Article

Unsupervised Monocular Depth Estimation Based on Residual Neural Network of Coarse–Refined Feature Extractions for Drone

Tao Huang, Shuanfeng Zhao * , Longlong Geng and Qian Xu

School of Mechanical Engineering, Xi'an University of Science and Technology, Xi'an 710054, China; ht775628@163.com (T.H.); SDMLJC@163.com (L.G.); xuqian1882934@126.com (Q.X.)

* Correspondence: zsf@xust.edu.cn; Tel.: +86-029-8558-31

Received: 24 September 2019; Accepted: 15 October 2019; Published: 17 October 2019

Abstract: To take full advantage of the information of images captured by drones and given that most existing monocular depth estimation methods based on supervised learning require vast quantities of corresponding ground truth depth data for training, the model of unsupervised monocular depth estimation based on residual neural network of coarse–refined feature extractions for drone is therefore proposed. As a virtual camera is introduced through a deep residual convolution neural network based on coarse–refined feature extractions inspired by the principle of binocular depth estimation, the unsupervised monocular depth estimation has become an image reconstruction problem. To improve the performance of our model for monocular depth estimation, the following innovations are proposed. First, the pyramid processing for input image is proposed to build the topological relationship between the resolution of input image and the depth of input image, which can improve the sensitivity of depth information from a single image and reduce the impact of input image resolution on depth estimation. Second, the residual neural network of coarse–refined feature extractions for corresponding image reconstruction is designed to improve the accuracy of feature extraction and solve the contradiction between the calculation time and the numbers of network layers. In addition, to predict high detail output depth maps, the long skip connections between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions are designed. Third, the loss of corresponding image reconstruction based on the structural similarity index (SSIM), the loss of approximate disparity smoothness and the loss of depth map are united as a novel training loss to better train our model. The experimental results show that our model has superior performance on the KITTI dataset composed by corresponding left view and right view and Make3D dataset composed by image and corresponding ground truth depth map compared to the state-of-the-art monocular depth estimation methods and basically meet the requirements for depth information of images captured by drones when our model is trained on KITTI.

Keywords: unsupervised; residual neural network; image reconstruction; monocular depth estimation

1. Introduction

In recent decades, drones are widely used in various fields due to low cost, high flexibility and reliability, such as drone aerial photography, drone rescue, drone plant protection and so on [1]. To take full advantage of the information of images captured by drones, obtaining depth information from a monocular camera in the drone is studied. It is known that image depth information can be easily obtained by camera binoculars, yet they are not suitable for drones due to high price and large size [2]. However, monocular depth estimation from a single image is an ill-posed problem in essence since an image has an infinite number of 3D scenes [3].

Monocular depth estimation is challenging, since there may be many distinct 3D scenes depicted in the same 2D image [4]. To address the problem of monocular depth estimation, a great deal of work focuses on depth estimation from a single image. Though shape from motion (SFM) [5], shape from shading (SFS) [6] and depth from focus or defocus (DFF/DFD) [7,8] are considered to be classical algorithms for monocular depth estimation, these methods are not widely used due to the device cost, high standard requirement for taking images, and the result is susceptible to occlusion and correspondence matching and so on. With the continuous improvement of computer performance and the gradual maturity of deep learning theory, deep learning is widely used in various fields [9,10], and it is possible to get depth information from a simple image like a binocular view [11,12]. Currently, to obtain the depth information of 2D images methods fall mainly into two major categories based on deep neural network according to whether vast quantities of corresponding ground truth depth data for training are required: one is the monocular depth estimation based supervised learning, another is the monocular depth estimation based unsupervised learning.

In recent years, supervised deep learning approaches have demonstrated promising results for single image depth prediction, which appear to capture the statistical relationship between appearance and distance to objects well. To train neural networks through vast quantities of corresponding ground truth depth data, multiple large-scale depth-contained datasets have been published, such as NYUD [13] presented indoor images and Make3D [14] recorded outdoor images. Given enough ground truth depth training data, deep neural networks-based approaches have achieved very promising performance of depth estimation from single images. Eigen et al. [15] proposed a multi-scale convolution neural network (CNN) architecture that integrates coarse-scale depth prediction with fine-scale prediction to predict depth in 2014. However, the depth map was extremely poor compared to ground truth depth map. According to the essence of depth information that is a classification from far to near layer, Cao et al. [16] proposed a simple effective approach to turn the problem of depth estimation into the task of a pixel-wise classification based on the advantage of deep residual network in 2016. Li et al. [17] combined deep learning features on image patches with hierarchical conditional random field (CRF) defined on a super pixel segmentation of the image in 2015, who use pre-trained AlexNet [18] features of image patches to predict depth at the center of the super-pixels. Liu et al. [19] proposed to use the combination of random field model and convolutional neural network to achieve monocular depth estimation in 2014. The core idea of the method is that the first-order convolutional neural network and second-order convolutional neural network are trained by a combination of first-order and second-order terms of conditional random field (CFR). Roy et al. [20] presented a novel deep architecture, called neural regression forest (NRF) that combines random forests and a convolutional neural network (CNN), for depth estimation from a single image in 2016. However, supervised learning models require ground truth depth data which are not flexible to deploy cross-application scenarios. Furthermore, supervised monocular depth estimation can be regarded as an ill-posed problem due to the scale ambiguity issue [21].

