sensors

MDPI

Article

Unsupervised Stereo Matching with Surface Normal Assistance for Indoor Depth Estimation

Xiule Fan [1], Ali Jahani Amiri [2], Baris Fidan [1,*] and Soo Jeon [1]

[1] Department of Mechanical and Mechatronics Engineering, University of Waterloo, 200 University Ave. W., Waterloo, ON N2L 3G1, Canada; x54fan@uwaterloo.ca (X.F.); soojeon@uwaterloo.ca (S.J.)
[2] Avidbots Corp., 45 Washburn Dr., Kitchener, ON N2R 1S1, Canada; jahaniam@ualberta.ca
* Correspondence: fidan@uwaterloo.ca

Abstract: To obtain more accurate depth information with stereo cameras, various learning-based stereo-matching algorithms have been developed recently. These algorithms, however, are significantly affected by textureless regions in indoor applications. To address this problem, we propose a new deep-neural-network-based data-driven stereo-matching scheme that utilizes the surface normal. The proposed scheme includes a neural network and a two-stage training strategy. The neural network involves a feature-extraction module, a normal-estimation branch, and a disparity-estimation branch. The training processes of the feature-extraction module and the normal-estimation branch are supervised while the training of the disparity-estimation branch is performed unsupervised. Experimental results indicate that the proposed scheme is capable of estimating the surface normal accurately in textureless regions, leading to improvement in the disparity-estimation accuracy and stereo-matching quality in indoor applications involving such textureless regions.

Keywords: stereo matching; unsupervised learning; indoor applications; normal estimation

check for updates

Citation: Fan, X.; Amiri, A.J.; Fidan, B.; Jeon, S. Unsupervised Stereo Matching with Surface Normal Assistance for Indoor Depth Estimation. *Sensors* **2023**, *23*, 9850. https://doi.org/10.3390/s23249850

Academic Editor: Denis Laurendeau

Received: 4 November 2023
Revised: 4 December 2023
Accepted: 13 December 2023
Published: 15 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Stereo cameras have been widely used by robotic and other intelligent systems to obtain depth information. In such systems, a stereo camera captures a pair of stereo images, from which a stereo-matching algorithm computes the disparity that corresponds to the depth to be estimated. Hence, the accuracy of the stereo-matching algorithm directly affects the quality of the depth estimates.

In the past decades, various stereo-matching algorithms have been proposed. In the early attempts, traditional algorithms [1–4] were well-studied. Their estimated disparity maps often contain inaccurate or missing estimates. With the help of recent advances in computer hardware technologies as well as developments in deep neural network (DNN) learning, learning-based stereo-matching approaches [5–9] that are trained with large datasets have gained popularity. These data-driven approaches often provide more accurate and denser disparity maps than traditional algorithms do. However, most of these methods are evaluated on either synthetic datasets [10] or outdoor datasets [11,12] collected in driving scenarios.

Estimating depth in an indoor environment using data-driven approaches has been studied previously by adopting various monocular depth-estimation networks trained in a supervised [13–15] or unsupervised manner [16,17]. However, existing stereo counterparts are still limited to supervised learning [18] for indoor scenarios. Recently, surface normal has been incorporated into a supervised stereo-based indoor depth-estimation approach [19]. Although supervised approaches may result in high accuracy, obtaining the ground-truth depth labels required for training is a time-consuming and complex process. When the neural network is deployed in an unseen environment, fine tuning with new data is often necessary to maintain its accuracy. The possibility of missing ground-truth

information in such new data increases the difficulty of deploying supervised approaches and fine tuning the schemes developed via these approaches. Unsupervised monocular depth-estimation approaches do not rely on expensive ground-truth labels for training; however, they can only estimate depth maps that are up to scale, or otherwise additional information is needed in order to properly scale the estimated depth. The aforementioned challenges can alternatively be addressed by the adoption of an unsupervised stereo-matching approach, where the training does not require ground-truth information and the disparity estimation is not affected by scaling factors.

Unsupervised indoor depth perception is not a trivial task. Compared to outdoor driving scenarios, indoor environments typically consist of more textureless regions. The photometric loss, which is the main supervisory signal in unsupervised monocular and stereo depth estimation, is often ambiguous for these textureless regions [16,17]. Therefore, training the neural network with photometric loss for indoor applications often leads to sub-optimal performance. To reduce the ambiguity due to photometric loss, researchers have attempted to incorporate other information to obtain more reliable supervisory signals. In unsupervised indoor monocular depth estimation, optical flows [16] and superpixels extracted from the input RGB images [17] have been considered. However, this unsupervised strategy is yet to be extended to indoor stereo matching.

In this paper, we study surface normal estimation and its incorporation into unsupervised stereo-matching-based indoor depth estimation. Motivated by the supervised surface-normal-assisted stereo indoor depth-estimation approach that was recently proposed in [19], we design a novel unsupervised scheme consisting of a neural network with three modules and a two-stage training strategy for stereo-depth estimation in indoor environments. The scheme first uses a feature extractor to obtain high-level features from RGB stereo-image inputs and then estimates the surface normal by using the extracted high-level features through its normal-estimation branch. Using the high-level features and the estimated normal maps, the scheme's disparity-estimation branch generates the disparity estimates. We follow a two-stage strategy to train the DNNs within the proposed scheme in order to achieve unsupervised learning for indoor disparity estimation. First, the feature extractor and normal-estimation-branch DNNs are pre-trained in a supervised manner with the ground-truth surface normal from the NYU v2 dataset [20]. In the second stage after pre-training, we only train the disparity-estimation branch in an unsupervised manner with guidance from the estimated surface normal. The proposed scheme is tested for analysis and performance verification on the NYU v2 dataset for surface normal estimation and on the IRS dataset [18] and InStereo2K dataset [21] for disparity estimation.

The rest of the paper is organized as follows: Section 2 provides a literature review on computer-vision-based indoor depth estimation and the use of stereo matching and surface normal estimation for this purpose. Section 3 describes the overall structure of the neural network in our proposed scheme. Section 4 presents the proposed two-stage training strategy. Section 5 is dedicated to the implementation and evaluation of the proposed scheme using different datasets. The conclusion and final remarks are provided in Section 6.

2. Related Work

2.1. Surface Normal Estimation

Surface normal estimation has been an important research topic in the computer-vision-research community for more than a decade. In an early attempt, Fouhey et al. [22] designed a support vector machine (SVM) to estimate the surface normal. By grouping pixels according to their geometries and exploiting various cues, the surface normal can also be estimated given an image [23]. The method in [24] combines the information provided by image pixels and segments based on the input images for normal estimation.

Recently, various DNNs have been designed for surface normal estimation. Wang et al. [25] estimated a coarse global normal map and surface normal for image patches and then combined them with a fusion network. Eigen and Fergus [26] developed a multi-scale

DNN for multiple computer-vision tasks including surface normal estimation. The surface normal, depth, and information for planar regions predicted by a DNN from an input image are processed by a conditional random field to refine the predictions in [27]. A skip network architecture has also been adopted for normal estimation [28]. GeoNet [29] and its successor GeoNet++ [30] both estimate the depth and surface normal, which are used to refine each other to obtain better estimates. Zhang et al. [31] designed a multi-task network and studied the similarity between estimations at different pixel locations. Such a similarity helps diffuse the surface normal estimates to obtain better results. Liao et al. [32] adopted a spherical regression strategy by using DNN to predict the surface normal. The method introduced in [33] is capable of predicting the normal with a tilted image input. Bae et al. [34] proposed a neural network to first estimate a coarse normal map and its corresponding uncertainty, both of which are combined to form a refined normal map. The encoder–decoder network in [35] learns a discretized representation of high-level features from an input image to support depth estimation and surface normal estimation. Instead of estimating the surface normal from a single image, photometric stereo is another approach that performs normal estimation based on images of the same object in different lighting conditions. Under this formulation, the attention mechanism is used in [36] to estimate a more accurate surface normal of an object with fewer input images. Ju et al. [37] estimated high-resolution normal maps with low-resolution input images.

2.2. Stereo Matching

Stereo matching typically consists of four stages [38]: matching cost computation, cost aggregation, disparity computation based on optimization, and disparity refinement. Different traditional stereo-matching algorithms have been proposed by following these steps. These algorithms can be categorized into more efficient local methods [1] and more accurate global methods [2] at the cost of more expensive global optimization. By approximating global optimization in multiple local regions, semi-global methods [3,4] provide a tradeoff between accuracy and efficiency.

