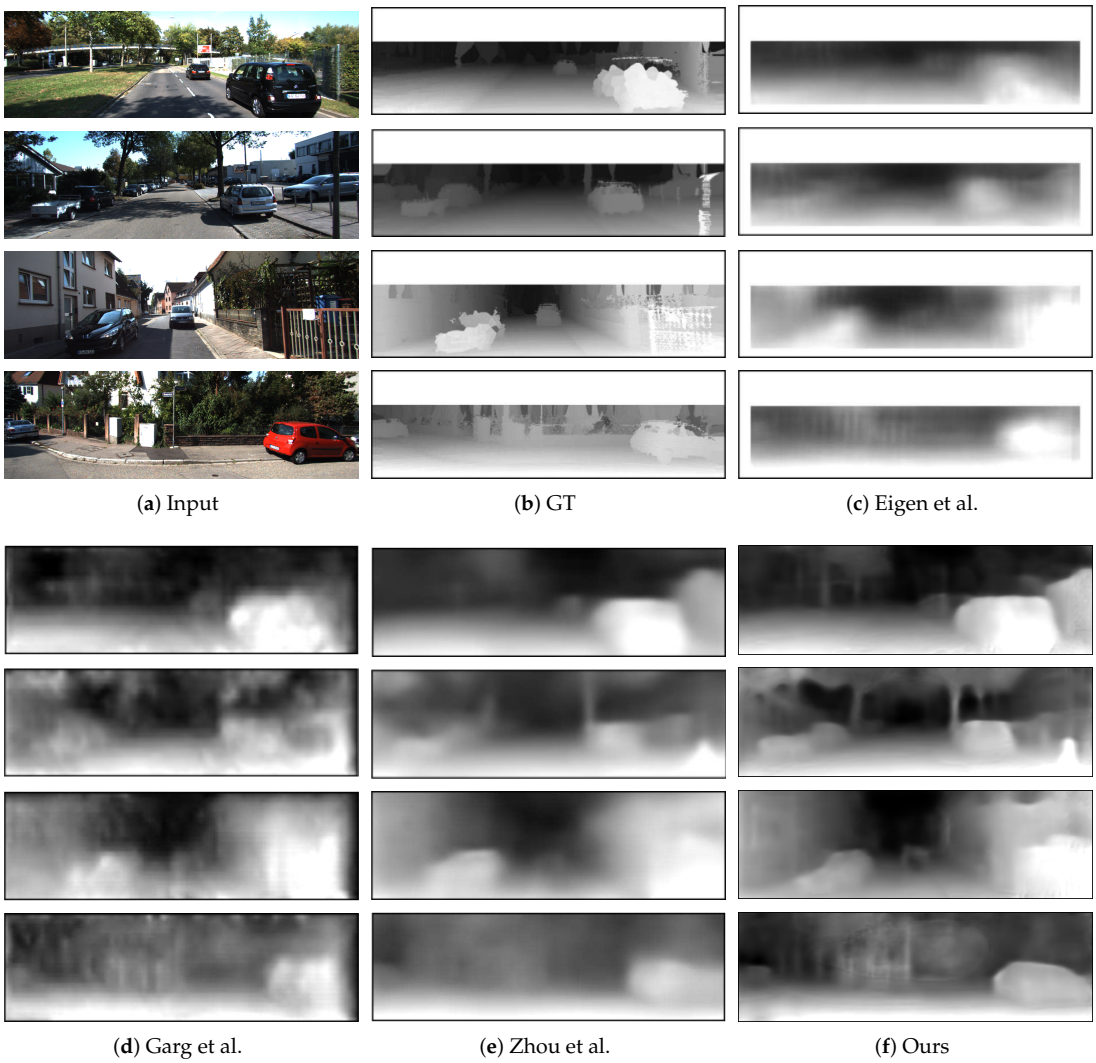
Method	Supervise	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen [10] Coarse	Depth	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen [10] Fine	Depth	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu [27]	Depth	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Godard [11]	Stereo	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhou [15]	No	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Mahjourian [18]	No	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Yin [19]	No	0.155	1.296	5.857	0.233	0.793	0.931	0.973
Ours (VGG)	No	0.157	1.229	5.960	0.238	0.799	0.932	0.973
Ours (ResNet)	No	0.149	1.223	5.732	0.225	0.829	0.944	0.978

Table 2. Monocular depth results on KITTI by the split of Eigen et al. [10] capped 50 m.