Very recently, unsupervised deep learning approaches have been introduced that learn to predict depth maps directly from the intensity images in a stereo setup—without the need for an additional supplementary modality for capturing the ground truth depth data. These models operate by learning the correspondence field between the two different images views of a calibrated stereo camera using only the rectified left and right images [22]. A remarkable advantage of unsupervised depth estimation lies in avoiding the use of costly ground truth depth annotations in training—in contrast left and right views from the same scene captured by a calibrated stereo camera that are easy to get compared to the ground truth depth data are used in training. To train neural network based unsupervised learning, binocular stereo images datasets have been published, such as KITTI [23] and Cityscapes [24] which are collected in the outdoors with calibrated stereo cameras. Garg et al. [25] proposed an encoder–decoder architecture similar to FlowNet in 2016, which was trained to predict a single image depth map through the method that used binocular stereo pairs (for which the inter-camera transformation is known) and trained a network to predict the depth that minimizes the photometric difference between the

true right image and one synthesized by warping the left image into the right's viewpoint, using the predicted depth. However, depth map is not clear enough and the outline of the small target is almost invisible through the method proposed by Garg et al. Xie et al. [26] proposed a network of Deep3d to addresses the problem of monocular depth estimation in 2016, where their goal was to generate the corresponding left view or right view from an input right view or left view in the context of binocular pairs, and again using an image reconstruction loss, the method produces disparities for each pixel. The disadvantage of their method is that by increasing the number of candidate disparity values, the memory consumption of the algorithm is also increased greatly, making it difficult to scale their approach to bigger output resolutions. Zhou et al. [27] presented an unsupervised monocular model for the task of monocular depth and camera motion estimation from unstructured video sequences in 2017, which use single-view depth and multi-view pose networks, with a loss based on warping nearby views to the target using the computed depth and pose. The drawbacks of the unsupervised monocular model proposed by Zhou et al. lie in the failure to fully exploit the relations between camera pose and depth prediction and ignorance of the fact that pose estimation from depth is a well-studied problem with known geometric properties and well performing algorithms. Inspired by recent advances in direct visual odometry (DVO), Wang et al. [28] proposed a simple normalization strategy that circumvents many of above-mentioned problems and delivers noticeably improved performance than that of Zhou et al.

Above all, the biggest drawback of monocular depth estimation based on supervised learning in References [15–17,19,20] is that these approaches rely on having high quality, pixel aligned, ground truth depth data which is not easy to get at training time and can be regarded as an ill-posed problem due to the scale ambiguity issue. The depth map is not clear enough and the outline of the small target is almost invisible through the methods of unsupervised monocular depth estimation proposed by References [25–28]. In addition, none of these works about unsupervised monocular depth estimation considered the principle of binocular depth estimation as the constraint and united training loss functions to improve the quality for better depth estimation. In this paper, we followed the unsupervised learning training and proposed a novel model of unsupervised monocular depth estimation for drones based on existing unsupervised learning depth estimation methods and inspired by the principle of binocular depth estimation, as illustrated in Figure 1. Our model introduced a virtual camera through deep residual convolution neural network based on coarse–refined feature extractions. The depth map was then calculated according to the image and correspondingly reconstructed the image based on the principle of binocular depth estimation. Specifically, we proposed the following contributions: First, the pyramid processing for input image was proposed to build the topological relationship between input image size and depth of input image, which can improve the sensitivity of depth information from a single image and reduce the impact of input image size on depth estimation. Second, the residual neural network of stereo images coarse–refined feature extractions for corresponding image reconstruction was designed to improve the accuracy of feature extraction and solve the contradiction between the calculation time and the numbers of network layer. In addition, the long skip connections between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions were designed to predict high detail output depth maps. Finally, we proposed a novel training loss to improve accuracy of monocular depth estimation that was a combination of the loss of corresponding image reconstruction based on the structural similarity index (SSIM) [29], the loss of approximate disparity smoothness and the loss of depth map based on the consistency between the right depth map generated according to right view and generated left view and the left depth map generated according to left view and generated right view.

The article is composed as follows. Section 2 introduces our method of unsupervised monocular depth estimation which is divided into three parts: (a) The principle of our model for monocular depth estimation. (b) The outstanding advantages of our model. (c) The details of our model. In Section 3, we discuss the implementation and experimental results. To verify the superiority of our model in monocular depth estimation, we analyzed our model from four aspects including comparison with

state-of-the-art method on the KITTI dataset, an ablation study that confirms the superior performance of our model, generalizing to other datasets and the result on a real scene. Finally, Section 4 presents the conclusion.

Figure 1. The schematic of unsupervised monocular depth estimation. X indicates that the three-dimensional scene is mapped to the two-dimensional plane. Z is the depth of the image. x_l is the coordinate of the left camera imaging and x_r is the coordinate of the right camera imaging. f is the focal length of the binocular camera. b is the distance of baseline between the left and right cameras of the binocular camera.

2. Method

Though monocular depth estimation is essentially a morbid issue, we proposed a novel model that could get the depth information from a single image influenced by Garg et al. [25], Xie et al. [26], Zhou et al. [27] and Wang et al. [28]. This section describes the principle of our model for monocular depth estimation, the outstanding advantages of our model and the details of our model.

2.1. The Principle of Our Model for Monocular Depth Estimation

Figures 1 and 2 show the principle of our model for monocular depth estimation. First, the input image was processed into a fixed size, then the input image was processed by the method of pyramid. Second, an innovative residual neural network of coarse–refined feature extractions for corresponding image reconstruction with long skip connections between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions was proposed, through which the left approximate disparity map d^l or right approximate disparity map d^r was generated by the process that features outputted by the coarse residual neural network were multiplied by a fixed scale factor, and then the corresponding right image $\tilde{I}^r$ or left image $\tilde{I}^l$ was reconstructed based on the generated left approximate disparity map d^l and input left image I^l or the generated right approximate disparity map d^r and input right image I^r according to the method of bilinear interpolation. Finally, the left depth map z^l or right depth map z^r was generated based on input left image I^l and corresponding reconstructed right image $\tilde{I}^r$ or the input right image I^r and corresponding reconstructed left image $\tilde{I}^l$ according to the principle of binocular depth estimation that is shown in Figure 1 and Formulas (1–4). According to some [30,31], the basic principle of binocular depth estimation is as follows: (a) Formulas (1,2) are derived according to the principle of triangle similarity. (b) Formula (3) that is used to calculate disparity between the left view and corresponding right view is derived according to Formulas (1,2). (c) Formula (4) that is used to calculate depth from

single image is derived according to the baseline distance b between the cameras and the camera focal length f .

$$x - b = \frac{z}{f} * x_r, \quad (1)$$

$$x = \frac{z}{f} * x_l, \quad (2)$$

$$d = x_l - x_r = \frac{f * b}{z}, \quad (3)$$

$$z = \frac{d^* z}{b}. \quad (4)$$

To improve the training effect of our model, we proposed a novel training loss function that was a combination of the loss of corresponding view reconstruction based on SSIM [29], the loss of approximate disparity smoothness and the loss of depth map.