The advancement in deep learning has introduced data-driven solutions to the stereo-matching problem, which often lead to better performance. The first attempt in deep stereo matching [39] utilizes a DNN to extract image features, which are then processed by using a traditional method to obtain the estimated disparity. The first end-to-end stereo-matching network was proposed in [5]. Spatial pyramid pooling is adopted in [6] to address ambiguous regions in stereo matching. Redesigning the cost-aggregation module in the neural network also improves the accuracy significantly [7]. Cheng et al. [8] utilized a neural architecture search to identify a design that leads to high-quality results. The attention mechanism and transformer architecture have also been adopted in deep stereo matching [40,41]. Li et al. [42] addressed stereo matching under non-ideal conditions, such as thin structures in the scene and inaccurate image rectification. In addition to accuracy in stereo matching, some other approaches were designed to achieve high-quality estimates in real time by eliminating the stereo-matching cost volume [43] or by performing cost aggregation for inference with 2D convolutions only [44].

Besides supervised stereo matching, research on unsupervised stereo-matching solutions is also popular since they do not depend on additional ground-truth-disparity labels. Unsupervised stereo matching was first studied in [45] by only using confidential regions in the stereo images as inputs. Li and Yuan [46] designed a two-part unsupervised neural network, which estimates an occlusion mask first and then computes disparity in an occlusion-aware manner. Liu et al. [47] explored the use of stereo images captured at different time steps to train their unsupervised neural network. Wang et al. [9] incorporated the recent development in the attention mechanism into their design. A spatially adaptive self-similarity module is introduced in [48] to solve unsupervised stereo matching by using left and right stereo images with different visual properties.

2.3. Indoor Depth Estimation

Indoor depth estimation has been studied in both monocular and stereo settings. With the indoor NYU v2 dataset [20], Eigen et al. [13] designed a two-stage neural network to predict a coarse and fine depth map with a monocular RGB input. Other researchers explored different architectures, including conditional random fields [49], random forests [50], adversarial networks [51], and vision transformers (ViT) [15], to improve the estimated depth. Wofk et al. [14] designed a lightweight monocular depth-estimation approach to perform inferences on embedded systems. In addition to the aforementioned supervised monocular methods, unsupervised indoor monocular depth estimation was also studied. Zhou et al. [16] proposed to predict optical flows from temporally consecutive frames captured indoors and use these flows as additional supervisory signals for unsupervised indoor monocular depth estimation. Yu et al. [17] first extracted superpixels from the RGB image and then enforced the planar consistency between the predicted depth map and the superpixels.

In the stereo setting, Kusupati et al. [19] regressed a depth map and surface normal from stereo inputs. Apart from the difference between the ground truth and estimated values, the consistency between the estimated depth and surface normal is also enforced as a training signal. To address the lack of large datasets with stereo images and ground-truth disparity in indoor scenes, a synthetic indoor stereo dataset with 100k frames is proposed in [18]. A smaller but real dataset is also published in [21].

3. Proposed Neural Network Design

The proposed neural network architecture as shown in Figure 1 consists of three modules: the feature extractor, normal-estimation branch, and disparity-estimation branch. These modules of the proposed scheme can be trained and evaluated in two different modes. In the normal-estimation mode, the feature extractor and normal-estimation branch are used together to produce a surface normal map from an input image. In the disparity-estimation mode, the feature extractor receives stereo images and computes two sets of image features. When training the neural network in the disparity-estimation mode, we use the normal-estimation branch to estimate two surface normal maps by using each set of image features. The disparity-estimation branch then estimates both the left and right disparity maps given the image features and surface normal maps. However, in the evaluation stage, only the left image features are processed by the normal-estimation branch to obtain the left normal map. Using the left and right image features and the left surface normal, the disparity-estimation branch then estimates the left disparity map.

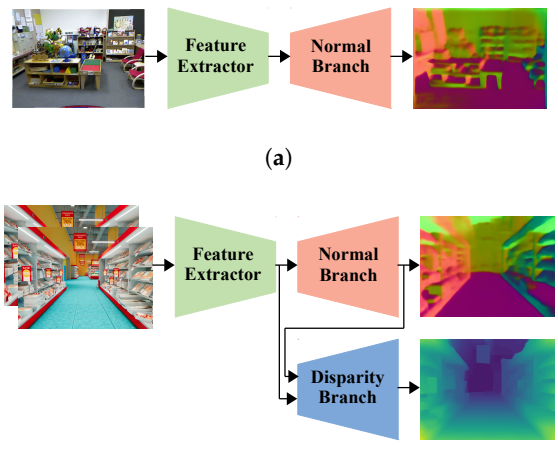


Figure 1. Overview of our proposed approaches for (**a**) normal estimation and (**b**) disparity estimation.

3.1. Feature Extraction

The feature extractor is used to downsample the input images and extract a set of high-level features $\{\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\}$. Its design is inspired by ResNet-50 [52] with three stages, as shown in Figure 2a. We denote the input feature at each stage as $\mathbf{F}_i^{'} \in \mathbb{R}^{H/2^i \times W/2^i \times C_i}$, where H and W are the height and width of the input image, respectively; $i \in \{0, 1, 2\}$; and C_i denotes the number of channels. $\mathbf{F}_i^{'}$ is downsampled by a 5×5 convolutional layer with a stride of two, padding of two, batch normalization, and leaky ReLU activation. The output from this layer has half of the spatial resolution compared to $\mathbf{F}_i^{'}$ and a higher number of channels C_{i+1}. This output is then processed by a series of 3×3 residual layers with leaky ReLU to obtain an intermediate feature $\mathbf{F}_{i+1}^{'} \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times C_{i+1}}$. The output feature $\mathbf{F}_{i+1} \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times C_{i+1}}$ from this stage is computed by applying a 3×3 convolution to $\mathbf{F}_{i+1}^{'}$ without an activation function. In the first stage, we set $\mathbf{F}_0^{'} = \mathbf{F}_0 = \mathbf{I}$, where $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ denotes the input image. C_i is set as 3, 32, 64, and 128 for $i = 0, 1, 2, 3$, respectively.

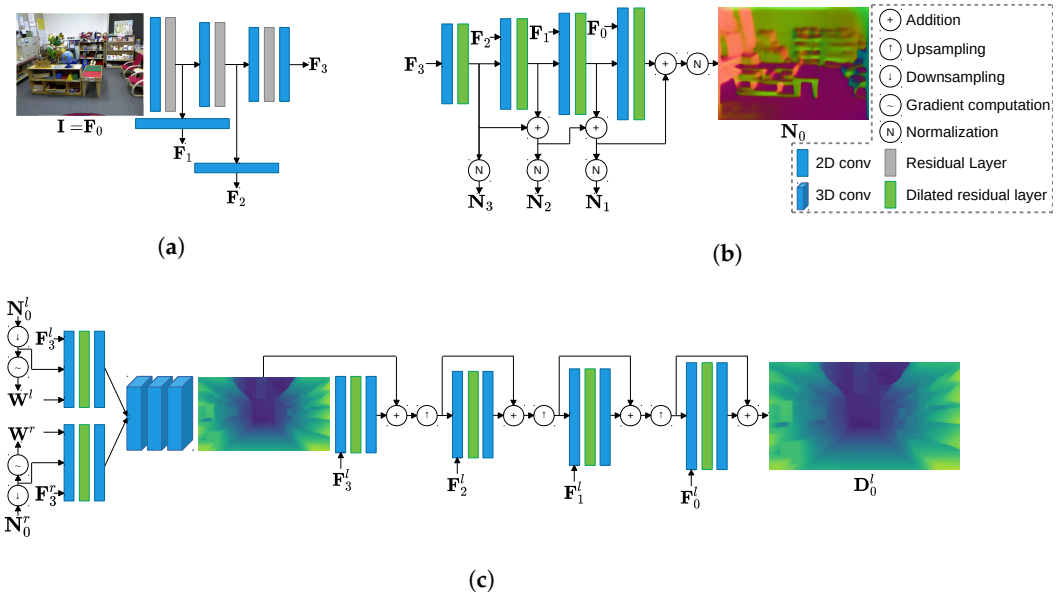


Figure 2. Schematics of different modules in the proposed neural network: (**a**) feature extraction, (**b**) normal-estimation branch, and (**c**) disparity-estimation branch.