Method	Supervise	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Garg [21]	Stereo	0.169	1.080	5.104	0.273	0.740	0.904	0.962
Mahjourian [18]	No	0.155	0.927	4.549	0.231	0.781	0.931	0.975
Yin [19]	No	0.147	0.936	4.348	0.218	0.810	0.941	0.977
Ours (Resnet)	No	0.142	0.909	4.306	0.203	0.832	0.949	0.978

Table 3. Monocular depth results on Cityscapes and KITTI by the split of Eigen et al. [10] capped 80 m.

Method	Supervise	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Godard [11]	Stereo	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Zhou [15]	No	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Mahjourian [18]	No	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Yin [19]	No	0.153	1.328	5.737	0.232	0.802	0.934	0.972
Ours (Resnet)	No	0.146	1.253	5.614	0.224	0.838	0.941	0.973



(**a**) Input (**b**) GT (**c**) Eigen et al.

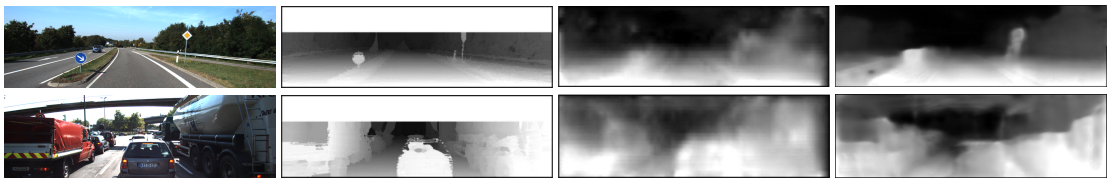
(**d**) Garg et al. (**e**) Zhou et al. (**f**) Ours

Figure 3. Comparison of monocular depth estimation between Eigen et al. [10] (supervised by depth), Garg et al. [21] (supervised by stereo), Zhou et al. [15] (unsupervised) and ours (unsupervised). Our method captures more details in the whole scene and particularly in thin structures for both close and distant regions.

Table 4. Ablation results where individual components are left out on KITTI dataset when capped 80 m.

Method	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
No Occ-Mask	0.161	1.367	6.017	0.236	0.805	0.934	0.972
No 3D ICP Loss	0.160	1.594	5.775	0.226	0.826	0.945	0.976
No 3D Flow Loss	0.157	1.353	5.971	0.232	0.811	0.938	0.975
All losses	0.149	1.223	5.732	0.225	0.829	0.944	0.978

To demonstrate the generalization ability of our model, we test the model on Make3D [29] dataset which trained only on KITTI and Cityscape. The images of Make3D dataset are in different aspect ratio thus we evaluate on a central crop of these images. Errors are only computed where depth is less than 70 m in the central image crop similar as Godard [11] and Zhou [15]. As shown in Table 5 and Figure 5, we compare with several methods including supervised ones using Make3D groundtruth depth. Though there leaves a performance gap between our method and supervised ones, our predictions beat Zhou [15] in three metrics, preserving well global scene layout and thin structures.

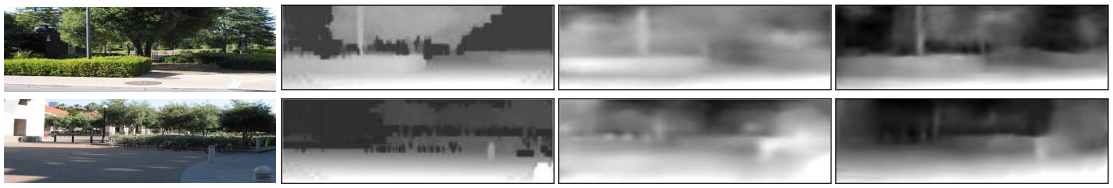


(**a**) Input (**b**) GT (**c**) Zhou et al. (**d**) Ours

Figure 4. Comparison of monocular depth estimation between failed cases of Zhou et al. [15] and same cases of ours. Our method preserves considerable predictions in vast open scenes and objects close to the front of the camera.