Figure 2. The flow chart of our monocular depth estimation.

2.2. The Outstanding Advantages of Our Model

According to the principle of our model for monocular depth estimation in Section 2.1, we could clearly see that the most outstanding advantage of our model was that depth map was predicted through unsupervised monocular depth rather than supervised monocular depth. As supervised learning-based monocular depth estimation requires vast quantities of corresponding ground truth depth data, it is difficult to carry out training. However, unsupervised learning-based monocular depth estimation can be trained with easier-to-obtain binocular stereo footage. In addition, a more accurate depth map was predicted by our model thanks to four outstanding innovations, which are as follows: First, the pyramid processing for input image was proposed to build the topological relationship between input image size and depth of input image, which could improve the sensitivity of depth information from a single image and reduced the impact of input image size on depth estimation. Second, the residual neural network of coarse-refined feature extractions for corresponding view reconstruction was designed to improve the accuracy of feature extraction and solve the contradiction

between the calculation time and the number of network layers. In addition, to predict high detail output depth maps, long skip connections between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions were designed. Third, the loss of corresponding image reconstruction based on the structural similarity index (SSIM), the loss of approximate disparity smoothness and the loss of depth map were united as a novel training loss to better train our model. Four, the depth map was more accurate and reliable based on the mature and reliable principle of monocular depth estimation compared with predicted direct disparity.

2.3. The Details of Our Model

To clearly explain the principle of the innovation of our model, this section details the innovation of our model including the pyramid processing for input image, the residual neural network of coarse–refined feature extractions for corresponding view reconstruction and the function of training loss.

2.3.1. The Pyramid Processing for Input Image

To reduce the influence of input image resolution on the result of the monocular depth estimation and given a fixed size of input image, the pyramid processing was proposed to build the topological relationship between the resolution of input image and the depth of input image. Figure 3 shows the specific pyramid processing for the input image. First, the values of Height and Width of the input image are obtained, then $i = 1$ was used as the initial value of i , while $i \leq 3$, $\text{Height} = \text{Height}/2$ and $\text{Width} = \text{Width}/2$ were executed cyclically. Finally, four different resolutions were generated, such as, the input image in our model was processed into four resolutions of 256×512 , 128×256 , 64×128 and 32×64 with the algorithm of the pyramid.

Figure 3. The flow chart of pyramid processing.

2.3.2. The Residual Neural Network of Coarse–Refined Feature Extractions for Corresponding View Reconstruction

The residual network of coarse–refined feature extractions that the corresponding right or left image was reconstructed based on the input left or right image was the core of our model for monocular depth estimation. Inspired by the principle of binocular depth estimation that the disparity map can be generated based on the left view and corresponding right view, we proposed a model of residual network of coarse–refined feature extractions for corresponding view reconstruction that is shown in Figure 4. The core idea of the corresponding image reconstruction of our model was as follows: First, features of the pyramid image processed in Section 2.1 were extracted by residual neural network, then the approximate disparity map was generated by setting the scale factor for images with features being extracted based on the residual neural network. Finally, the corresponding view was reconstructed according to the input image and corresponding approximate disparity map by the method of bilinear interpolation. To improve the quality of image reconstruction, the training loss was continuously optimized by the stochastic gradient descent method. Compared with other models,

our model was designed to take the advantages of residual blocks shown in Figure 5 to improve the accuracy of feature extraction and solve the contradiction between the calculation time and the number of network layers. The idea that the corresponding approximate disparity map obtained by the process that features outputted by the coarse residual neural network were multiplied by a fixed scale factor and the corresponding views were reconstructed based on the method of bilinear interpolation according to the corresponding approximate disparity map and the input image were raised. To facilitate fine detail predictions at the output resolution, our architecture included long skip connections between corresponding layers in the neural network of coarse feature extractions and the deconvolution neural network of refined feature extractions.

Figure 4. The structure of corresponding image reconstruction residual network model.

To clarify the details of our model, the formulas of *Conv*, *Resconv*, *Resblock*, *Upconv*, *Disp* and so on are defined. Formula (5) is the convolution in which x is the input of the convolution operation, num_layers is the number of channels of the output layer of the convolution operation, k is the filter size, $stride$ is the stride of convolution operation and $activation$ is the activation function for the convolution operation.

$$Conv = C(x, num_layers, k, stride, activation), \quad (5)$$

The details of the residual block of our model are shown in Figure 5, and according to the structure of residual block, the formula for the residual block is defined as:

$$Resconv = Resconv(x, num_layers, stride), \quad (6)$$

where x is the input of the residual operation, num_layers is the number of channels of the output layer of residual operation and $stride$ is the stride of residual operation. Figure 6 is *Resblock* of our model, according to the Figure 6 and the formula of *Resconv*, the formula of *Resblock* is defined:

$$Resblock = Resblock(x, num_layers, num_blocks), \quad (7)$$

where, x is the input of *Resblock*, num_layers is the number of channels of the output layer of *Resblock*, num_blocks is the number of residual block in *Resblock*.

Figure 5. Residual block of our model. The residual block was obtained from three successive convolutions, and $Conv1 = C(x, num_layers, 1, 1, RELU)$, $Conv2 = C(Conv1, num_layers, 3, 1, RELU)$, $Conv3 = C(Conv2, 4 \times num_layers, 1, 1, RELU)$, $Shortcut = Conv4 = C(x, 4 \times num_layers, 1, 1, s)$. Therefore the output of the residual block in our model was $RELU(Conv3 + Shortcut)$.

Figure 6. *Resblock* of our model. First, x , num_layers and Num_blocks are input and the initial value of i is 0. $Out = x$ is defined. When $i < Num_block - 1$, $Out = Resconv(Out, num_layers, 1)$ and $i = i + 1$ are executed, otherwise $Out = Resconv(Out, num_layers, 2)$ is executed.