In the normal-estimation mode, we apply this module to one input image $\mathbf{I}$ to obtain $\{\mathbf{F}_0, \mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3\}$. In the disparity mode, two sets of image features $\left\{\mathbf{F}_0^l, \mathbf{F}_1^l, \mathbf{F}_2^l, \mathbf{F}_3^l\right\}$ and $\left\{\mathbf{F}_0^r, \mathbf{F}_1^r, \mathbf{F}_2^r, \mathbf{F}_3^r\right\}$ are extracted based on the left and right stereo images, $\mathbf{I}^l \in \mathbb{R}^{H \times W \times 3}$ and $\mathbf{I}^r \in \mathbb{R}^{H \times W \times 3}$, respectively.

3.2. Normal-Estimation Branch

After obtaining the high-level image features, we use our proposed modular normal-estimation branch shown in Figure 2b to estimate the surface normal. The normal-estimation branch gradually upsamples the estimated normal maps. Additionally, instead of estimating the surface normal at a higher resolution in each stage, our normal-estimation branch is inspired by a previous stereo-matching network [53] to estimate the surface normal residual at a higher resolution.

At stage i of the normal-estimation branch, the image feature $\mathbf{F}_i$ and an unnormalized surface normal $\mathbf{N}_{i+1}^{'} \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1} \times 3}$ from the previous stage $i+1$ of this branch are used as the inputs. $\mathbf{N}_{i+1}^{'}$ is first bilinearly upsampled to match the resolution of $\mathbf{F}_i$ and then concatenated with $\mathbf{F}_i$ along the channel dimension to form a feature volume. The

feature volume is processed by six 3×3 residual blocks with the leaky ReLU activation function while maintaining the same resolution and number of channels. The residual blocks are designed with dilation factors 1, 2, 4, 8, 1, and 1. Next, a 3×3 convolution with no activation functions is applied to the feature volume to compute the surface normal residual $\Delta\mathbf{N}_i \in \mathbb{R}^{H/2^i \times W/2^i \times 3}$. $\Delta\mathbf{N}_i$ is then added to the upsampled $\mathbf{N}'_{i+1}$ to compute the unnormalized normal $\mathbf{N}'_i \in \mathbb{R}^{H/2^i \times W/2^i \times 3}$. $\mathbf{N}'_i$ is used in the next stage of estimation and normalized to $\mathbf{N}_i \in \mathbb{R}^{H/2^i \times W/2^i \times 3}$ as the output of stage i.

There are four stages in the normal-estimation branch in total. To start the normal-estimation process, the upsampling and concatenation steps in stage 3 are neglected. Furthermore, since there is no estimated surface normal at the beginning of this stage, we only use $\mathbf{F}_3$ as the input and process it with the dilated residual blocks directly. After four stages of computation, the outputs of the normal-estimation branch include $\{\mathbf{N}_3, \mathbf{N}_2, \mathbf{N}_1, \mathbf{N}_0\}$. $\mathbf{N}_0$ is considered the final output of the normal-estimation branch.

3.3. Disparity-Estimation Branch

The design of the disparity-estimation branch, as shown in Figure 2c, follows the general architecture adopted by existing data-driven stereo-matching methods [5,6,9,53]. This architecture includes the matching cost construction, cost aggregation, and disparity refinement. To exploit the benefit of the estimated surface normal, we propose an additional normal integration component to combine the surface normal with the matching cost. To introduce our design, we only consider the left stereo view and all estimations derived from this view as examples, unless otherwise stated. The same components can be applied to the right view easily.

3.3.1. Normal Integration

In order to integrate the surface normal information, we treat it as additional features that can be combined with the high-level image features extracted from the feature-extraction module. From the normal-estimation branch, we can obtain the surface normal maps $\mathbf{N}_0^l \in \mathbb{R}^{H \times W \times 3}$ and $\mathbf{N}_0^r \in \mathbb{R}^{H \times W \times 3}$ for the left and right stereo images, respectively. Using the left view as an example, we first downsample $\mathbf{N}_0^l$ to $\mathbf{N}_{0\to3}^l \in \mathbb{R}^{H/8 \times W/8 \times 3}$ with nearest sampling so that its spatial resolution matches that of $\mathbf{F}_3^l$.

From our experiments, we also observe that the estimated surface normal is generally more accurate in regions with smooth estimates than in areas with rapid changes in the surface normal. Integrating inaccurate surface normal information into the matching cost may introduce negative effects in stereo matching. Therefore, it is important that the neural network focuses on accurate normal estimates and ignores the inaccurate ones. To achieve this goal, we propose a weighting mask based on our observation. This weighting mask places higher weights at smooth regions and lower weights when the surface normal changes significantly. In image processing, the Laplacian filter is commonly used to capture edges or intensity changes, which means it can also be used to identify image patches with minimal variations. By using this filter, the weighting mask that we design is:

$$\mathbf{W}^l = \exp\left(-\lambda_w \sum_{j=1}^{3} \left| \nabla^2 \mathbf{N}_{0\to3}^l(\cdot, \cdot, j) \right| \right) \in \mathbb{R}^{H \times W}, \tag{1}$$

where $\lambda_w = 5$ is a constant to control the sensitivity and ∇^2 denotes a 3×3 Laplacian filter. The resulting values from the Laplacian filters have lower magnitudes at regions with a smoother surface normal. To remove the ambiguity introduced by signs, we consider the absolute value of these resulting features. Then, we perform summation along the channel dimension to combine the surface normal smoothness in different directions. Lastly, the exponential function constrains the weighting mask to be between 0 and 1.

After obtaining the downsampled estimated surface normal and weighting mask, we concatenate $\mathbf{F}_3^l$, $\mathbf{N}_{0\to3}^l$, and $\mathbf{W}^l$ along the channel dimension and process this volume by a

3×3 convolution followed by batch normalization and leaky ReLU activation to change its number of channels to 256. Then, we apply dilated residual blocks, which follow the same design as introduced in Section 3.2, to balance the values in the combined feature while maintaining the same spatial resolution and number of channels. Lastly, another 3×3 convolution without batch normalization or an activation function computes the output volume $\mathbf{F}_3^{\prime l} \in \mathbb{R}^{H/8 \times W/8 \times 256}$ from this component. This volume contains both information obtained directly from the input image and the estimated surface normal. Building a stereo-matching cost with this volume allows us to take advantage of accurate normal estimates.

3.3.2. Matching Cost Construction

From the left and right combined features $\mathbf{F}_3^{\prime l}$ and $\mathbf{F}_3^{\prime r}$, we construct a stereo-matching cost volume by considering one of them as the reference feature, while the other feature is considered the target feature. The difference between the reference feature and the target feature that shifted according to all disparity candidates is computed as the cost volume [53]. If we assume that the number of disparity candidates at the original image resolution is D, there are $d = D/8$ candidates at the lowest image resolution. When using $\mathbf{F}_3^{\prime l}$ as the reference feature, we obtain a left matching cost $\mathbf{C}^l \in \mathbb{R}^{H/8 \times W/8 \times 256 \times d}$.

3.3.3. Cost Aggregation

To enable more robust stereo matching, we perform cost aggregation on the matching costs. Cost aggregation in a data-driven stereo-matching approach is achieved by applying 3D convolutions to the cost volume along the spatial and disparity dimensions [5,6,53]. We follow [53] to design a lightweight cost-aggregation module with five 3D $3 \times 3 \times 3$ convolutional layers. The first four 3D convolutions are followed by batch normalization and leaky ReLU activation. They also maintain the number of channels for the cost volume at 256. The last convolution reduces the channel number to one to obtain an aggregated cost, from which a left initial disparity $\mathbf{D}_{init}^l \in \mathbb{R}^{H/8 \times W/8}$ is regressed through the differentiable soft argmin introduced in [5].

3.3.4. Disparity Refinement

Although the cost-aggregation module can compute an initial disparity map, $\mathbf{D}_{init}^l$ may not include detailed estimates. To remedy this problem, we design a disparity-refinement module to gradually upsample $\mathbf{D}_{init}^l$ while introducing more details. Similar to the normal-estimation branch, the refinement module adopts a modular design with multiple stages.