Table 5. Depth estimation results on the Make3D dataset.

Method	Supervision	Abs Rel	Sq Rel	RMSE	RMSE Log
Train set mean	Depth	0.876	13.98	12.27	0.307
Karsch [30]	Depth	0.428	5.079	8.389	0.149
Liu [31]	Depth	0.475	6.562	10.05	0.165
Laina [32]	Depth	0.204	1.840	5.683	0.084
Godard [11]	Pose	0.544	10.94	11.76	0.193
Zhou [15]	No	0.383	5.321	10.47	0.478
Ours	No	0.406	4.071	9.624	0.385



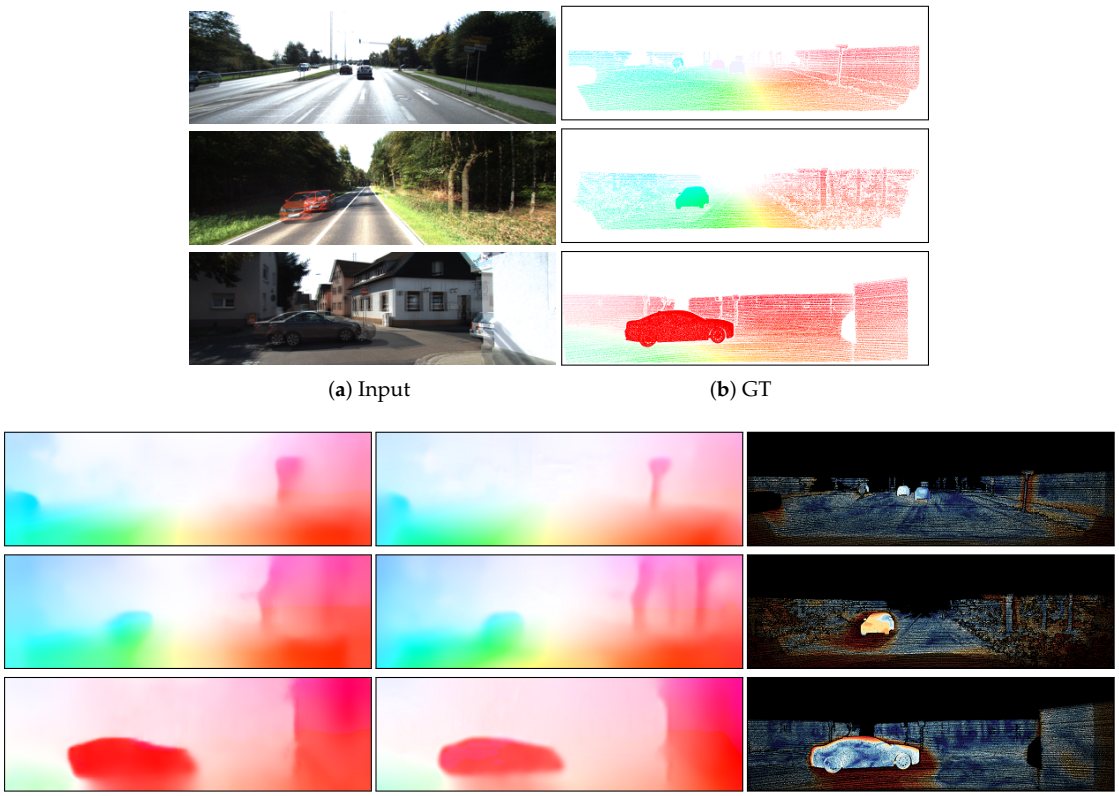
(**a**) Input (**b**) GT (**c**) Zhou et al. (**d**) Ours

Figure 5. Comparison of monocular depth estimation between Zhou et al. [15] and ours on Make3D [29] dataset. The model is trained on KITTI + Cityscapes and tested on Make3D [29] directly. Our method captures the global scene layout reasonably well and preserves better thin structures.