Formulas (5–7) are the basic units of the residual network in our model. As precise image reconstruction cannot be implemented by the feature extraction only based on the residual neural network, deconvolution neural network-based refined feature extractions for corresponding view reconstruction was proposed by drawing on the idea of self-encoding neural network for image reconstruction. The purpose of the deconvolution operation was to restore the feature size of input image through the feature extraction of the convolutional neural network. In our model, the formula of *Upconv* was used as a deconvolution operation, which shown in Formula (8):

$$Upconv = Upconv(upsample(x, scale), num_layers, k, 1), \quad (8)$$

where *upsample* is the function that scales up or down the feature size, *x* is input of *upsample*, *scale* is the scale factor, *num_layers* is the number of the channels of the output layer of *Upconv*, *k* is the filter size and 1 is the stride. To reconstruct the image according to the disparity and corresponding single image, the idea of the approximate disparity map was proposed, the formula of which is defined:

$$Disp = 0.3C(x, 2, 3, 1, Sigmoid), \quad (9)$$

where 0.3 is the scale, *C* is the convolution operation, *x*, 2, 3, 1, *Sigmoid* are the parameters of the convolution operation of input, filter size, stride and activation function respectively.

Table 1 details the various layers in our network. The specific process of corresponding view reconstruction based on our model according to Figure 4 and Table 1 is as follows:

- Step 1: Residual neural network-based coarse feature extractions. Input to our network was the RGB camera image of pyramid. The residual neural network based on coarse feature extractions (without the final fully connected layer) successively extracted low resolution high-dimensional features from the input image of 256×512 . The residual neural network based on coarse feature extractions subsampled the input image in six stages: the output was $128 \times 256 \times 64$ through Conv1; the output was $64 \times 128 \times 64$ through Pool 1; the output was $32 \times 64 \times 256$ through Resblock_1; the output was $16 \times 32 \times 512$ through Resblock_2; the output was $8 \times 16 \times 1024$ through Resblock_3; the output was $4 \times 8 \times 2018$ through Resblock_4.
- Step 2: Deconvolution neural network-based refined feature extractions. Long skip connections between corresponding layers in the encoder and decoder were added to improve the performance on all metrics without affecting convergence. The deconvolution neural network based on refined feature extractions subsampled the Resblock_4 in six stages: (1) The output was $8 \times 16 \times 512$ through Upconv5; the output was $8 \times 16 \times 1536$ through Contact5; the output was $8 \times 16 \times 512$ through Iconv5. (2) The output was $16 \times 32 \times 256$ through Upconv4; the output was $16 \times 32 \times 768$ through Contact4; the output was $16 \times 32 \times 256$ through Iconv4. (3) The output was $32 \times 64 \times 128$ through Upconv3; the output was $32 \times 64 \times 384$ through Contact3; the output was $32 \times 64 \times 128$ through Iconv3; the output was $32 \times 64 \times 2$ through Disp3; the output was $64 \times 128 \times 2$ through Udisp3. (4) The output was $64 \times 128 \times 64$ through Upconv2; the output was $64 \times 128 \times 128$ through Contact2; the output was $64 \times 128 \times 64$ through Iconv2; the output was $64 \times 128 \times 2$ through Disp2; the output was $128 \times 256 \times 2$ through Udisp2. (5) The output was $128 \times 256 \times 32$ through Upconv1; the output was $128 \times 256 \times 96$ through Contact1; the output was $128 \times 256 \times 32$ through Iconv1; the output was $128 \times 256 \times 2$ through Disp1; the output was $256 \times 512 \times 2$ through Udisp1. (6) The output was $256 \times 512 \times 16$ through Upconv0; the output was $256 \times 512 \times 18$ through Contact0; the output was $256 \times 512 \times 16$ through Iconv0; the output was $256 \times 512 \times 2$ through Disp0.
- Step 3: Corresponding view reconstruction based on the method of bilinear interpolation. Right or left view was reconstructed according to the pyramid disparity map including Disp0, Disp1, Disp2, Disp4 and the corresponding input left or right image processed by the method of the pyramid.

Table 1. Our architecture of image reconstruction. The encoder neural network based on ResNet-50 was composed of Conv1, Pool 1, Resblock_1, Resblock_2, Resblock_3 and Resblock_4. The decoder neural network was composed of convolution layers from Upconv5 to Disp0.

Layer	Channel I/O	Kernel_Size	Stride	Num_Layers	Num_Blocks/Skip	Scale	Input
Conv1	3/64	7	2	64			RGB
Pool 1	64/64	3					Conv1
Resblock_1	64/256			64	3		Pool 1
Resblock_2	256/512			128	4		Resblock_1
Resblock_3	512/1024			256	6		Resblock_2
Resblock_4	1024/2048			512	3		Resblock_3
Upconv5	2048/512	3	1	512		2	Resblock_4
Contact5	512/512				Resblock_3		Upconv5
Iconv5	512/512	3	1	512			Contact5
Upconv4	512/256	3	1	256		2	Iconv5
Contact4	256/256				Resblock_2		Upconv4
Iconv4	256/256	3	1	256			Contact4
Upconv3	256/128	3	1	128		2	Iconv4
Contact3	128/128				Resblock_1		Upconv3
Iconv3	128/128	3	1	128			Contact3
Disp3	128/2	3	1	2			Iconv3
Udisp3	2/2					2	Disp3
Upconv2	128/64	3	1	64		2	Iconv3
Contact2	64/64				Pool 1		Upconv2
Iconv2	64/64	3	1	64			Contact2
Disp2	64/2	3	1	2			Iconv2
Udisp2	2/2					2	Disp2
Upconv1	64/32	3	1	32		2	Iconv2
Contact1	32/32				Conv1		Upconv1
Iconv1	32/32	3	1				Contact1
Disp1	32/2	3	1	2			Iconv1
Udisp1	2/2					2	Disp1
Upconv0	32/16	3	1	16		2	Iconv1
Contact0	16/16				Udisp1		Upconv0
Iconv0	16/16	3	1	16			Contact0
Disp0	16/2	3	1	2			Iconv0

