

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



## Chapter

# Depth Learning Methods for Bridges Inspection Using UAV

*Hicham Sekkati and Jean-Francois Lapointe*

## Abstract

This paper is investigating learning methods using depth as a cue measurement that can be used for bridge inspection. We investigate learning methods based on mono, stereo, and multiview image input and discuss the constraints that allow some methods to perform better than others in various scenarios. We go over the state-of-the-art deep learning methods, including supervised and unsupervised methods. These methods will be compared and evaluated, based on constraints, performance, and accuracy, and how top methods should be selected for each scenario. The same database should be used for fair comparison between all methods ensuring that evaluations are unbiased, replicable, and meaningful.

**Keywords:** depth, 3D reconstruction, deep learning, bridge inspection, UAV

## 1. Introduction

Tragedies such as the recent collapse of the Morandi bridge in Italy [1] remind us of the importance of good and regular bridge inspections. Such inspections are conducted mainly manually but the advent of new technologies such as drones and depth estimation using deep learning paradigms has the potential to automate part of the task. Depth map estimation for bridge inspection can provide valuable information about the three-dimensional structure of the bridge. It allows for the identification of surface irregularities, cracks, deformations, and overall conditions. This process provides valuable information for analyzing the structural integrity and identifying potential issues. Several techniques can be used for depth map estimation in bridge inspection, including Time-of-Flight (ToF) cameras, Structured Light scanning (SL), Laser scanning or Light Detection and Ranging (LiDAR) systems, as well as photogrammetry-based depth estimation. ToF cameras emit infrared light and measure the time it takes for the light to bounce back from the bridge surface. This information is used to estimate the distance to each point on the surface, creating a depth map. ToF cameras can provide real-time depth information, making them suitable for dynamic inspections. SL scanning involves projecting a pattern of light onto the bridge surface and capturing the deformed pattern using a camera. By analyzing the distortions in the pattern, depth information can be calculated. This technique is effective for capturing detailed depth maps of bridge surfaces and can be performed using handheld devices or mounted on vehicles or drones. LiDAR

technology can also be utilized for depth map estimation, by emitting laser pulses and measuring their time of flight, LiDAR scanners can generate accurate depth information of the bridge's surface. High-resolution LiDAR scanners can capture detailed depth maps, facilitating precise analysis of the bridge structure. In photogrammetry-based techniques, and by employing computer vision algorithms, depth maps can be estimated from regular images captured by Unmanned Aerial Vehicles (UAVs) or drones during bridge inspections. Various techniques, such as depth from single image, structure from motion (SfM), or multi-view stereo, can be employed to extract depth information from the image data.

Each of the above techniques has its advantages and limitations, and the choice depends on factors such as the desired level of accuracy, resolution, portability, and budget. It is advisable to consult with experts in the field of bridge inspection or 3D imaging to determine the most suitable depth estimation method for a specific bridge inspection project. However, this paper focuses on photogrammetry-based methods and specifically the last advances using deep learning techniques to generate depth maps from images.

This paper first talks about depth perception and then discuss various ways of obtaining depth, be it by using pictorial cues, from monocular video, or from stereo and multi-view.

## **2. Depth perception**

Perception refers to the ability to interpret and organize stimuli from the surrounding environment, enabling effective understanding and behavior. The visual system plays a crucial role as one of the primary sources of stimuli for human beings. It comprises more than one million axons from each eye, responsible for capturing light reflected by objects. Research on human perception suggests that the visual system utilizes multiple sources of information to comprehend and infer the depth structure of scenes. The human visual system relies on various monocular or binocular cues present in two-dimensional retinal images to gather information that helps in perceiving the depth of the scene. Monocular cues can be divided into two categories: pictorial cues and motion-based cues.

Pictorial cues, or image cues, are derived from visual features observed in a static view of a scene. The most common pictorial cues used in computer vision methods for depth estimation from a single image are texture variations [2], shading [3], and defocus [4]. Texture variations are translated such as objects that are closer to the viewer tend to exhibit more detailed and distinct textures, while objects that are farther away appear to have less detailed or blurred textures. This texture gradient helps us infer depth. The distribution of the direction of edges or lines in a scene changes as objects recede into the distance. The spacing between these lines becomes smaller as objects get farther away, giving us a sense of depth. Depth from shading is a technique used to estimate the depth or 3D structure of a scene based on the shading or variations in brightness and contrast within an image. It relies on the principle that the distribution of light and shadows on objects can provide valuable information about their shape and depth. Depth from defocus is a depth estimation technique that utilizes the blur or defocus information in an image to infer the distance of objects. These algorithms take advantage of the fact that objects in the focus plane of a camera appear sharper, while objects that are out of focus exhibit varying degrees of blur. By analyzing the amount of blur in an image, these algorithms can infer the relative depth

of different objects in the scene. A good survey on methods using depth from defocus can be found in ref. [5]. Let us remind that depth from defocus technique is fundamentally different from depth from focus in the sense that the later uses a stack of images to model the blur in image while the former technique uses a single image. The stack of images can be obtained by varying the camera aperture like in ref. [6] or the focal length like in refs. [7–11]. In the next section, we only review learning depth from a single image.

On the other hand, motion-based cues make use of observer motion and leverage motion parallax, that is, nearby objects appear to move faster in the retinal image compared to distant objects. In contrast, binocular cues rely on the perception of depth through disparities between two different viewpoints of the same scene. By comparing the differences in the views from each eye, the brain can accurately triangulate the distance to an object. Binocular cues offer a high level of precision in estimating distances. The aim of the following sections is to go over the last advanced research on deep learning techniques for each category to estimate depth.