The inputs of stage i include the refined disparity from the previous refinement stage $\mathbf{D}_{i+1}^l \in \mathbb{R}^{H/2^{i+1} \times W/2^{i+1}}$ and the left high-level feature $\mathbf{F}_i^l$, while its output is the refined disparity map at a higher resolution $\mathbf{D}_i^l \in \mathbb{R}^{H/2^i \times W/2^i}$. In this refinement stage, $\mathbf{D}_{i+1}^l$ is first bilinearly upsampled to match the resolution of $\mathbf{F}_i^l$. The upsampled disparity and $\mathbf{F}_i^l$ are then concatenated and processed by a 3×3 convolution without batch normalization or activation functions to reduce its channel number to 32. Dilated residual blocks as described in Section 3.2 are also applied to this volume. Following the residual blocks, the volume undergoes another 3×3 convolution with no batch normalization or activation functions, resulting in a disparity residual. The disparity residual is added to the upsampled disparity. After addition, this refined disparity map passes through a ReLU activation function to obtain a $\mathbf{D}_i^l$ whose values are all non-negative.

Similar to the normal-estimation branch, the refinement module also includes four stages. At the first stage of refinement, which is stage 3, the upsampling step is neglected and the upsampled disparity is replaced by $\mathbf{D}_{init}^l$. $\mathbf{D}_0^l$ at the original image resolution is used as the final output of the disparity-estimation branch.

4. Training Strategy

4.1. Training for Normal Estimation

In the normal mode, the neural network is trained in a supervised manner. The supervised learning of surface normal estimation commonly relies on either the cosine similarity loss [26,37] or the L2 loss [30,33,36]. We adopt the latter alternative since it yields a better performance. With the set of estimated surface normal maps $\{\mathbf{N}_3, \mathbf{N}_2, \mathbf{N}_1, \mathbf{N}_0\}$ from an input image, the supervised loss is

$$\mathcal{L}_n = \sum_{i=0}^{3} \left(\frac{1}{2^i HW} \sum_{\mathbf{p}} \| \mathbf{N}_{i\to 0}(\mathbf{p}) - \mathbf{N}^*(\mathbf{p}) \|_2 \right), \tag{2}$$

where $\mathbf{N}_{i\to 0}$ denotes the estimated surface normal $\mathbf{N}_i$ bilinearly upsampled to the same resolution as the input image, $\mathbf{N}^*$ denotes the ground-truth normal, and $\mathbf{p}$ denotes an arbitrary pixel. The weighting term $1/2^i$ enforces the training loss to focus on estimates at higher image resolutions. Note that only the feature extractor and normal-estimation branch are utilized to estimate the surface normal. Hence, only the parameters in these two modules are updated with (2).

4.2. Training for Disparity Estimation

After the neural network obtains preliminary knowledge on surface normal estimation, we further train it for disparity estimation in a fully unsupervised manner. In this stage of training, the parameters of the feature extractor and surface normal-estimation branch are frozen. Therefore, back propagation is only allowed in the disparity-estimation branch. This training stage involves multiple training losses whose definitions are given below by using the left view as an example. By applying a similar formulation, these losses can be expanded to the right view.

4.2.1. Photometric Loss

The photometric loss quantifies the differences between one stereo image and a reconstructed image based on the other stereo view and disparity. If the disparity is accurate, the stereo image and the reconstructed view are visually similar. Hence, the photometric loss will be close to zero. The photometric loss of a left-view pixel is defined as

$$\mathcal{L}^l_{ph,i}(\mathbf{p}) = \frac{\alpha}{2}\left(1 - SSIM\left(\mathbf{I}^l(\mathbf{p}), \hat{\mathbf{I}}^l_i(\mathbf{p})\right)\right) + (1-\alpha)\left\|\mathbf{I}^l(\mathbf{p}) - \hat{\mathbf{I}}^l_i(\mathbf{p})\right\|, \tag{3}$$

where $\alpha = 0.85$ and $SSIM(\cdot)$ denotes the structural similarity index measure [54]. $\hat{\mathbf{I}}^l_i \in \mathbb{R}^{H\times W\times 3}$ is a bilinearly reconstructed image according to the right stereo view $\mathbf{I}^r$ and a disparity map $\mathbf{D}^l_{i\to 0} \in \mathbb{R}^{H\times W}$, which is bilinearly upsampled from the estimated left disparity map $\mathbf{D}^l_i$ at refinement stage i.

4.2.2. Disparity Smoothness Loss

To prevent the neural network from estimating noisy disparity maps, a disparity smoothness loss is widely used to regularize the estimates. This smoothness loss is given as

$$\mathcal{L}^l_{ds,i}(\mathbf{p}) = \left|\nabla_x \mathbf{D}^l_{i\to 0}(\mathbf{p})\right| e^{-\left\|\nabla_x \mathbf{I}^l(\mathbf{p})\right\|} + \left|\nabla_y \mathbf{D}^l_{i\to 0}(\mathbf{p})\right| e^{-\left\|\nabla_y \mathbf{I}^l(\mathbf{p})\right\|}, \tag{4}$$

where ∇_x and ∇_y are the gradients of an image with respect to the horizontal and vertical direction, respectively. The gradients in (4) emphasize disparity smoothness at textureless regions since these regions are more likely to exhibit smooth disparity.

4.2.3. Normal Consistency Loss

In addition to the photometric and disparity smoothness losses, we further exploit the consistency between the estimated normal and disparity to improve estimation at ambiguous regions. The normal consistency loss is defined as

$$\mathcal{L}^l_{n,i}(\mathbf{p}) = \mathbf{W}^l_{i\to 0}(\mathbf{p}) \left\| \mathbf{N}^l_{i\to 0}(\mathbf{p}) - \mathbf{N}^l_{D,i\to 0}(\mathbf{p}) \right\|_2, \tag{5}$$

where $\mathbf{N}^l_{D,i\to 0} \in \mathbb{R}^{H\times W\times 3}$ denotes the surface normal converted from the upsampled disparity map $\mathbf{D}^l_{i\to 0}$ according to [18], and the weight $\mathbf{W}^l_{i\to 0} \in \mathbb{R}^{H\times W}$ is obtained by applying (1) to the upsampled left estimated surface normal map $\mathbf{N}^l_{i\to 0}$. The weight can constrain the normal consistency loss at smoother regions, which usually contain more accurate normal estimates.

4.2.4. Left–Right Consistency Loss

To address occlusion, which is a common problem in stereo matching, a left–right consistency loss is used. This loss is given as

$$\mathcal{L}^l_{lr,i}(\mathbf{p}) = \left| \mathbf{D}^l_{i\to 0}(\mathbf{p}) - \hat{\mathbf{D}}^l_{i\to 0}(\mathbf{p}) \right|, \tag{6}$$

where $\hat{\mathbf{D}}^l_{i\to 0} \in \mathbb{R}^{H\times W}$ is a reconstructed left disparity map by bilinearly sampling the upsampled right disparity map $\mathbf{D}^r_{i\to 0}$ according to the upsampled left disparity map $\mathbf{D}^l_{i\to 0}$.

Moreover, since our network can estimate multi-scale disparity and normal maps, we utilize estimates at all scales to train the disparity-estimation branch. The combined training loss based on left and right estimates at scale i is

$$\begin{aligned} \mathcal{L}_{d,i} = & \sum_{\mathbf{p}} \alpha_{ph} \left(\mathcal{L}^l_{ph,i}(\mathbf{p}) + \mathcal{L}^r_{ph,i}(\mathbf{p}) \right) + \alpha_{ds} \left(\mathcal{L}^l_{ds,i}(\mathbf{p}) + \mathcal{L}^r_{ds,i}(\mathbf{p}) \right) \\ & + \alpha_n \left(\mathcal{L}^l_{n,i}(\mathbf{p}) + \mathcal{L}^r_{n,i}(\mathbf{p}) \right) + \alpha_{lr} \left(\mathcal{L}^l_{lr,i}(\mathbf{p}) + \mathcal{L}^r_{lr,i}(\mathbf{p}) \right), \end{aligned} \tag{7}$$

where the superscript r denotes that the losses are based on the right-view images, and the α's are the weights for different terms. By collecting the training losses at all scales, the final loss for disparity training is

$$\mathcal{L}_d = \frac{1}{4HW} \sum_{i=0}^{3} \left(\frac{1}{2^i} \mathcal{L}_{d,i} \right). \tag{8}$$

5. Experimental Results

5.1. Implementation Details

We train and evaluate our proposed scheme on multiple datasets for normal and disparity estimations. For normal estimation, we apply our design to the NYU v2 dataset [20]. The availability of large public datasets with indoor stereo images and ground-truth disparity is limited. Therefore, we train our network by using the large synthetic IRS dataset [18] for indoor stereo matching. The IRS dataset consists of images rendered in both bright and dark lighting conditions. Since low-light scenarios are out of the scope of this study, we only include images rendered in normal lighting in training and evaluation. To evaluate our method's generalization ability, we further test it with a smaller real indoor dataset, InStereo2K [21].