2.3.3. The Function of Training Loss

To improve the clarity of the depth map, a novel function of training loss that included the loss of the corresponding view reconstruction based on the structural similarity index (SSIM) and the approximate disparity smoothness, and the loss of depth map based on the consistency between right depth map generated according to right view and generated left view and left depth map generated according to left view and generated right view was proposed to improve the precision of monocular neural estimation. The specific function of training loss is defined as

$$C_s = a_{ap}(C_{ap}^l + C_{ap}^r) + a_{ds}(C_{ds}^l + C_{ds}^r) + a_{lr}C_{lr}, \quad (10)$$

$$C = \sum_{i=1}^4 C_s, \quad (11)$$

where C_s is loss of each output scale, C is the total loss of training, C_{ap}^l is the loss of corresponding image reconstruction based on the left image, C_{ap}^r is the loss of corresponding image reconstruction based the right image, C_{ds}^l is the smoothness loss of disparity generated by the left image, C_{ds}^r is the smoothness loss of disparity generated by right image, C_{lr} is the loss of depth map based on the consistency between the right depth map generated according to right view and generated left view and the left depth map generated according to left view and generated right view, α_{ap} , α_{ds} , α_{lr} represent the proportion of corresponding image reconstruction loss, disparity smoothness loss and depth map loss, respectively.

The function of disparity smoothness loss mainly reduced the influence of the smoothness of the approximate disparity map generated by the convolutional neural network on the corresponding image reconstruction. To improve the smoothness of the disparity map, it was proposed that a gradient derivation on the x and y directions of the disparity map was performed. The specific function of disparity smoothness loss is defined as:

$$C_{ds}^l = \frac{1}{N} \sum_{i,j} \left| \partial_x d_{ij}^l \right| e^{-\|\partial_x I_{ij}^l\|} + \left| \partial_y d_{ij}^l \right| e^{-\|\partial_y I_{ij}^l\|}, \quad (12)$$

where $\partial_x d_{ij}^l$ is the gradient of the x direction of the disparity map, $\partial_y d_{ij}^l$ is the gradient of the y direction of the disparity map, I_{ij}^l is the input pyramid image, $e^{-\|\partial_x I_{ij}^l\|}$ is the smoothness coefficient in the direction of x , $e^{-\|\partial_y I_{ij}^l\|}$ is the smoothness coefficient in the direction of y , N is the number of pixels.

Inspired by Reference [29] that introduced the structural similarity index (SSIM) into the corresponding image reconstruction loss function, we used a combination of the absolute value of difference between the reconstructed image and the input image and SSIM based on the reconstructed image and the input image as the loss function of corresponding image reconstruction. The specific function of corresponding image reconstruction is shown in Formula (13).

$$C_{ap}^l = \frac{1}{N} \sum_{i,j} a \frac{1 - SSIM(I_{ij}^l, \tilde{I}_{ij}^l)}{2} + (1 - a) \|I_{ij}^l - \tilde{I}_{ij}^l\|, \quad (13)$$

where N is the number of pixels, I_{ij}^l is the input pyramid image, $\tilde{I}_{ij}^l$ is the reconstructed corresponding pyramid image, a is the proportion of structural similarity loss to image reconstruction loss.

Inspired by Formula (13), we used a combination of absolute value of the difference between the left and right depth map and structural similarity index (SSIM) as the loss function of depth map, with the main purpose of ensuring the consistency between the depth map generated by the left image and right image reconstructed by left image and the depth map generated by the right image and left image reconstructed by the right image. The specific loss function of disparity consistency is shown in Formula (14).

$$C_{lr} = \frac{1}{N} \sum_{i,j} a \frac{1 - SSIM(d_{ij}^l, d_{ij}^r)}{2} + (1 - a) |d_{ij}^l - d_{ij}^r|, \quad (14)$$

where d_{ij}^l is the depth map generated by the left image and right image reconstructed by the left image, d_{ij}^r is the depth map generated by the right image and left image reconstructed by the right image.

3. Experimental Results

In this section, we show an extensive experiment for evaluating the performance of our proposed model. We evaluated our approach on the raw sequences of the KITTI benchmark which is a popular dataset for single image depth map prediction. To demonstrate the effectiveness of the proposed unsupervised monocular depth estimation model, we compared the performance of our model to both the supervised and unsupervised single view depth estimation method in Section 3.3.1. In Section 3.3.2, several factors affecting the performance of monocular depth estimation were discussed and analyzed as follows: (a) We compared our model in which the input images were processed by the pyramid method with the one that the pyramid method had not used, thereby demonstrating that input images processed by the pyramid method could improve the sensitivity of depth information in monocular images and reduce the influence of input image size on the estimation result. (b) To test the outstanding advantages of residual neural networks, we compared the performance of our residual network with VGG-16 in our model. (c) Our model was compared with the one without long skip connections

between corresponding layers in the neural network of coarse feature extractions and the deconvolution neural network of refined feature extractions to show that long skip connection can predict high detail output depth maps. (d) To reflect the superiority of our proposed loss function, our novel loss function was compared with other well-known training losses. In addition, to verify the cross-dataset generalization ability of our model, our model and other monocular depth estimation methods were tested on Make3D that consisted of RGB/depth pairs in Section 3.3.3. Finally, the real outdoor scene was used to testify the superior performance of our model in Section 3.3.4.

3.1. Experimental Setup

3.1.1. Platform of Drone Experience

The drone shown in the Figure 7a was assembled with flight control of open source Pixhawk and related components of the drone. Given the technical packaging, drones on the market are not conducive to secondary development. NVIDIA Jetson TX2, an AI computer in the embedded field with a six-core Tegra processor and 256-core Pascal architecture GPU was installed on the drone. The connection between NVIDIA Jetson TX2 and the drone is shown in Figure 7b.

Figure 7. Platform of the drone experiment. (a) The drone assembled with flight control of open source Pixhawk and related components. (b) The connection diagram of NVIDIA Jetson TX2 and the drone.

3.1.2. Evaluation Metrics

Following previous work, we adopted the following evaluation metrics to quantitatively assess the performance of our unsupervised monocular depth estimation model. Specifically, we considered:

Average relative error (REL):

$$REL = \frac{1}{N} \sum_i \frac{|y_i^* - y_i|}{y_i^*},$$

Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_i (y_i^* - y_i)^2},$$

Average $\log_{10}$ error (log RMSE):

$$\log RMSE = \sqrt{\frac{1}{N} \sum_i \|\log y_i^* - \log y_i\|^2},$$

And threshold:

The percentage of pixels in the number of pixels that satisfy the condition that

$$\max = \left(\frac{y_i^*}{y_i}, \frac{y_i}{y_i^*} \right) = \delta < threshold,$$

where y_i^* is the ground truth depth, y_i is the predicted depth, N is the sum of the pixels of all test images.