### **3. Depth from pictorial cues**

Deep learning methods have been successfully applied to estimate depth from a single image, leveraging the power of neural networks to learn complex mappings between image features and depth information. The first deep-learning method to estimate depth from a single image was proposed in ref. [12]. Image cues are learned as multi-scale features. The method uses two-step process involving two deep neural networks to predict depth information for a given scene. The first step is performed by a coarse-scale network. This network takes an input image as its input and predicts the depth of the scene at a global level. The second step involves a fine-scale network. This network takes the coarse depth map (output of the coarse-scale network) and refines it within local regions. The method achieves state-of-the-art results on both NYU Depth [13] and KITTI [14] datasets. The authors in ref. [15] proposed a framework to model the conditional probability on depth with conditional random field (CRF) and learn the probability distributions using deep convolutional neural network (CNN). The method has outperformed the classical methods on both indoor and outdoor scenes using both the public datasets NYU depth and the Make3D range image [16]. In ref. [17], the method also uses two CNN to capture both global and local scales while jointly estimating depth and semantic segmentation from a single image. The method in ref. [18] has trained a CNN to learn the relative depth ordering between pairs of points in the image. The same network was trained to learn independently the reflectance and shading in the image, however, no interaction between these metrics was taken into account. A better structural relationship between points in the image was learned by a CNN in ref. [19]. This method involves training a neural network to characterize the local geometry of a scene by predicting depth derivatives of various orders, orientations, and scales at every image location. In ref. [20], a method that combines a CNN and regression forest was presented to regress depth in the continuous domain. In ref. [21], the authors proposed a fully convolutional architecture (ResNet) for depth prediction enabling the generation of dense output maps with higher resolution, while significantly reducing the number of parameters required. Furthermore, the model can be trained using one-tenth of the data compared to the previous state-of-the-art approaches. An improvement of the previous method's accuracy was presented in ref. [22] by applying

a post-processing *via* fully-connected conditional random fields (CRF). More improvements using CRFs in cascades was presented in ref. [23]. In ref. [24], two cascade-deep fully connected CNNs were proposed to learn both global and local feature maps that are propagated to estimate depth. Most methods learn depth as a regression model and code implicit structure of the scene with CNNs features, but in ref. [25], a method was presented that explicitly modeled the defocus blur in an image and link it to image depth. In ref. [26], a method was presented that learns depth from defocus, unfortunately only qualitative results on NYU depth dataset were shown. Quantitative comparisons with state-of-arts learning methods on this dataset were not reported. **Tables 1** and **2** summarize the evaluation of depth estimation from state-of-the-art pictorial-based methods using both Make3D [16] and NYU Depth [13] datasets, respectively.

Method	Error ( $C_1$ )			Error ( $C_2$ )		
	AbsRel	log10	RMS	AbsRel	log10	RMS
Saxena et al. [2]	—	—	—	0.370	—	—
Roy et al. [20]	—	—	—	0.260	<b>0.119</b>	<b>12.400</b>
Liu et al. [15]	0.314	0.119	8.600	0.307	0.125	12.890
Anwar et al. [24]	0.213	0.075	<b>2.560</b>	<b>0.202</b>	<b>0.312</b>	<b>0.079</b>
Laina et al. [21]	0.176	0.072	4.460	—	—	—
Xu et al. [23]	<b>0.184</b>	<b>0.065</b>	<b>4.380</b>	<b>0.198</b>	<b>4.530</b>	<b>8.560</b>

**Table 1.**

Result comparisons of depth evaluation from pictorial-based methods on the Make3D dataset. Best performance is marked with bold fonts.

Method	Error			Accuracy		
	AbsRel	log10	RMS	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zoran et al. [18]	0.400	0.420	1.200	—	—	—
Liu et al. [15]	0.230	0.095	0.824	0.614	0.883	0.971
Wang et al. [17]	0.220	0.094	0.745	0.605	0.890	0.970
Eigen et al. [12]	0.215	0.285	0.907	0.611	0.887	0.971
Roy et al. [20]	0.187	0.078	0.744	—	—	—
Chakrabarti et al. [19]	0.149	0.205	0.620	0.806	0.958	0.987
Cao et al. [22]	0.141	0.060	0.540	0.819	0.965	0.992
Laina et al. [21]	0.127	0.055	0.573	0.811	0.953	0.988
Xu et al. [23]	0.121	0.052	0.586	0.811	0.954	0.987
Anwar et al. [24]	0.094	0.039	0.347	—	—	—
Carvalho et al. [25]	<b>0.036</b>	<b>0.016</b>	<b>0.144</b>	<b>0.993</b>	<b>1.000</b>	<b>1.000</b>

**Table 2.**

Result comparisons of depth evaluation from pictorial-based methods on the NYU depth dataset. Best performance is marked with bold fonts.