In both training stages, the neural network is trained by using an Adam optimizer. Data augmentation is applied to all training images by randomly modifying their brightness, contrast, saturation, and hue. All images are normalized by the ImageNet mean and variance. During training for normal estimation, the images are randomly cropped to a resolution of 416 × 552. The neural network is then trained by using data from the NYU v2 dataset with a batch size of eight for 20 epochs. The initial learning rate in the first stage is

0.001. This learning rate is later reduced by half at the 10th epoch. After training for normal estimation is completed, the disparity-estimation branch is fine tuned on the IRS dataset for another 20 epochs with a batch size of four. The initial learning rate is 0.0001, which is multiplied by 0.1 at the 10th epoch. The input images are randomly cropped to a resolution of 256 × 512. The constants chosen for (7) are $\alpha_{ph} = 5$, $\alpha_{ds} = 0.05$, $\alpha_n = 0.5$, and $\alpha_{lr} = 0.01$. The negative slope of all leaky ReLU activation functions is chosen as 0.2.

5.2. NYU v2 Dataset

We compare the performance of our approach on normal estimation with existing methods on the NYU v2 [20] test set. We report the performance by using error and accuracy metrics, both of which are based on the angular difference between the estimated and ground-truth normal vectors. The error metrics include the mean error, median error, and root mean squared error (RMSE) of the angular differences at all pixel locations. The accuracy metrics are the percentages of pixels with angular differences lower than 11.25°, 22.5°, and 30°, respectively.

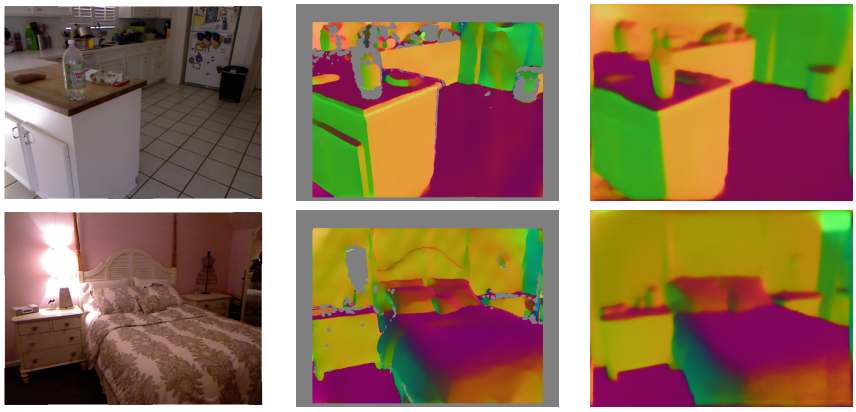
The quantitative results are summarized in Table 1. Although the main focus of our work is indoor stereo matching instead of surface normal estimation, our feature extractor and normal-estimation branch can still compute accurate surface normal estimates. Compared to the majority of the existing methods in Table 1, our approach achieves a lower error and higher accuracy. The performance of our method only falls behind that of [34,35] even though we did not specifically tune our neural network or utilize an intricate design tailored for surface normal estimation. Although it is possible to utilize the surface normal estimated from [34,35] to guide the downstream disparity-estimation process, this approach will significantly increase the complexity. For instance, we can no longer use the same feature-extraction module for both tasks, which implies a possible higher memory footprint and computational power requirement to train and use the entire architecture. Moreover, the prediction step in [34] relies on both the estimated normal and an uncertainty map, which introduces additional complexity compared to our approach. In [35], an extra internal discretization module is needed in addition to the regular encoder–decoder design. On the other hand, our method offers a simplistic solution to surface normal estimation while maintaining high accuracy. Since the main goal of our proposed scheme is disparity estimation and surface normal estimation only serves as a support role to our main goal, it is important to keep the surface-normal-estimation solution simple to avoid adding unnecessary overhead to the overall scheme.

Table 1. Error and accuracy metrics of different surface-normal-estimation methods on the NYU v2 test set.

Method	Error (°)			Accuracy (%)		
	Mean	Median	RMSE	11.25°	22.5°	30°
Wang et al. [25]	26.9	14.8	-	42.0	61.2	68.2
Multi-scale [26]	20.9	13.2	-	44.4	67.2	75.9
SURGE [27]	20.6	12.2	-	47.3	68.9	76.6
Bansal et al. [28]	19.8	12.0	28.2	47.9	70.0	77.8
GeoNet [29]	19.0	11.8	26.9	48.4	71.5	79.5
Nekrasov et al. [55]	24.0	17.7	-	-	-	-
PAP [31]	18.6	11.7	25.5	48.8	72.2	79.8
Liao et al. [32]	19.7	12.5	-	45.8	72.1	80.6
Bae et al. [34]	14.9	7.5	23.5	62.2	79.3	85.2
GeoNet++ [30]	18.5	11.2	26.7	50.2	73.2	80.7
iDisc [35]	14.6	7.3	22.8	63.8	79.8	85.6
Ours	17.8	11.3	25.4	52.1	74.2	81.6

In addition to the quantitative comparison, we present some qualitative results obtained by our approach in Figure 3. According to these results, our method can compute high-quality surface normal estimates, especially at smooth and often textureless regions

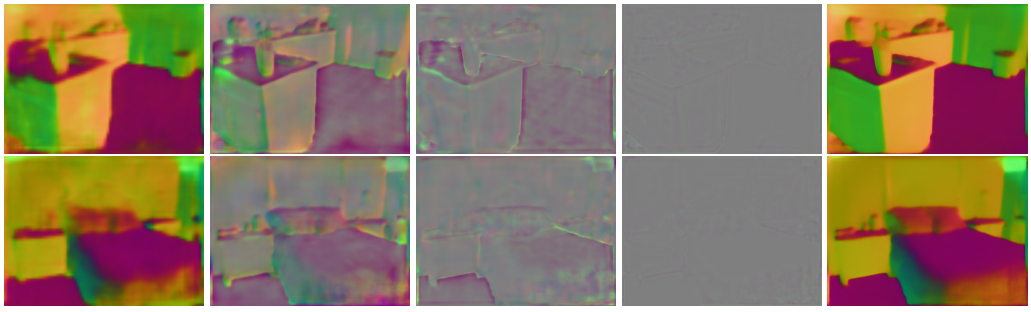
that are commonly seen in indoor environments. Examples of these regions can be found in Figure 3 in the top image on the counter and in the bottom image on the wall and floor areas. These regions often lead to ambiguous results in unsupervised stereo matching. This observation suggests that our estimated normal may contain useful information to address ambiguity in unsupervised indoor disparity estimation.



(**a**) (**b**) (**c**)

Figure 3. Sample qualitative results of surface normal estimation by our approach using the NYU v2 test set: (**a**) input RGB images, (**b**) ground-truth surface normal, (**c**) estimated surface normal.

One unique design of our normal-estimation branch is its ability to estimate surface normal residuals to refine normal estimates from the previous stage. We provide visualization in Figure 4 to demonstrate these residuals. It can be seen that the surface normal residuals recover a substantial amount of missing information at an earlier stage (e.g., stage 2) of the normal-estimation branch, especially at large flat regions. At stage 0, which is close to the end of the network, the residuals only need to correct the normal estimates at object boundaries.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 4. Visualization of the estimated normal residuals from the normal-estimation branch: (**a**) initial estimated surface normal at stage 3 of the normal-estimation branch, (**b**–**d**): surface normal residuals obtained at stage 2 to 0, (**e**) final estimated normal.