4.2.2. Optical Flow Evaluation

Only one of prior works include optical flow into depth and ego-motion estimation as a 3D geometric understanding scheme, thus we conduct the similar evaluation to make a fair comparison. On the KITTI flow 2015 dataset we evaluated our optical flow component. We adopted the training images as a testing set. We compared our method with previous state-of-the-art deep learning single task methods, including handcrafted EpicFlow [33], supervised FlowNetS [14] and FlowNet2 [20], unsupervised DSTFlow [34], UnFlow-C [23] and the joint learning method GeoNet [19]. By looking at the EPE of pixels over non-occluded area, we can see that our method beats most deep learning approaches including carefully designed supervised learning method FlowNet2, and comparable to UnFlow-C when trained on KITTI. As demonstrated in Figure 6 and Table 6, our system achieves significant improvement in non-occluded regions against other baselines due to the bidirectional flow occlusion mask and regularities extracted from different tasks.

By avoiding any improper constraint effecting on occluded regions, the predictions over these regions tend to be primitive and not refined. However, leveraging our 3D alignment loss and mutually supervised term, our method still performs comparable to other baselines in overall regions.



(**a**) Input (**b**) GT

(**c**) Yin et al. (**d**) Ours (**e**) Ours Error

Figure 6. Comparison of optical flow estimation between Yin et al. [19] and ours on KITTI Flow 2015 dataset. The ground truth is interpolated for visualization purpose. Our method performs better in both occluded and overall regions.

Table 6. The reported Average end-point error (AEE) on KITTI flow 2015 over all pixels (All) and over non-occluded pixels only (Noc). C denotes dataset FlyingChairs, S denotes Sintel and T is FlyingThings3D.

Method	Dataset	Noc	All
EpicFlow [33]	-	4.45	9.57
FlowNetS [14]	C+S	8.12	14.19
FlowNet2 [20]	C+T	4.93	10.06
DSTFlow [34]	K	6.96	16.79
UnFlow-C [23]	K	4.29	8.80
GeoNet (FlowNetS) [19]	K	6.77	12.21
GeoNet [19]	K	8.05	10.81
Ours	K	4.41	9.24

4.2.3. Pose Evaluation

During the training process, these three tasks are learned jointly and their accuracy is inter-dependent. For evaluation of estimated camera pose accuracy, we test our model on KITTI visual odometry split which contains 11 sequences with ground truth. To compare with prior works, we train our joint network on the sequences as the same setting as [15], that sequences 00–08 are for training and 09–10 are for testing. We also compare our method with a full version and a short version of traditional SLAM approach ORB-SLAM [8]. The input to the full version takes entire sequence while the short version takes 5 frames. We also employ another experiment where all the loss weights are naive. As shown in Table 7, our method outperforms all of the competing

baselines. Our occlusion-aware geometric constrained model is able to capture high-level and more reliable details.

Table 7. Absolute Trajectory Error (ATE) on KITTI odometry Dataset. The results of other baselines are taken from [15,18,19].

Method	Seq.09	Seq.10
ORB-SLAM (full)	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
Zhou [15]	0.021 ± 0.017	0.020 ± 0.015
Mahjourian [18]	0.013 ± 0.010	0.012 ± 0.011
Yin [19]	0.012 ± 0.007	0.012 ± 0.009
Ours (naive)	0.011 ± 0.007	0.011 ± 0.009
Ours	0.011 ± 0.006	0.010 ± 0.008

5. Conclusions

In this paper, we proposed a joint learning framework to discuss three basic vision tasks of the long-standing scene understanding problem in an unsupervised manner. We explicitly take the occlusion and 3D structure of the scene into consideration. By preserving extensive geometric cues to lead learning, we obtained impressive results which demonstrated how geometry can benefit deep learning significantly in geometric reasoning.

However, our method is still limited to certain type of scenes where there are not many dynamic motions. Moreover, lower computational burden and system complexity are in need of discovery. In the future, we would like to explore the modeling of various motion masks for a more dynamic environment. Enlarged search space for warping-based loss to solve gradient locality is in the process of exploration. Also, introducing semantic segmentation into our system can offer more advantages. Meanwhile, the potential multi-task architecture, which has shared representation and computation in learning both low-level and high-level vision tasks, could be promising.

Author Contributions: Methodology, Experimental analysis and Paper Writing, Q.T.; Writing-review and Data analysis, Q.T. and Y.C.; Data and Writing Correction, C.H. The work was done under the supervision and guidance of Y.C.