### 3.1 Loss functions

Learning depth from a single image is a challenging task due to the inherent ambiguity. However, there are techniques that leverage deep learning models to estimate depth from a single image. When training such models, a common choice for the loss function is the depth regression loss, which measures the difference between the predicted depth map and the ground truth depth map. One popular loss function for depth regression is the mean squared error (MSE) loss, given by:

$$L_1 = \sum_{x,y} (Z(x,y) - Z^*(x,y))^2 \quad (1)$$

where  $Z^*(x,y)$  is the depth map predicted by the model and  $Z(x,y)$  is the ground truth depth map. The MSE loss penalizes large errors between the predicted and ground truth depth values. Minimizing this loss helps the model learn to estimate depth accurately. Alternatively, you can also use other variations of the loss function such as the Huber loss or the smooth L1 loss, which provide a balance between the absolute and square losses and can be less sensitive to outliers. These loss functions can be advantageous when dealing with noisy or sparse depth measurements. When training models to learn depth from a single image, additional constraints or regularization terms might be necessary to improve the quality of the estimated depth. Some common techniques include incorporating geometric or semantic information, enforcing local smoothness, or using multi-scale depth supervision. Overall, the choice of the loss function depends on the specific requirements and characteristics of depth estimation task. Experimentation with different loss functions and regularization techniques can help find the most suitable approach for different applications.

## 4. Depth from monocular video

Estimating 3D interpretation from a monocular video is a fundamental and challenging topic in visual perception. Two common techniques used for this purpose are structure from motion (SfM) and simultaneous localization and mapping (SLAM). In the context of monocular video, SfM involves jointly estimating the camera motion and depth map of the scene, while SLAM involves jointly estimating the camera trajectory and the 3D structure of the scene. Monocular depth prediction using pairs of frames or more can be particularly challenging. This is because it requires reasoning about the relative camera pose, as well as estimating the disparity or optical flow between the frames. Furthermore, there is an inherent ambiguity in scale when using only monocular input, unless additional information or a consistent SLAM reconstruction pipeline is employed. The relative camera pose estimation is crucial for understanding the spatial relationship between frames and is necessary to compute accurate depth maps. Determining the camera motion accurately becomes more difficult when dealing with larger displacements, occlusions, or scene dynamics. Incorrectly estimated camera poses can lead to inaccurate depth predictions. Additionally, estimating disparity or optical flow between frames is challenging due to factors such as textureless regions, occlusions, and large displacements. These factors can introduce errors and ambiguities in the depth estimation process. Moreover, when using only monocular input, there is an inherent scale ambiguity. That is, without additional information, it is challenging to determine the absolute scale of the scene, leading to

depth maps that are only accurate up to an unknown scale factor. Despite the difficulties, researchers continue to develop methods that leverage monocular depth prediction using pairs of frames. These methods often combine deep learning techniques with geometric constraints and SLAM-like approaches to improve the accuracy of depth estimation and mitigate the inherent challenges. Ongoing research in this area aims to push the boundaries of monocular depth prediction and address the inherent limitations of single-camera input.

#### **4.1 Supervised deep learning methods**

Supervised deep learning methods have made significant progress in addressing the problem of determining 3D interpretation from monocular video. By training neural networks on large-scale annotated datasets, these methods can learn to estimate depth, motion, and other geometric properties from single-camera input. One popular approach to estimate depth is to use convolutional neural networks (CNNs) applied independently at each frame of the video. These networks take an image as input, as seen in the previous section, and output a depth map that represents the scene's 3D structure. By leveraging the large amounts of labeled data, CNNs can learn to infer depth cues such as perspective, texture gradients, and occlusion patterns. A common dataset used to compare methods in this category is KITTI [14]. In ref. [27], given a pair of frames and camera intrinsics, a deep architecture, computes depth, 3D camera motion, a set of 3D rotations and translations for the dynamic objects in the scene, and corresponding pixel assignment masks. However, the method uses a single image deep architecture for depth estimation.

Additionally, recurrent neural networks (RNNs) and particularly the convolutional LSTM networks have been employed to capture temporal dependencies and motion information in video sequences. By incorporating temporal context into the learning process, these models can estimate not only depth but also camera motion, object motion, and scene dynamics. In ref. [28], the proposed ConvLSTM network learns depth maps from a set of  $N$  consecutive video frames in a depth-supervised setting, allowing the ConvLSTM network to perform spatiotemporal reasoning about the image-depth map relationship.

To further enhance performance, supervised methods often make use of additional cues, such as optical flow or semantic segmentation. Optical flow provides dense pixel-level motion information [29], which can aid in-depth estimation and object tracking. Semantic segmentation helps in understanding the scene's layout and can guide the depth estimation process by leveraging object boundaries and semantic context.

However, it is important to note that despite the advancements, challenges remain in accurately determining 3D interpretation from monocular video. Factors such as occlusions, lighting variations, and scene complexity can still pose difficulties for supervised methods. Nonetheless, ongoing research and the continuous development of more sophisticated deep learning architectures hold promise for further improvements in tackling this problem.

##### *4.1.1 Supervised loss function*

When learning supervised depth from motion, a common approach is to use a loss function that compares the predicted depth map with the ground truth depth map. One such loss function is the photometric loss, which measures the difference

between the rendered image using the predicted depth map and the actual input image. The photometric loss can be defined as follows:

$$L_1 = \sum_{x,y} (I_1(x,y) - I_2(x,y))^2 \quad (2)$$

where  $I_1$  is the original input image and  $I_2$  is the image rendered using the predicted depth map and camera parameters, which can be done using techniques such as differentiable warping or inverse depth warping. In addition to the photometric loss, you can also incorporate smoothness regularization to encourage smooth depth predictions. The smoothness loss penalizes large depth gradients and helps produce more visually coherent depth maps. One common smoothness regularization term is the total variation loss, which can be defined as follows:

$$L_2 = \sum_{x,y} \|\nabla Z\| \quad (3)$$

where  $\|\nabla Z\|$  is the gradient of the predicted depth map in the x and y directions. The total loss for learning depth from motion can be a combination of the photometric loss and the smoothness regularization term:

$$L = L_1 + \alpha L_2 \quad (4)$$

where  $\alpha$  is a weighting factor that controls the relative importance of the photometric loss and the smoothness regularization term. By minimizing this total loss using techniques like gradient descent, you can train a model to learn depth from motion. Keep in mind that this is just one possible approach, and depending on your specific requirements and constraints, you may need to modify or customize the loss function accordingly.