3.1.3. Implementation Details

We implemented the system using the publicly available Tensorflow framework. For our experiments, the hardware of our training for unsupervised monocular depth estimation was the computer with the processor of Intel(R)Core(TM)i7-7700K CPU @4.20 GHz, the graphics card of NVIDIA GeForce GTX 1080, the memory of 16 GB. We used the computer with the operating system of Ubuntu16.04 and the used Tensorflow with the GPU developed by Google for our framework of deep learning. We trained our model with the popular KITTI datasets which are divided into 280,000 pairs of training sets and 1600 pairs of verification sets and resized to 256×512 . The maximum training step was 160,000 times, batchsize was eight, epoch was 50, training loss was saved for every 100 steps in training, training weight was saved for every 10,000 steps in training, and training initial learning rate was 0.00001 in our model. At the same time, learning rate was reduced to 50% of the initial learning rate when training accounted for 60% of the total step size, and learning rate was reduced to 25% of the initial learning rate when training accounted for 80% of the total step size. During optimization, we set the weighting of the different loss component to $\alpha_{ap} = 1$, $\alpha_{ds} = 0.85$, $\alpha_{lr} = 1$, and we used the Adam Optimizer to optimize weights of the network.

3.2. Training

It took 36.5 h to train our model with the computer and parameter set in Section 3.1 and each image took about 0.05 s. The specific results of the training loss are shown in Figure 8, in which we can clearly know that the loss of each part oscillates slightly and tends to stabilize and the total loss drops to around 0.4 after multiple training. It was found that the training effect based on the unsupervised monocular depth estimation model proposed in this paper was excellent.

Figure 8. Loss function of each part of the training process. (a) The disparity smoothness loss, (b) the corresponding image reconstruction loss, (c) the loss of depth map, (d) the total loss of our model that was the sum of losses of (a, b, c).

3.3. Experimental Results and Analysis

3.3.1. Comparison with State-of-the-Art Method on KITTI Dataset

In recent years, the monocular depth estimation based on supervised learning has been studied and enjoyed great success, such as Eigen et al. [15] and Cao et al. [16]. In addition, the monocular depth estimation based on unsupervised learning also has been studied and enjoyed great success because monocular depth estimation based on supervised learning have been restricted to scenes where large image collections and their corresponding pixel depths are available, such as Garg et al. [25], Xie [26], Zhou et al. [27] and Wang et al. [28]. To highlight the superiority of our model for monocular depth estimation, our model was compared with the state-of-the-art methods on the KITTI dataset, including the ones with supervised learning, such as Eigen et al. [15] and Cao et al. [16], and the ones with unsupervised learning, such as Zhou et al. [27] and Wang et al. [28]. Table 2 shows our results in relation to the state-of-the-art methods on the test image of the KITTI dataset. For all evaluation metrics, our model for depth estimation was close to or even better than the model of monocular depth estimation based on supervised learning (Eigen et al. [15] and Cao et al. [16]), where the error is lower and the accuracy is higher compared with the monocular depth estimation, while our model was unsupervised learning without using any ground-truth depth data in training, which showed that unsupervised depth estimation has higher feasibility. For comparison with the unsupervised monocular depth estimation methods, we were also very close to or even better than the best competitor (Garg et al. [25], Xie [26], Wang et al. [28]). Above all, our model significantly outperformed their results with current supervised or unsupervised monocular depth estimation, further demonstrating the effectiveness of the means we considered.

We also qualitatively compared the output of our method with the state-of-the-art methods in Figure 9. The comparison between the results of the state-of-the-art methods (such as the depth estimation method based on supervised learning proposed by Cao et al. [16] and the depth estimation methods based on unsupervised learning proposed by Garg et al. [25] and Wang et al. [28]) and the ground true depth map showed that: First, the sharpness of the depth map predicted by our model was better than others and closer to the ground true depth map compared with other models. The predictions of our model appeared more detailed and our depth maps seemed to be smoother. Second, the contours of relatively small objects could be detected by our model, such as cars, people, utility poles, etc. shown in Figure 9. Third, the accuracy of predicting close depth was higher than that in the long distance, which explained that the depth estimation of our model was similar to the depth estimation of human eye. In summary, the depth map prediction of our model, which had no requirement for vast quantities of corresponding ground truth data for training, achieved excellent results.

Table 2. Quantitative results of our model and comparison with state-of-the-art method on the KITTI dataset.

Method	Supervised	Error (Lower is Better)			Accuracy (Higher is Better)		
		REL	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [15]	Yes	0.203	6.307	0.282	0.702	0.890	0.958
Cao et al. [16]	Yes	0.202	6.523	0.275	0.678	0.895	0.965
Garg et al. [25]	No	0.208	6.856	0.283	0.678	0.885	0.957
Xie et al. [26]	No	0.189	6.642	0.254	0.752	0.904	0.961
Wang et al. [27]	No	0.159	5.789	0.234	0.796	0.923	0.963
Our model	No	0.135	5.446	0.215	0.895	0.979	0.986

Figure 9. Quantitative results of our model and comparison with state-of-the-art method on KITTI dataset. (a) The input image, (b) the ground truth depth map, (c) the depth map of monocular depth estimation based on supervised learning proposed by Cao et al., (d) the depth map of monocular depth estimation based on the unsupervised learning proposed by Garg et al., (e) the depth map of monocular depth estimation based on the unsupervised learning proposed by Wang et al. and (f) the depth map of our model for depth estimation.