5.3. IRS Dataset

The evaluation of disparity estimation is first performed on the IRS dataset's [18] test set after training the neural network for stereo matching. Since indoor stereo matching is a less-explored topic, existing work on this topic is limited. Based on the availability of existing work and open-source code, we select FADNet [18], GwcNet [56], and PASMnet [9] for comparison. The first two approaches are supervised methods, while the last one is based on unsupervised training. The quantitative comparison is outlined in Table 2 based on two metrics: the endpoint error (EPE) and percentage of pixels with an error of more than 3 px (>3 px). The former metric quantifies the error, while the latter quantifies the

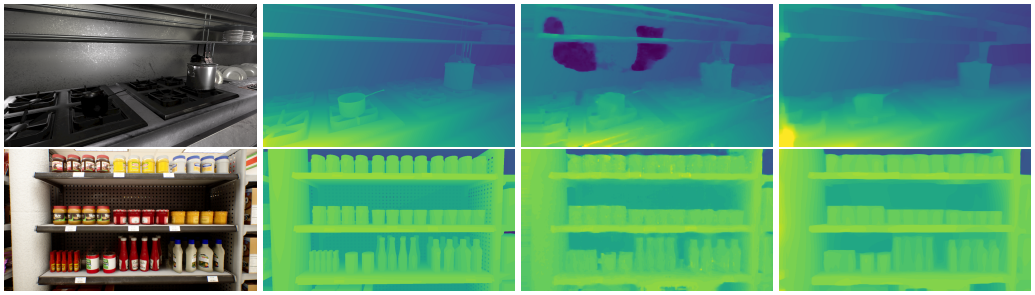
accuracy of different approaches. Furthermore, we compute these two metrics in two scenarios: using all the pixels in the images (EPE-a and >3 px-a) and using textureless pixels in the images (EPE-t and >3 px-t). To extract the textureless pixels, we first apply an 11×11 Laplacian filter to the input RGB images. After calculating the absolute value and summation across the channel dimension, we label a pixel as textureless if its resulting value is less than or equal to one.

Table 2. Quantitative comparison for disparity estimation on the test set of the IRS dataset in terms of error and accuracy. Different methods are separated into supervised (Sup.) methods and unsupervised (Unsup.) ones.

Method	Training	EPE-a (px)	>3 px-a (%)	EPE-t (px)	>3 px-t (%)
FADNet [18]	Sup.	0.75	-	-	-
GwcNet [56]	Sup.	3.01	-	-	-
PASMnet [9]	Unsup.	2.91	15.08	2.97	14.96
Ours	Unsup.	2.89	14.33	2.92	14.43

Among all four methods included in Table 2, the supervised FADNet [18] achieves the lowest EPE. Even though our approach is an unsupervised method, it still outperforms another supervised method [56] with a lower error. Compared to the recent open-source unsupervised stereo-matching approach [9], our method estimates disparity with a lower EPE and fewer outliers (>3 px). Our approach results in a decrease in the EPE and in >3 px by 0.69% and 4.97%, respectively, when all the pixels are considered in comparison with [9]. At textureless regions, the EPE and >3 px are lower by 1.68% and 3.54%, respectively. These results demonstrate that our approach is effective at estimating more accurate disparity at textureless regions, especially in terms of the EPE.

Apart from the quantitative results, we present sample qualitative results in Figure 5. The qualitative results demonstrate that the estimated disparity using our method is significantly better at planar and textureless regions than the estimates from [9]. This observation is supported by the back of the stove shown in the top image of Figure 5. In this example, PASMnet fails to understand that the gray wall at the back is a planar region and computes holes in the estimated disparity map. Our method successfully estimates a smooth disparity transition in that area to represent a plane.



(**a**) (**b**) (**c**) (**d**)

Figure 5. Sample qualitative results of disparity estimation on the IRS dataset test set: (**a**) input RGB images, (**b**) ground-truth disparity, (**c**) estimated disparity using [9], and (**d**) estimated disparity from our method.

5.4. InStereo2K Dataset

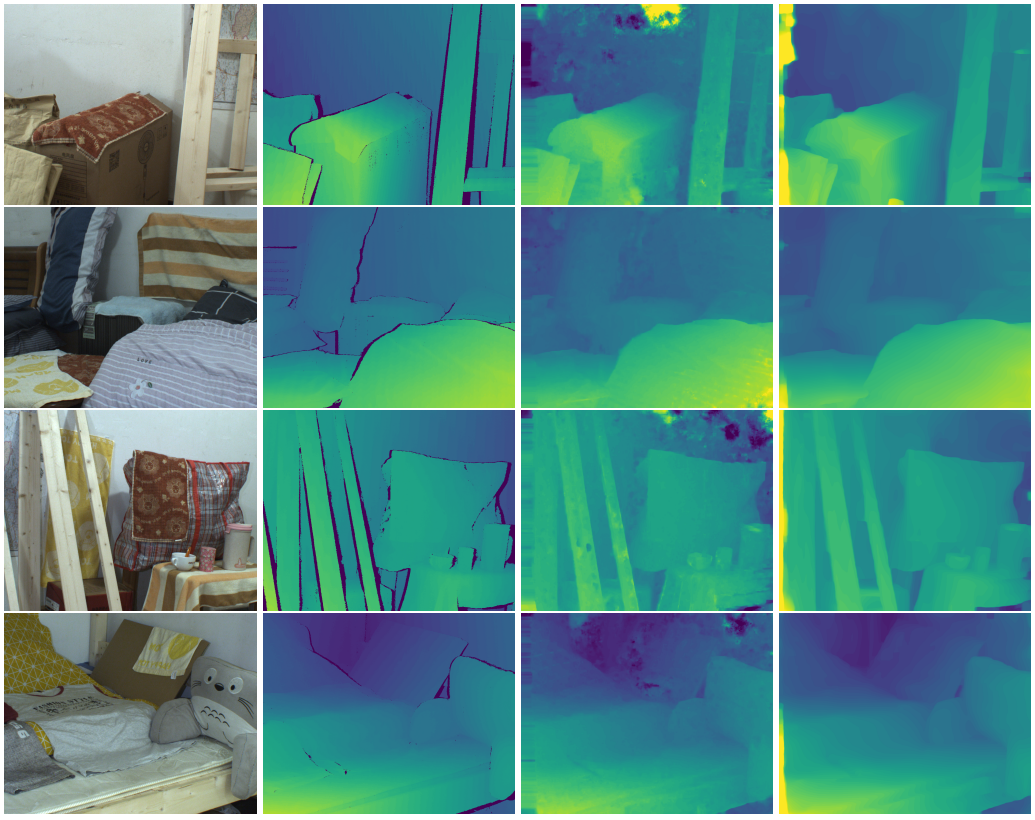
To further evaluate the performance of our approach in indoor stereo matching, we further study its generalization ability. This study is completed by performing inference directly on the InStereo2K [21] test set by using our method and [9], both of which are only trained on the IRS dataset for stereo matching. The quantitative results are shown in Table 3.

It can be seen that our method outperforms [9] with a lower error and percentage of outliers using both all pixels and textureless pixels. The results indicate that our approach improves the EPE and >3 px by 7.35% and 14.23%, respectively, when using all the pixels, as well as 5.90% and 13.04% in the EPE and >3 px, respectively, when considering the textureless pixels. These results demonstrate the better generalization ability of our approach.

Table 3. Quantitative results for disparity estimation on InStereo2K test set.

Method	EPE-a (px)	>3 px-a (%)	EPE-t (px)	>3 px-t (%)	FPS
PASMnet [9]	3.13	15.18	3.90	19.32	7.20
Ours	2.90	13.02	3.67	16.80	14.01

The qualitative results from Figure 6 show that both methods have difficulties estimating accurate disparity at the leftmost occluded areas, which are generally challenging to estimate correctly. However, the estimates at textureless and planar regions using our method are smoother and more accurate. These estimates can be found in the wall areas in the first and third row from the top in Figure 6. Additionally, our approach also captures object boundaries more clearly, which can be seen at the wood sticks and pillows in Figure 6.



(**a**) (**b**) (**c**) (**d**)

Figure 6. Sample qualitative results of disparity estimation on the InStereo2K test set: (**a**) input RGB images, (**b**) ground-truth disparity, (**c**) estimated disparity using [9], and (**d**) estimated disparity using our method.