## 4.2 Unsupervised deep learning methods

While supervised deep learning methods have achieved notable progress in determining 3D interpretation from monocular video, unsupervised deep learning methods have also shown promise in tackling this problem. Unsupervised approaches aim to learn 3D representations from unlabeled or self-supervised data, eliminating the need for costly manual annotations. One popular technique in unsupervised learning is based on the concept of “self-supervision.” By leveraging the temporal coherence of consecutive video frames, unsupervised methods can learn representations that capture the underlying 3D structure of the scene. These methods often utilize techniques such as photometric consistency, geometric consistency, or depth and ego-motion prediction.

In *photometric consistency-based* methods, the network learns to generate a synthesized view of the input frame from a different viewpoint using estimated depth or motion. The consistency between the synthesized view and the actual input frame is maximized during training, encouraging the network to learn meaningful depth representations. In ref. [30], the photometric consistency was achieved by training the network in a manner analogous to an autoencoder.

*Geometric consistency-based* methods exploit the geometric relationship between frames. They aim to minimize the disparity or reprojection error between multiple views of the same scene. By leveraging geometric constraints, the network can learn to



estimate depth and camera motion. The method in ref. [31] takes into account this geometry by learning camera pose between each two frames.

*Depth and ego-motion prediction* methods train the network to directly predict depth maps or motion vectors from single images or consecutive frames. These predictions are compared to ground truth or photometrically warped frames, respectively, to supervise the learning process. In ref. [32], photometric consistency was taken into account in training step while 3D geometry consistency was achieved by reconstructing 3D points cloud from depth and directly comparing the points cloud in a common reference frame. Nonrigid motion of dynamic objects in the scene was taken into account in ref. [33] by adding ResFlowNet architecture [34]. Photometric and geometric consistencies were combined in ref. [35] in a way to minimize the discrepancy between the reconstructed optical flow obtained from depth and egomotion, and the optical flow generated using FlowNet [36]. Dynamic scenes were handled in ref. [37] by learning objects motion independently from the egomotion without an explicit motion segmentation. Likewise, the motion model of moving objects in the work [38] is tackled by optical flow estimation using view synthesis objective as supervision, again with the assumption of photometric consistency. In ref. [39], a method was presented by adding another term that explicitly segments the scene into competing background and foreground masks. In most unsupervised methods, including mono or stereo-SfM (Structure from Motion) approaches, photometric consistency is a crucial principle used to guide the learning process. Photometric consistency is based on the assumption that the appearance of the same point in different views should remain consistent under different camera poses. In the context of monocular or stereo video, this consistency is expressed using a warping function and is often referred to as the view-synthesis loss [31]. In ref. [40], a generalization of the photometric loss was used by coupling the spatiotemporal variations in image sequence to the scene geometry with the goal to supervise both camera motion and depth in a new learning framework. **Table 3** summarizes the evaluation of depth estimation from state-of-the-art monocular motion-based methods using KITTI [14] dataset.

Method		Error				Accuracy $\delta$		
		AbsRel	SqRel	RMS	RMSlog	<1.25	<1.25 <sup>2</sup>	<1.25 <sup>3</sup>
Supervised	Vijayanarasimhan et al. [27]	—	<b>0.770</b>	—	—	—	—	—
	Kumar et al. [28]	<b>0.137</b>	<b>1.019</b>	<b>5.187</b>	<b>0.218</b>	<b>0.809</b>	<b>0.928</b>	<b>0.971</b>
Unsupervised	Zhou et al. [31]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
	Garg et al. [30]	0.169	1.080	5.104	0.273	0.740	0.904	0.962
	Mahjourian et al. [32]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
	Yin et al. [33]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
	Zou et al. [35]	0.150	1.124	5.507	0.223	0.806	0.933	0.973
	Casser et al.(M) [37]	0.141	1.026	5.291	0.215	0.816	0.9452	0.979
	Rajan et al. [39]	0.140	1.070	5.326	0.217	0.826	0.941	0.975
	Sekkati et al. [40]	0.137	<b>0.947</b>	<b>5.019</b>	0.216	0.838	0.933	0.970
Chen et al. [38]	<b>0.135</b>	1.070	5.230	<b>0.210</b>	<b>0.841</b>	<b>0.948</b>	<b>0.980</b>	

**Table 3.**

*Result comparisons of depth evaluation from monocular motion-based methods on the KITTI dataset. Best performance is marked with bold fonts.*

By training on vast amounts of unlabeled video data, unsupervised methods can capture 3D cues and learn to interpret depth, motion, and other scene properties. They can also handle challenging scenarios, such as dynamic scenes, textureless regions, and varying lighting conditions. While unsupervised methods have shown promising results, there are still challenges to overcome. The quality and accuracy of the learned representations heavily depend on the design of the self-supervision tasks and the data distribution. Additionally, the unsupervised learning process can be computationally expensive and may require substantial computational resources. Nevertheless, the development of unsupervised deep learning methods for 3D interpretation from monocular video holds great potential for advancing our understanding of the 3D world and enabling applications in robotics, augmented reality, and autonomous systems.