3.3.2. Ablation Study

We presented an extensive ablation study on the KITTI dataset and analyzed the contributions of the various design choice in our model (see Table 3). We had several baseline models for the ablation study to confirm the superior performance of our model, including (1) our model without pyramid processing for input image, in which the input of training model was the image that has only a fixed resolution; (2) our model with VGG-16 for feature extractions, which was a convolutional neural network without residual block; (3) our model with only the loss of C_{ap}^l , in which the network was trained by the training loss C_{ap}^l that was the loss of corresponding view reconstruction; (4) our model with only the loss of C_{lr} , in which the network was trained by the training loss C_{lr} that was the loss of depth map; (5) our model without long skip connection, which lacks of the long skip connections between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions.

Compared with our model without pyramid processing for the input image with our model, it was very obvious that pyramid processing for input image could improve the performance for monocular depth estimation. Comparing our model with VGG-16 for feature extractions with our model with residual neural network, our results clearly demonstrated the benefit of using a deep residual neural network for feature extractions. To show that our approach benefitted from a novel function of training

loss that included the loss of corresponding view reconstruction, the loss of disparity smoothness and the loss of the depth map, we also compared our model with only loss of C_{ap}^l and our model with only loss of C_{lr} to our model. Here, the error of our model was lower and the accuracy was higher. Comparing our model without long skip connection with our model, our results clearly demonstrated the benefit of using a deep residual neural network based coarse–refined feature extractions with long skip connections for the task of single image depth map prediction. Above all, pyramid processing of the input image, the convolutional neural network based residual block, the long skip connections between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions and the combined loss based on the loss of disparity smoothness, the loss of corresponding image reconstruction, the loss of depth map with structural similarity index (SSIM) were introduced our model to greatly improve the performance of depth estimation from a single image.

Table 3. Results of the various design choices in our model on the KITTI dataset.

Method	Supervised	Error (Lower is Better)			Accuracy (Higher is Better)		
		REL	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Our model (No pyramid)	No	0.352	6.056	0.395	0.752	0.889	0.926
Our model (VGG-16)	No	0.452	8.976	0.556	0.646	0.735	0.806
Our model (loss of C_{ap}^l)	No	0.396	7.152	0.465	0.672	0.796	0.845
Our model (loss of C_{lr})	No	0.383	7.096	0.452	0.684	0.806	0.850
Our model (no skip)	No	0.362	6.758	0.425	0.696	0.854	0.895
Our model	No	0.135	5.446	0.215	0.895	0.979	0.986

3.3.3. Generalizing to Other Datasets

To illustrate that our model could generalize to other datasets, here our model was compared with the state-of-the-art methods, including monocular depth estimation based on supervised learning and monocular depth estimation based on unsupervised learning on the Make3D [29] that consists of only RGB/depth pairs and no stereo images. Using our model only trained on the KITTI dataset, we tested it on the Make3D. The experimental results are shown in Table 4. The following conclusions were obtained according to the comparison of Tables 2 and 4: First, the depth estimation accuracy of our model trained by KITTI dataset on KITTI was higher than that of Make3D. Second, the accuracy of the monocular depth estimation of our model was higher than the other methods, where it resulted in 0.328 REL, 7.529 RMSE, 0.348 Log RMSE.

Figure 10 gives qualitative results of our model on test images of Make3D. Qualitatively, our model could capture the general scene layout and object such as cars, trees well in images that shared similarities with the KITTI dataset. According to the Figure 10 and Table 4, the conclusion we could draw was that our model has excellent generalization ability.

Table 4. Quantitative results of our model and approaches on Make3D datasets.

Method	Supervised	Error (Lower is Better)			Accuracy (Higher is Better)		
		REL	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. [15]	Yes	0.417	8.526	0.403	0.692	0.899	0.948
Cao et al. [16]	Yes	0.462	9.972	0.456	0.656	0.887	0.945
Garg et al. [25]	No	0.443	8.326	0.398	0.662	0.885	0.932
Xie et al. [26]	No	0.410	8.125	0.378	0.683	0.895	0.938
Wang et al. [27]	No	0.387	7.895	0.354	0.704	0.899	0.946
Our model	No	0.328	7.529	0.348	0.751	0.924	0.962

Figure 10. Quantitative results of our model on Make3D.

3.3.4. Result on a Real Scene

In addition, to better reflect the generalization ability of our model, a real outdoor scene was used for experimental verification. The result of depth estimation in a real outdoor environment is shown in Figure 11. The following conclusions error reached according to the results shown in Figure 11: First, the contour of the scene could be predicted well when the distance was relatively close, while the contour of the scene was not clear when the distance was far away. The depth estimation accuracy was similar to the accuracy of human eye. Second, compared with the result of depth estimation on KITTI, a good effect of depth estimation could still be obtained for our model in real outdoor scenes. Third, the estimated depth of each frame was 0.058 seconds for our model, and in other words 18 frames per second were predicted for our model. In summary, our model could still obtain a good effect of depth estimation in the real outdoor scenes, which proved that our model had strong ability of generalization.

Figure 11. Examples of depth map predictions on real outdoor scenes. (a) The input image. (b) The depth map predicted by our model on real outdoor scenes.

4. Conclusions

We have presented a novel approach for unsupervised monocular depth estimation. According to the principle of binocular depth estimation, we introduced a virtual camera to achieve monocular estimation, thus turning the problem of unsupervised monocular depth estimation into the problem of corresponding image reconstruction. In our model, the pyramid processing was proposed to build the topological relationship between the input image resolution and depth map, which can improve the sensitivity of depth information in monocular images and reduce the influence of the input image size

on the estimation result. Compared with other models, our model was designed to take the advantages of residual blocks to improve the accuracy of feature extraction and solve the contradiction between the calculation time and the numbers of network layer. In addition, the idea that the corresponding approximate disparity map obtained by setting the scale factor for feature extractions outputted by the coarse residual neural network and the corresponding view was reconstructed according to input image and the corresponding approximate disparity map based on the method of bilinear interpolation was raised. To facilitate fine detail predictions at the output resolution, our architecture included long skip connections between corresponding layers in the neural network of coarse feature extractions and the deconvolution neural network of refined feature extractions. The experiment of Section 3.3.1 showed that our model exhibited excellent depth estimation performance compared with state-of-the-art method on the KITTI dataset. The experiment of Section 3.3.2 showed that the pyramid processing for the input image, the residual deep neural network based on coarse–refined feature extractions, the long skip between corresponding layers in the neural network of coarse feature extractions and deconvolution neural network of refined feature extractions and the novel training loss combined by the loss of corresponding image reconstruction, the loss of disparity smoothness and the depth map based on the consistency were introduced to our model to greatly improve the performance of monocular depth estimation. The experiment of Section 3.3.3 showed that our model still had good depth estimation performance for Make3D datasets. The experience of Section 3.3.4 showed that our model could achieve better depth estimation on a real scene. Above all, the experimental results showed that our model achieved state-of-the-art performance in single image depth map prediction on the popular KITTI dataset and has excellent generalization ability.