Funding: This work is partially supported by Shanghai International Cooperation Fund Project (No.12510708400) and Shanghai Innovation Action Plan Project (No.16511101200) of Science and Technology Committee of Shanghai Municipality.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yang, L.; Cheng, H.; Hao, J.; Ji, Y.; Kuang, Y. *A Survey on Media Interaction in Social Robotics*; Springer: Cham, Switzerland, 2015; pp. 181–190.
2. Chen, C.; Seff, A.; Kornhauser, A.; Xiao, J. DeepDriving: Learning Affordance for Direct Perception in Autonomous Driving. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; Volume 00, pp. 2722–2730.
3. Carmigniani, J.; Furht, B.; Anisetti, M.; Ceravolo, P.; Damiani, E.; Ivkovic, M. Augmented reality technologies, systems and applications. *Multimed. Tools Appl.* **2011**, *51*, 341–377. [CrossRef]
4. Torresani, L.; Hertzmann, A.; Bregler, C. Nonrigid Structure-from-Motion: Estimating Shape and Motion with Hierarchical Priors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 878–892. [CrossRef] [PubMed]
5. Wu, C. Towards Linear-Time Incremental Structure from Motion. In Proceedings of the International Conference on 3dtv-Conference, Seattle, WA, USA, 29 June–1 July 2013; pp. 127–134.
6. Agudo, A.; Morenonoguer, F.; Calvo, B.; Montiel, J.M. Sequential Non-Rigid Structure from Motion Using Physical Priors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 979–994. [CrossRef] [PubMed]
7. Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1052–1067. [CrossRef] [PubMed]

8. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
9. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
10. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2366–2374.
11. Godard, C.; Aodha, O.M.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 6602–6611.
12. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 66–75.
13. Brahmbhatt, S.; Gu, J.; Kim, K.; Hays, J.; Kautz, J. MapNet: Geometry-Aware Learning of Maps for Camera Localization. *arXiv* **2017**, arXiv:1712.03342.
14. Dosovitskiy, A.; Fischery, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Smagt, P.V.D.; Cremers, D.; Brox, T. FlowNet: Learning Optical Flow with Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
15. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised Learning of Depth and Ego-Motion from Video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6612–6619.
16. Ummenhofer, B.; Zhou, H.; Uhrig, J.; Mayer, N.; Ilg, E.; Dosovitskiy, A.; Brox, T. DeMoN: Depth and Motion Network for Learning Monocular Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5622–5631.
17. Vijayanarasimhan, S.; Ricco, S.; Schmid, C.; Sukthankar, R.; Fragkiadaki, K. SfM-Net: Learning of Structure and Motion from Video. *arXiv* **2017**, arXiv:1704.07804.
18. Mahjourian, R.; Wicke, M.; Angelova, A. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. *arXiv* **2018**, arXiv:1802.05522.
19. Yin, Z.; Shi, J. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. *arXiv* **2018**, arXiv:1803.02276.
20. Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1647–1655.
21. Garg, R.; Vijay, K.B.G.; Carneiro, G.; Reid, I. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 740–756.
22. Li, R.; Wang, S.; Long, Z.; Gu, D. UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning. *arXiv* **2017**, arXiv:1709.06841.
23. Meister, S.; Hur, J.; Roth, S. UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss. *arXiv* **2017**, arXiv:1711.07837.
24. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2017–2025.
25. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3061–3070.
26. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2015**, arXiv:1603.04467.
27. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2024–2039. [CrossRef] [PubMed]
28. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

29. Saxena, A.; Sun, M.; Ng, A.Y. Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 824–840. [CrossRef] [PubMed]
30. Karsch, K.; Liu, C.; Kang, S.B. Depth Transfer: Depth Extraction from Video Using Non-Parametric Sampling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *36*, 2144. [CrossRef] [PubMed]
31. Liu, M.; Salzmann, M.; He, X. Discrete-Continuous Depth Estimation from a Single Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 716–723.
32. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper Depth Prediction with Fully Convolutional Residual Networks. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
33. Revaud, J.; Weinzaepfel, P.; Harchaoui, Z.; Schmid, C. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7–12 June 2015; pp. 1164–1172.
34. Ren, Z.; Yan, J.; Ni, B.; Liu, B.; Yang, X.; Zha, H. Unsupervised Deep Learning for Optical Flow Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.