#### 4.2.1 Unsupervised loss functions

Let us consider two nonconsecutive frames from the image sequence  $I(x, y, t)$  at time  $t_1$  and  $t_2$ , denoted by  $I_1 = I(x, y, t_1)$  and  $I_2 = I(x, y, t_2)$ , respectively. The general idea followed by a previous work [41], and subsequently by others, is to minimize the photometric loss generated by the image difference:

$$L_1 = \sum_{x,y} |I_1(x, y) - I_2(\tau(x, y))| \quad (5)$$

where  $\tau$  is the warping function that maps pixel from  $I_2$  to  $I_1$ . Using image warping and image similarity metrics to supervise learning frameworks has certain limitations, particularly when dealing with large baseline views, occlusions, and image gradients. While these approaches can be effective in many cases, they may not fully capture the complexities of the scene geometry and structural edges, leading to some shortcomings. Several methods have been proposed to address these limitations, but they may not always be explicitly related to scene geometry. For example, image similarity metrics can help guide the learning process, but they might not directly capture the underlying scene geometry. Estimating depth and understanding the 3D structure of the scene is inherently related to scene geometry, which involves estimating accurate depths and surface orientations. Simple image similarity metrics may not fully encapsulate these geometric properties.

Now let us consider two consecutive frames of the image sequence  $I(x, y, t)$  at times  $t$  and  $t + 1$ . We denote the spatiotemporal derivatives of the image sequence by  $(I_x, I_y, I_t)$ . Then, the 3D brightness constraint for rigid objects' motion can be expressed by ref. [42]

$$\Gamma(\mathbf{T}, \boldsymbol{\omega}, Z) = I_t + \mathbf{s} \cdot \frac{\mathbf{T}}{Z} + \mathbf{q} \cdot \boldsymbol{\omega} = 0 \quad (6)$$

where  $\mathbf{s}$  and  $\mathbf{q}$  are two quantities expressed in terms of image gradients and camera intrinsic parameters. Then, the problem of learning jointly the depth  $Z$  and egomotion, parameterized by translational and rotational motions  $(\mathbf{T}, \boldsymbol{\omega})$ , can be stated as the following loss minimization

$$L_2 = \sum_{x,y} |\Gamma(\mathbf{T}, \boldsymbol{\omega}, Z)| + \mu |\nabla Z| \quad (7)$$

where the first term reduces the loss when the prediction deviates from the 3D brightness constraint, and the second term stands for smoothing depth to avoid both overfitting and the trivial null solution. Minimization of the loss  $L_2$  will overcome all the above shortcomings related to minimizing  $L_1$  and by adding other constraints instead.

## 5. Depth from stereo and multi-view

Deep learning methods have been widely employed for depth estimation from stereo or multiview images. These methods leverage convolutional neural networks (CNNs) to learn the mapping between input image pairs or sets and their corresponding depth maps. DispNet is a popular deep-learning architecture specifically designed for stereo depth estimation. It consists of a CNN-based encoder-decoder network that takes a stereo pair of images as input and predicts a dense disparity map, which can be converted to depth. The network is trained using a supervised learning framework with ground truth depth maps. Pyramid Stereo Matching Network (PSMNet) is another deep learning architecture for stereo depth estimation. It introduces a spatial pyramid pooling module to capture multi-scale information and a stacked hourglass network structure to refine the disparity estimation. PSMNet has demonstrated excellent performance in stereo depth estimation tasks. MC-CNN is a deep learning method that takes advantage of multiple views of a scene to estimate depth. It takes a set of calibrated images as input and processes them through a shared CNN architecture to predict the depth map. MC-CNN exploits the inter-view geometric relationships to improve depth estimation accuracy. Generative Adversarial Networks (GANs) have also been utilized for depth estimation from stereo or multiview images. GAN-based methods often involve training a generator network to generate depth maps from input images and a discriminator network to distinguish between real and synthesized depth maps. This adversarial training helps improve the quality and realism of the predicted depth maps. The disparity estimation method in [41] uses a CNN network for computing matching distances between image patches followed by a cross-based aggregation to compute the disparity map. In ref. [43], a CNN was trained in a supervised way to estimate disparity between stereo images from stereo video datasets. An implementation in GPU was presented in ref. [44] to learn feature correspondences faster. In ref. [45], stereo matching is enhanced using conditional random fields (CRF) to improve the accuracy and coherence of the depth estimates. CRF is a probabilistic graphical model that models the dependencies between variables in a structured manner. In the context of Semi-Global Matching (SGM) [46], the spatial-variant penalty parameters were learned by regularization terms applied to the disparity map to enforce smoothness and coherence. SGM employs a penalization approach where the disparity differences between neighboring pixels are penalized and controlled by the penalty parameters. In ref. [47], a CNN method with differentiable layers was presented that learns an end-to-end mapping from an image pair to disparity map. A refinement by adding another CNN stage was presented in ref. [48].