We will investigate unsupervised monocular depth estimation for further tasks such as monocular depth estimation from videos and also consider using unsupervised learning to solve more problems about neural networks.

Author Contributions: The manuscript was written through contributions of all authors. Conceptualization, S.Z.; data curation, L.G.; formal analysis, T.H.; funding acquisition, S.Z.; investigation, L.G. and Q.X.; methodology, S.Z. and T.H.; project administration, S.Z.; software, T.H.; supervision, S.Z.; validation, T.H.; visualization, Q.X.; writing—original draft, T.H.; writing—review and editing, S.Z.

Funding: This research was supported by the Shaanxi Provincial Education Department serves Local Scientific Research Plan in 2019 (Project NO.19JC028), Shaanxi Provincial Key Research and Development Program (Project NO.2018ZDCXL-G-13-9) and Shaanxi Provincial special project of technological innovation guidance (fund) (Program No.2019QYPY-055).

Conflicts of Interest: We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

References

1. Shangjie, J.; Bin, L.; Jun, L.; Yun, Z. Real-time detection of vehicle targets based on drones. *Bullet. Sur. Map.* **2017**, *1*, 164–168.
2. Zhenqiang, B.; Aihua, L.; Zhiqiang, C.; Meng, Y. Research progress of deep learning in visual localization and three-dimensional structure recovery. *Laser Optoelectron. Prog.* **2018**, *55*, 050007.
3. Jiang, G.; Jin, S.; Ou, Y.; Zhou, S. Depth Estimation of a Deformable Object via a Monocular Camera. *Appl. Sci.* **2019**, *9*, 1366. [[CrossRef](#)]
4. Tongneng, H.; Jiageng, Y.; Defu, C. Monocular image depth estimation based DenseNet. *Comput. Meas. Cont.* **2019**, *27*, 233–236.
5. Noah, S.; Steven, M.S.; Richard, S. Skeletal Graphs for Efficient Structure from Motion. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 24–26 June 2018; pp. 45–56.
6. Zhang, R.; Tsai, P.S.; Cryer, J.E.; Shah, M. Shape from Shading: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 690–706. [[CrossRef](#)]

7. Nayar, S.; Nakagawa, Y. Shape from Focus. *IEEE Trans. Pattern Anal. Mach. Intell.* **1994**, *16*, 824–831. [[CrossRef](#)]
8. Favaro, P.; Soatto, S. A Geometric Approach to Shape from Defocus. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 406–417. [[CrossRef](#)]
9. Shuanfeng, Z. Study on Driver Model Parameters Distribution for Fatigue Driving Levels Based on Quantum Genetic Algorithm. *Open Cybern. Syst. J.* **2015**, *9*, 1559–1566.
10. Shuanfeng, Z.; Liang, L.; Guanghua, X.; Jing, W. Quantitative diagnosis of a spall-like fault of a rolling element bearing by empirical mode decomposition and the approximate entropy method. *Mech. Syst. Sign. Process.* **2013**, *40*, 154–177.
11. Cang, Y.; He, H.; Qiao, Y. Measuring the Wave Height Based on Binocular Cameras. *Sensors* **2019**, *19*, 1338. [[CrossRef](#)]
12. He, L.; Yang, J.; Kong, B.; Wang, C. An Automatic Measurement Method for Absolute Depth of Objects in Two Monocular Images Based on SIFT Feature. *Appl. Sci.* **2017**, *7*, 517. [[CrossRef](#)]
13. 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.
14. Saxena, A.; Sun, M.; Ng, A.Y. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 824–840. [[CrossRef](#)]
15. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *arXiv* **2014**, arXiv:1406.2283.
16. Cao, Y.; Wu, Z.; Shen, C. Estimating depth from monocular images as classification using deep fully convolution residual network. *arXiv* **2016**, arXiv:1605.02305.
17. Li, N.B.; Shen, N.C.; Dai, N.Y.; Hengel, A.V.D.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1119–1127.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
19. Liu, F.; Shen, C.; Lin, G. Deep Convolutional Neural Fields for Depth Estimation from a Single Image. *arXiv* **2014**, arXiv:1411.6387.
20. Roy, A.; Todorovic, S. Monocular Depth Estimation Using Neural Regression Forest. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5506–5514.
21. Sunok, K.; Sunghwan, C.; Kwanghoon, S. Learning depth from a single images using visual-depth words. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec, QC, Canada, 27–30 September 2015; pp. 1895–1899.
22. Wenjie, L.; Alexander, G.S.; Raquel, U. Efficient deep learning for stereo matching. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5695–5703.
23. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI Vision Benchmark Suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
24. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3354–3361.
25. Garg, R.; BG, K.G.; Carneiro, G.; Reid, I. Unsupervised CNN for single view depth estimation: Geometry to the rescue. *arXiv* **2016**, arXiv:1603.04992.
26. Xie, J.; Girshick, R.; Farhadi, A. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolution neural networks. *arXiv* **2016**, arXiv:1604.03650.
27. Zhou, T.; Brown, M.; Suavely, N.; Lowe, D.G. Unsupervised learning of depth and ego-motion from video. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6612–6619.
28. Wang, C.; Buenaposada, J.M.; Zhu, R.; Lucey, S. Learning depth from monocular videos using direct methods. *arXiv* **2017**, arXiv:1712.00175.

29. Wang, C.; 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](#)]
30. Hirschm, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [[CrossRef](#)]
31. Peris, M. Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps. *Tech. Rep. IEICE PRMU* **2012**, *111*, 117–118.

© 2019 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 (<http://creativecommons.org/licenses/by/4.0/>).