In addition to the evaluation of accuracy, we performed a time study on both methods by using the same dataset. The results are shown in Table 3. Our approach can process the images at an average rate of 14.01 frames per second (FPS) on an NVIDIA RTX 3060 GPU, which is significantly faster than the 7.20 FPS achieved by the PASMnet.

5.5. Ablation Study

Based on the previous experimental results, it can be seen that our proposed unsupervised stereo-matching scheme is effective in indoor stereo matching. To further understand how the surface normal contributes to our problem, we study the effectiveness of each design component related to normal estimation. In our proposed scheme, surface normal information is incorporated through three main design components: pre-training the feature extractor and normal-estimation branch for the normal-estimation task, normal consistency loss in (5), and the normal integration component introduced in Section 3.3.1. In this ablation study, we design four different configurations based on our proposed scheme by disabling and enabling some of these design components.

In our baseline configuration (Configuration I), the neural network only consists of the feature extractor and the disparity-estimation branch without the normal integration component. Additionally, the feature extractor has not been pre-trained on the NYU v2 dataset. Unsupervised training of this configuration for disparity estimation only relies on (3), (4), and (6). This configuration represents an unsupervised stereo-matching scheme without any surface normal information. Building upon Configuration I, we introduce the normal-estimation branch in Configuration II. Additionally, both the feature extractor and normal-estimation branch are pre-trained with the NYU v2 dataset in this configuration. In Configuration III, we further include the normal consistency loss (5) in the unsupervised training process for disparity estimation. Lastly, the normal integration component is incorporated into the disparity-estimation branch in Configuration IV. Configuration IV also represents our proposed scheme. From Configuration I to IV, more and more surface normal information is included. Comparing these configurations can demonstrate that each additional piece of normal information is beneficial to unsupervised indoor stereo matching.

The results of the ablation study are summarized in Table 4 and Figure 7. According to the quantitative results in Table 4, Configuration I estimates disparity with a significantly higher error and more outliers compared to the other configurations. Pre-training the feature extractor in Configuration II improves the disparity estimation considerably by a 41.02% decrease in the EPE and a 45.55% decrease in >3 px. Configuration II mainly relies on the photometric loss as the main supervisory signal. Introducing (5) into the training loss results in a lower EPE and >3 px by 3.41% and 5.49%, respectively. Configuration IV further incorporates the normal integration component, which leads to the most accurate disparity estimates among all four configurations with a 0.31% and 2.12% decrease in the EPE and >3 px, respectively, compared to Configuration III. This configuration is also the one we adopt as our final design.

Table 4. Quantitative results from the ablation study for disparity estimation with different configurations.

Configuration	Pre-Train	$\mathcal{L}_n$	Normal Integration	EPE-a (px)	>3 px-a (%)
I				5.080	28.45
II	✓			2.996	15.49
III	✓	✓		2.894	14.64
IV	✓	✓	✓	2.885	14.33

From the qualitative results in Figure 7, we can see that the estimated disparity maps from Configuration I are blurry with many inaccurate disparity estimates, especially at textureless regions, such as the wall, whiteboard, and floor. After pre-training the neural network with the NYU v2 dataset in Configuration II, more-defined object boundaries are captured at the shelf and table areas. However, significant errors are still visible at objects with low textures. As the normal consistency loss and normal integration component are included, the quality of the estimated disparity maps increases, especially at large, flat, and textureless regions that are typically ambiguous for stereo matching. This can be observed

in the disparity maps computed by Configuration IV. These disparity maps contain smooth and accurate disparity estimates at textureless areas.

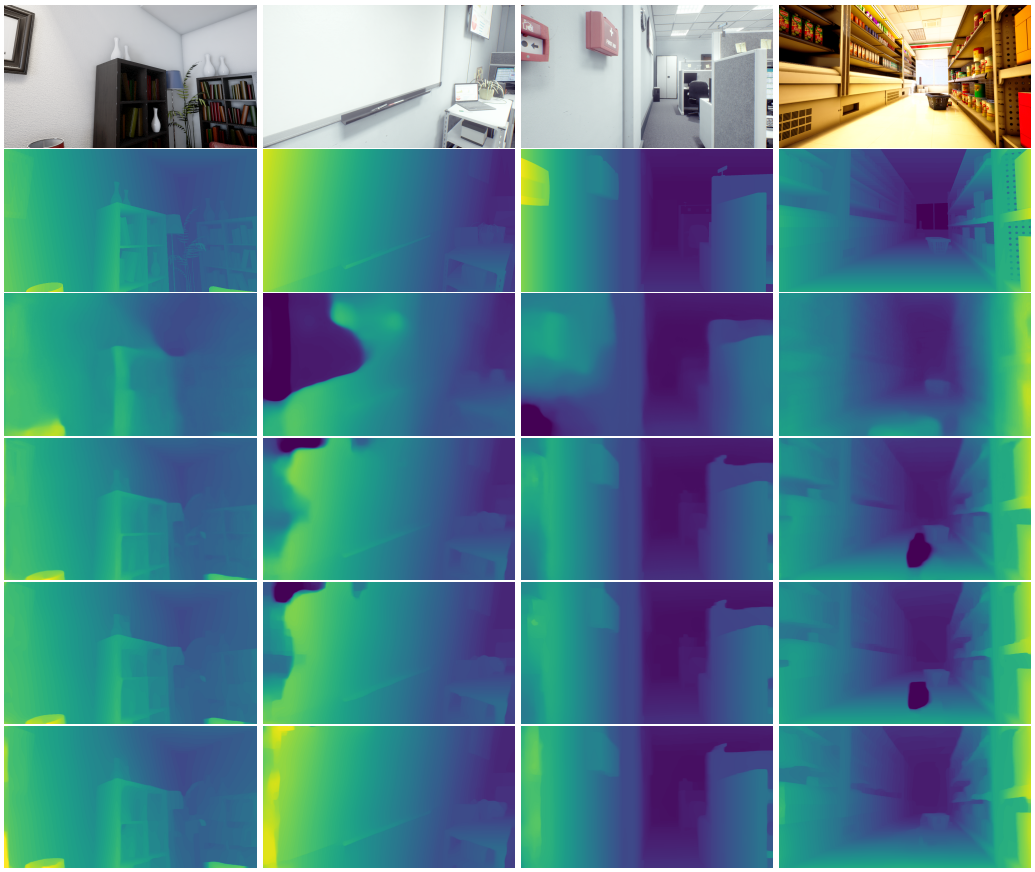


Figure 7. Sample qualitative results from the ablation study. The images from top to bottom are RGB image inputs, ground-truth disparity, and disparity estimates using Configurations I to IV as described in Table 4.

Overall, the above results demonstrate the significance of integrating surface normal information into unsupervised indoor stereo matching. Since pre-training the network for surface normal estimation, the normal consistency loss (5), and the normal integration component introduced in Section 3.3.1 involve standalone designs independent of other modules and training losses typically used in deep-learning-based stereo matching, we expect them to provide a similar performance improvement when they are integrated with supervised stereo-matching approaches for indoor applications. However, further experiments are required to formally demonstrate this.

6. Conclusions

In this work, we addressed the problem of unsupervised indoor stereo matching. We proposed a neural network design that consists of a feature extractor, a surface normal-estimation branch, and a disparity-estimation branch. The training of our network is performed in two stages. First, the extraction module and the normal-estimation branch are trained to estimate the surface normal with supervised learning by using the NYU v2 dataset. The disparity-estimation branch is then trained in an unsupervised manner while incorporating the surface normal estimated by the normal-estimation branch. Due to the lack of large datasets with real indoor stereo images, the second stage of training is carried out by using a large synthetic indoor stereo dataset. Experimental results demonstrate that the normal-estimation branch estimates the surface normal accurately. With the aid of normal estimation, the disparity-estimation branch estimates high-quality disparity for indoor scenes. Our method achieves higher accuracy in disparity estimation than a recent

unsupervised method. It also demonstrates a better generalization ability when it is applied to images that are visually different from the training images.

As a future direction, the proposed design may be further refined by jointly improving the normal-estimation branch and the disparity-estimation branch. Unsupervised surface normal estimation may be approached to reach a fully unsupervised training strategy. It is also important to quantify the effectiveness of integrating surface normal information into a supervised stereo-matching method to further understand its potential for indoor scenarios. Lastly, integrating this method with a robotic system for various applications is another future direction of study.