In ref. [49], a method was presented to train a CNN network that performs end-to-end unsupervised depth estimation with a training loss that enforces left-right depth consistency inside the network. Similarly, the method in ref. [50] learns self-supervised stereo matching as finding the disparity map that best warps between the stereo image pair. In ref. [51], CNN architecture was proposed to jointly unsupervise

Method		Error				Accuracy $\delta$			
		AbsRel	SqRel	RMS	RMSlog	$D1_{all}$	$<1.25$	$<1.25^2$	$<1.25^3$
Supervised	Mayer et al. [43]	—	—	—	—	4.34%	—	—	—
	Wang et al. [45]	—	—	—	—	4.32%	—	—	—
	Zbontar et al. [41]	—	—	—	—	3.89%	—	—	—
	Seki et al. [46]	—	—	—	—	3.09%	—	—	—
	Kendall et al. [47]	—	—	—	—	2.87%	—	—	—
	Luo et al. [43]	—	—	—	—	2.56%	—	—	—
	Pang et al. [48]	—	—	—	—	<b>2.67%</b>	—	—	—
Unsupervised	Zhong et al. [50]	—	—	—	—	3.57%	—	—	—
	Godard et al. [49]	0.148	1.344	5.927	0.247	—	0.803	0.922	0.964
	Yang et al. [52]	0.109	1.004	6.232	0.203	—	0.853	0.937	0.975
	Liu et al. [51]	0.051	0.532	3.780	0.126	—	0.957	0.982	0.991
	Wang et al. [53]	<b>0.049</b>	<b>0.515</b>	<b>3.404</b>	<b>0.121</b>	<b>5.943%</b>	<b>0.965</b>	<b>0.984</b>	<b>0.992</b>
	Jiao et al. [54]	<b>0.049</b>	<b>0.522</b>	<b>3.461</b>	<b>0.120</b>	—	<b>0.961</b>	<b>0.984</b>	<b>0.992</b>

**Table 4.**

Result comparisons of depth evaluation from stereo-based methods on the KITTI dataset. Best performance is marked with bold fonts.

learning optical flow and stereo depth map. By jointly estimating optical flow and stereo depth using unsupervised deep learning like in refs. [52, 53], the network can exploit the shared features and dependencies between the two tasks, leading to improved performance compared to separate estimation methods. Exploiting segmentation in the context of stereo motion learning can lead to further improvements in-depth estimation as in ref. [54]. **Table 4** summarizes the evaluation of depth estimation from state-of-the-art stereo-motion-based methods using KITTI [14] dataset.

## 5.1 Loss function

When learning depth from stereo or multi-view images, we can use a loss function that compares the predicted depth map with the ground truth depth map derived from stereo or multi-view disparity information. One commonly used loss function is the smooth L1 loss, which is defined as:

$$L_{1;smooth} = \begin{cases} |Z(x,y) - Z^*(x,y)| - \frac{\alpha}{2} & \text{if } |Z(x,y) - Z^*(x,y)| > \alpha \\ \frac{1}{2\alpha}(Z(x,y) - Z^*(x,y))^2 & \text{otherwise} \end{cases} \quad (8)$$

where  $Z(x,y)$  is the depth map predicted by the model and  $Z^*(x,y)$  is the ground truth depth map derived from stereo or multi-view disparity information. The smooth L1 loss provides a balance between the L1 loss (absolute difference) and the L2 loss (squared difference). It reduces the impact of outliers while still providing gradient information for training. In stereo depth estimation, the ground truth depth map can

be obtained by converting the disparity map (derived from stereo correspondence) to depth using the camera parameters and baseline distance. The disparity map represents the horizontal pixel shifts between corresponding points in the stereo images. For multi-view depth estimation, we can use multiple views (more than two) of the same scene to derive the ground truth depth map by triangulation. By estimating the disparity or correspondence between each view and a reference view, we can triangulate the 3D points and obtain the ground truth depth map. Additionally, we can incorporate other regularization terms or constraints into the loss function to further improve the depth estimation. Some common techniques include incorporating geometric consistency, enforcing smoothness or sparsity, or leveraging semantic information. Remember that the choice of the loss function and additional constraints may vary depending on the specific requirements and characteristics of stereo or multi-view depth estimation task.

## **6. Conclusion**

Deep learning methods have significantly advanced the field of depth estimation by providing effective approaches for inferring depth from various types of input data. They are highly data-driven and excel in learning complex patterns and representations from large-scale datasets, enabling them to capture intricate depth cues and generalize well to different scenes and scenarios. Deep learning allows for end-to-end learning, where the model learns to directly predict depth from input data, such as monocular images, stereo pairs, or multi-view images. This eliminates the need for explicitly designing handcrafted features or intermediate steps in the depth estimation pipeline. In the case of monocular case, learning models can estimate depth from a single image, which is a challenging task due to the inherent ambiguity. Despite the limitations, many approaches have achieved impressive results by leveraging large-scale annotated datasets and incorporating various techniques like multi-scale processing, context aggregation, and geometric constraints. Deep learning methods have also shown remarkable success in-depth estimation from stereo and multi-view images. By utilizing the correspondence or disparity information between multiple views, deep models can leverage geometric constraints to provide accurate depth estimation. The choice of loss functions and regularization techniques plays a crucial role in training deep learning models for depth estimation. Common loss functions include mean squared error (MSE) loss, smooth L1 loss, and photometric loss. Regularization techniques like smoothness regularization, geometric consistency, and semantic guidance can further enhance the quality of depth estimation. Pretrained models on large-scale datasets, such as KITTI, have been successfully applied to depth estimation tasks, while transfer learning allows leveraging the knowledge learned from a source task to improve performance on a target depth estimation task with limited data. The choice of deep learning methods for depth estimation depends on the specific application requirements. Factors such as real-time performance, accuracy, robustness to noise and occlusions, and memory efficiency need to be considered when selecting or designing deep learning models for depth estimation. Overall, deep learning methods have revolutionized depth estimation by providing powerful techniques that can learn depth from different input modalities, generalize well to diverse scenes, and achieve state-of-the-art performance. Ongoing research continues to refine and enhance these methods, making depth estimation an active and evolving area of study.

## **Acknowledgements**

This project was supported in part by collaborative research funding from the National Research Council of Canada's Artificial Intelligence for Logistics Program.

IntechOpen

## **Author details**

Hicham Sekkati<sup>1\*</sup> and Jean-Francois Lapointe<sup>2</sup>


1 National Research Council of Canada, Montreal, Canada

2 National Research Council of Canada, Ottawa, Canada

\*Address all correspondence to: [hicham.sekkati@nrc-cnrc.gc.ca](mailto:hicham.sekkati@nrc-cnrc.gc.ca)

## **IntechOpen**

---

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Calvi GM, Moratti M, O'Reilly GJ, Scattarreggia N, Monteiro R, Malomo D, et al. Once upon a time in Italy: The tale of the Morandi bridge. *Structural Engineering International*. 2019;**29**(2): 198-217, Taylor Francis
- [2] Saxena A, Sun M, Ng AY. Make3D: Learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2009;**31**(5):824-840
- [3] Zhang R, Tsai PS, Cryer JE, Shah M. Shape-from-shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1999;**21**(8): 690-706
- [4] Pentland AP. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. July 1987;**9**(4):523-531
- [5] Kulkarni JB, Sheelarani CM. Generation of Depth Map Based on Depth from Focus: A Survey, *International Conference on Computing Communication Control and Automation - ICCUBEA*. Pune, India: IEEE Xplore; 2015. pp. 716-720
- [6] Srinivasan P, Garg R, Wadhwa N, Ng R, Barron J. Aperture Supervision for Monocular Depth Estimation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR*. Salt Lake City, UT, USA; 2018. pp. 6393-6401
- [7] Suwajanakorn S, Hernandez C, Seitz SM. Depth from Focus with your Mobile Phone, *IEEE Conference on Computer Vision and Pattern Recognition - CVPR*. Boston, MA, USA; 2015
- [8] Hazirbas C, Soyer SG, Staab MC, Leal-Taixé L, Cremers D. Deep depth from focus. In: *14th Asian Conference on Computer Vision - ACCV*. Perth, Australia: ACCV; 2018
- [9] Gur S, Wolf L. Single Image Depth Estimation Trained Via Depth from Defocus Cues, *IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR*. Long Beach, CA, USA; 2019
- [10] Maximov M, Galim K, Leal-Taixé L. Focus on Defocus: Bridging the Synthetic to Real Domain Gap for Depth Estimation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR*. Seattle, WA, USA; 2020
- [11] Won C, Jeon HG. Learning Depth from Focus in the Wild, *17th European Conference Computer Vision – ECCV*. Tel Aviv, Israel; 2022
- [12] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale Deep Network. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - NIPS14*. Vol. 2. Montreal, Canada: NIPS14; 2014. pp. 2366-2374
- [13] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor Segmentation and Support Inference from RGBD Images. *Firenze, Italy: Computer Vision – ECCV*; 2012
- [14] Geiger A, Lenz P, S. C, Urtasun R. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*. 2013;**32**(11):1231-1237
- [15] Liu F, Shen C, Lin G. Deep Convolutional Neural Fields for Depth Estimation from a Single Image, *arXiv*, 2014

- [16] Saxena A, Chung SH, Andrew YN. 3-D depth reconstruction from a single still image. *International Journal of Computer Vision - IJCV*. 2008;**76**(1): 53-69
- [17] Wang P, Shen X, Lin Z, Cohen S, Price B, Yuille A. Towards unified depth and semantic prediction from a single image. In: *IEEE Conference on Computer Vision and Pattern Recognition - CVPR*. Boston, MA; 2015
- [18] Zoran D, Isola P, Krishnan D, Freeman WT. Learning ordinal relationships for mid-level vision. In: *IEEE International Conference on Computer Vision - ICCV*. Santiago, Chile; 2015
- [19] Chakrabarti A, Shao J, Shakhnarovich G. Depth from a Single Image by Harmonizing Overcomplete Local Network, *arXiv-CoRR*. 2016
- [20] Anirban R, Sinisa T. Monocular depth estimation using neural regression Forest. In: *IEEE Conference on Computer Vision and Pattern Recognition - CVPR*. Las Vegas; 2016
- [21] Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N. Deeper Depth Prediction with Fully Convolutional Residual Networks, *3DV*. California, USA; 2016
- [22] Cao Y, Wu Z, Shen C. Estimating Depth from Monocular Images as Classification Using Deep Fully Convolutional Residual Networks, *arXiv-CoRR*. 2016
- [23] Xu D, Ricci E, Ouyang W, Wang X, Sebe N. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation. In: *IEEE Conference on Computer Vision and Pattern Recognition - CVPR*. Honolulu, HI, USA; 2017
- [24] Anwar S, Hayder Z, Porikli F. Depth estimation and blur removal from a single out-of-focus image. In: *British Machine Vision Conference - BMVC*. London, UK; 2017
- [25] Carvalho M, Le Saux B, Trounev-Peloux P, Almansa A, Champagnat F. Deep Depth from Defocus: How Can Defocus Blur Improve 3D Estimation Using Dense Neural Networks. Munich, Germany: *ECCV*; 2018
- [26] Zhang A, Sun J. Joint depth and defocus estimation from a single image using physical consistency. *IEEE Transactions on Image Processing*. 2021; **30**:3419-3433
- [27] Vijayanarasimhan S, Ricco S, Schmid C, Sukthankar R, Fragkiadaki K. SfM-Net: Learning of Structure and Motion from Video, *arXiv-CoRR*. 2017
- [28] Kumar ACS, Bhandarkar SM, Prasad M. DepthNet: A recurrent neural network architecture for monocular depth prediction. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops - CVPRW*. Salt Lake City, USA; 2018
- [29] Fischer P, Dosovitskiy A, Ilg E, Hausser P, Hazirbas C, Golkov V. FlowNet: Learning Optical Flow with Convolutional Networks, *arXiv-CoRR*, 2015
- [30] Garg R, Kumar BG, Carneiro G, Reid I. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue, *Computer Vision – ECCV*. 2016
- [31] Zhou T, Brown M, Snavely N, Lowe DG. Unsupervised learning of depth and ego-motion from video. In: *IEEE Conference on Computer Vision and Pattern Recognition - CVPR*. Honolulu, HI, USA; 2017