Author Contributions: X.F. and A.J.A. conceptualized the design; X.F. developed the algorithms, performed the experiments, and prepared the draft; B.F. and S.J. supervised the research and reviewed the manuscript; B.F. acquired the funding. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Canadian Mitacs Accelerate Project IT16435 and Avidbots Corp., Kitchener, ON, Canada.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: A.J.A. was employed by Avidbots Corp. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Lee, Z.; Juang, J.; Nguyen, T.Q. Local Disparity Estimation With Three-Moded Cross Census and Advanced Support Weight. *IEEE Trans. Multimed.* **2013**, *15*, 1855–1864. [CrossRef]
2. Kolmogorov, V.; Zabih, R. Computing visual correspondence with occlusions using graph cuts. In Proceedings of the IEEE International Conference on Computer Vision, Vancouver, BC, Canada, 7–14 July 2001; Volume 2, pp. 508–515.
3. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef] [PubMed]
4. Mei, X.; Sun, X.; Zhou, M.; Jiao, S.; Wang, H.; Zhang, X. On building an accurate stereo matching system on graphics hardware. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011; pp. 467–474.
5. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 66–75.
6. Chang, J.R.; Chen, Y.S. Pyramid Stereo Matching Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
7. Zhang, F.; Prisacariu, V.; Yang, R.; Torr, P.H. GA-Net: Guided Aggregation Net for End-To-End Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 185–194.
8. Cheng, X.; Zhong, Y.; Harandi, M.; Dai, Y.; Chang, X.; Li, H.; Drummond, T.; Ge, Z. Hierarchical Neural Architecture Search for Deep Stereo Matching. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; Volume 33, pp. 22158–22169.
9. Wang, L.; Guo, Y.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W. Parallax Attention for Unsupervised Stereo Correspondence Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2108–2125. [CrossRef] [PubMed]
10. Mayer, N.; Ilg, E.; Häusser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
11. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
12. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3061–3070.

13. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2366–2374.
14. Wofk, D.; Ma, F.; Yang, T.; Karaman, S.; Sze, V. FastDepth: Fast Monocular Depth Estimation on Embedded Systems. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 6101–6108.
15. Farooq Bhat, S.; Alhashim, I.; Wonka, P. AdaBins: Depth Estimation Using Adaptive Bins. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4008–4017.
16. Zhou, J.; Wang, Y.; Qin, K.; Zeng, W. Moving Indoor: Unsupervised Video Depth Learning in Challenging Environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8617–8626.
17. Yu, Z.; Jin, L.; Gao, S. P^2Net: Patch-Match and Plane-Regularization for Unsupervised Indoor Depth Estimation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 206–222.
18. Wang, Q.; Zheng, S.; Yan, Q.; Deng, F.; Zhao, K.; Chu, X. IRS: A Large Naturalistic Indoor Robotics Stereo Dataset to Train Deep Models for Disparity and Surface Normal Estimation. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shenzhen, China, 5–9 July 2021; pp. 1–6.
19. Kusupati, U.; Cheng, S.; Chen, R.; Su, H. Normal Assisted Stereo Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2186–2196.
20. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 746–760.
21. Wei, B.; Wang, W.; Xu, Y.; Guo, Y.; Hong, S.; Zhang, X. InStereo2K: A large real dataset for stereo matching in indoor scenes. *Sci. China Inf. Sci.* **2020**, *63*, 1869–1919.
22. Fouhey, D.F.; Gupta, A.; Hebert, M. Data-Driven 3D Primitives for Single Image Understanding. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3392–3399.
23. Fouhey, D.F.; Gupta, A.; Hebert, M. Unfolding an Indoor Origami World. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 687–702.
24. Ladický, L.; Zeisl, B.; Pollefeys, M. Discriminatively Trained Dense Surface Normal Estimation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 468–484.
25. Wang, X.; Fouhey, D.F.; Gupta, A. Designing deep networks for surface normal estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 539–547.
26. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In Proceedings of the IEEE IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2650–2658.
27. Wang, P.; Shen, X.; Russell, B.; Cohen, S.; Price, B.; Yuille, A.L. SURGE: Surface Regularized Geometry Estimation from a Single Image. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 172–180.
28. Bansal, A.; Russell, B.; Gupta, A. Marr Revisited: 2D-3D Alignment via Surface Normal Prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5965–5974.
29. Qi, X.; Liao, R.; Liu, Z.; Urtasun, R.; Jia, J. GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 283–291.
30. Qi, X.; Liu, Z.; Liao, R.; Torr, P.H.S.; Urtasun, R.; Jia, J. GeoNet++: Iterative Geometric Neural Network with Edge-Aware Refinement for Joint Depth and Surface Normal Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 969–984. [CrossRef] [PubMed]
31. Zhang, Z.; Cui, Z.; Xu, C.; Yan, Y.; Sebe, N.; Yang, J. Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4101–4110.
32. Liao, S.; Gavves, E.; Snoek, C.G.M. Spherical Regression: Learning Viewpoints, Surface Normals and 3D Rotations on N-Spheres. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9751–9759.
33. Do, T.; Vuong, K.; Roumeliotis, S.I.; Park, H.S. Surface Normal Estimation of Tilted Images via Spatial Rectifier. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 265–280.
34. Bae, G.; Budvytis, I.; Cipolla, R. Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13117–13126.
35. Piccinelli, L.; Sakaridis, C.; Yu, F. iDisc: Internal Discretization for Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21477–21487.
36. Ikehata, S. PS-Transformer: Learning Sparse Photometric Stereo Network using Self-Attention Mechanism. In Proceedings of the British Machine Vision Conference, Virtual, 22–25 November 2021.

37. Ju, Y.; Jian, M.; Wang, C.; Zhang, C.; Dong, J.; Lam, K.M. Estimating High-resolution Surface Normals via Low-resolution Photometric Stereo Images. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. Available online: https://ieeexplore.ieee.org/document/10208243 (accessed on 3 November 2023). [CrossRef]
38. Scharstein, D.; Szeliski, R.; Zabih, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision, Kauai, HI, USA, 9–10 December 2001; pp. 131–140.
39. Zbontar, J.; LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1592–1599.
40. Li, Z.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F.X.; Taylor, R.H.; Unberath, M. Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6177–6186.
41. Zhao, H.; Zhou, H.; Zhang, Y.; Chen, J.; Yang, Y.; Zhao, Y. High-Frequency Stereo Matching Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1327–1336.
42. Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; Liu, S. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16242–16251. [CrossRef]
43. Tankovich, V.; Häne, C.; Zhang, Y.; Kowdle, A.; Fanello, S.; Bouaziz, S. HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14357–14367.
44. Wang, Q.; Xing, H.; Ying, Y.; Zhou, M. CGFNet: 3D Convolution Guided and Multi-scale Volume Fusion Network for fast and robust stereo matching. *Pattern Recognit. Lett.* **2023**, *173*, 38–44. [CrossRef]
45. Zhou, C.; Zhang, H.; Shen, X.; Jia, J. Unsupervised Learning of Stereo Matching. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1576–1584.
46. Li, A.; Yuan, Z. Occlusion Aware Stereo Matching via Cooperative Unsupervised Learning. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; pp. 197–213.
47. Liu, P.; King, I.; Lyu, M.R.; Xu, J. Flow2Stereo: Effective Self-Supervised Learning of Optical Flow and Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6647–6656.
48. Song, T.; Kim, S.; Sohn, K. Unsupervised Deep Asymmetric Stereo Matching with Spatially-Adaptive Self-Similarity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13672–13680.
49. Li, B.; Shen, C.; Dai, Y.; van den Hengel, A.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1119–1127.
50. Roy, A.; Todorovic, S. Monocular Depth Estimation Using Neural Regression Forest. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5506–5514.
51. Jung, H.; Kim, Y.; Min, D.; Oh, C.; Sohn, K. Depth prediction from a single image with conditional adversarial networks. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 1717–1721.
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Khamis, S.; Fanello, S.; Rhemann, C.; Kowdle, A.; Valentin, J.; Izadi, S. StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 573–590.
54. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
55. Nekrasov, V.; Dharmasiri, T.; Spek, A.; Drummond, T.; Shen, C.; Reid, I. Real-Time Joint Semantic Segmentation and Depth Estimation Using Asymmetric Annotations. In Proceedings of the International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 7101–7107.
56. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. Group-Wise Correlation Stereo Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3268–3277.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.