- [32] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: IEEE Conference on Computer Vision and Pattern Recognition - CVPR. Salt Lake City, Utah, USA; 2018
- [33] Yin Z, Shi J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: IEEE Conference on Computer Vision and Pattern Recognition - CVPR. Salt Lake City, Utah, USA; 2018
- [34] Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition - CVPR. Honolulu, HI, USA; 2017
- [35] Zou Y, Luo Z, Huang JB. Df-Net: Unsupervised Joint Learning of Depth and Flow Using Cross-Task Consistency, in European Conference on Computer Vision - ECCV. Munich, Germany; 2018
- [36] Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, et al. FlowNet. In: Learning Optical Flow with Convolutional Networks, IEEE International Conference on Computer Vision - ICCV. Santiago, Chile; 2015
- [37] Casser V, Pirk S, Mahjourian R, Angelova A. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Thirty-Third Conference on Artificial Intelligence - AAAI. Honolulu, Hawaii, USA; 2019
- [38] Chen Y, Schmid C, Sminchisescu C. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: IEEE/CVF International Conference on Computer Vision -ICCV. Seoul, Korea; 2019
- [39] Ranjan A, Jampani V, Balles L, Kim K, Sun D, Wulff J, et al. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Long Beach, California, USA; 2019
- [40] Sekkati H, Lapointe J-F. Back to old constraints to jointly supervise learning depth, camera motion and optical flow in a monocular video. In: IEEE International Conference on Image Processing -ICIP. Bordeaux, France; 2022
- [41] Žbontar J, LeCun Y. Stereo matching by training a convolutional neural network to compare image patches. arXiv-CoRR. 2016
- [42] Sekkati H, Mitiche A. Joint dense 3d interpretation and multiple motion segmentation of temporal image sequences: A variational framework with active curve evolution and level sets. In: International Conference on Image Processing -ICIP. Singapore; 2004
- [43] Mayer N, Ilg E, Hausser P, Fischer P, Cremers D, Dosovitskiy A, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Las Vegas, Nevada, USA; 2016
- [44] Luo W, Schwing AG, Urtasun R. Efficient deep learning for stereo matching. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Las Vegas, NV, USA; 2016
- [45] Wang Z, Zhu S, Li Y, Cui Z. Convolutional neural network based deep conditional random fields for stereo matching. Journal of Visual

Communication and Image Representation. 2016;**40**:739-750

[46] Seki A, Pollefeys M. SGM-nets: Semi-global matching with neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Honolulu, HI, USA; 2017

[47] Kendall A et al. End-to-end learning of geometry and context for deep stereo regression. In: IEEE International Conference on Computer Vision -ICCV. Venice, Italy; 2017

[48] Pang J, Sun W, Ren JSJ, Yang C, Yan Q. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: IEEE International Conference on Computer Vision Workshop -ICCVW. Seoul, Korea; 2019

[49] Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Honolulu, Hawaii, USA; 2017

[50] Zhong Y, Dai Y, Li H. Self-supervised learning for stereo matching with self-improving ability. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Honolulu, Hawaii, USA; 2017

[51] Liu L, Zhai G, Ye W, Liu Y. Unsupervised Learning of Scene Flow Estimation Fusing with Local Rigidity, 28th International Joint Conference on Artificial Intelligence -IJCAI. Macao, China; 2019

[52] Yang Z, Wang P, Wang Y, Xu W, Nevatia R. Every Pixel Counts: Unsupervised Geometry Learning with Holistic 3d Motion Understanding, in European Conference on Computer Vision -ECCVW. Germany: Munich; 2018

[53] Wang Y, Wang P, Yang Z, Yi CL. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In: IEEE Conference on Computer Vision and Pattern Recognition -CVPR. Long Beach, California; 2019

[54] Jiao Y, Tran T, Shi G. EffiScene: Efficient per-pixel rigidity inference for unsupervised joint learning of optical flow, depth. In: Camera Pose and Motion Segmentation, IEEE/CVF Conference on Computer Vision and Pattern Recognition -CVPR. Nashville, TN, USA; 2021