Self-Supervised Monocular Image Depth Learning and Confidence Estimation

Long Chen[a], Wen Tang[a], Tao Ruan Wan[b], Nigel W. John[c]

[a] *Bournemouth University, Poole, UK*
[b] *University of Bradford, Bradford, UK*
[c] *University of Chester, Chester, UK*

Abstract

We present a novel self-supervised framework for monocular image depth learning and confidence estimation. Our framework reduces the amount of ground truth annotation data required for training Convolutional Neural Networks (CNNs), which is often a challenging problem for the fast deployment of CNNs in many computer vision tasks. Our DepthNet adopts a novel fully differential patch-based cost function through the Zero-Mean Normalized Cross Correlation (ZNCC) to take multi-scale patches as matching and learning strategies. This approach greatly increases the accuracy and robustness of the depth learning. Whilst the proposed patch-based cost function naturally provides a 0-to-1 confidence, it is then used to self-supervise the training of a parallel network for confidence map learning and estimation by exploiting the fact that ZNCC is a normalized measure of similarity which can be approximated as the confidence of the depth estimation. Therefore, the proposed corresponding confidence map learning and estimation operate in a self-supervised manner and is a parallel network to the DepthNet. Evaluation on the KITTI depth prediction evaluation dataset and Make3D dataset show that our method outperforms the state-of-the-art results.

Keywords: Monocular Depth Estimation, Deep Convolutional Neural Networks, Confidence Map

Email addresses: alwaysunny@gmail.com (Long Chen), wtang@bournemouth.ac.uk (Wen Tang)

1. Introduction

The human vision system is amazingly complex and extremely delicate. It can perceive depth through stereopsis, which relies on the displacement of the same object between the images received by the left and right retinas [1]. With extensive visual experience and through trial and error, humans develop the ability to use contextual depth cues to achieve good and reliable perception of depth and better understanding of spatial structure. Among these depth cues, most of them do not rely on stereopsis (the perception of depth from binocular vision), such as object occlusion, perspective, familiar and relative size, depth from motion, lighting and shading. Therefore, if blind in one eye or if performing a monocular task such as endoscopic surgery, we can still judge distance from these many different intuitive depth cues. In contrast, when using machine vision it is hard to infer the non-stereopsis depth cues.

With the recent development of Deep Convolutional Neural Networks (DCNNs), machines can solve many computer vision problems when provided with very large human annotated datasets such as ImageNet [2], which is known as supervised learning. Acquisition of labelled datasets is one of the biggest challenges for supervised learning, however, which is an expensive, time-consuming and labour-intensive task.

In this paper, we propose a novel self-supervised computational framework that mimics the process of how a human learns varies of contextual depth cues from stereopsis. We propose to "teach" the neural networks to "learn" the depth

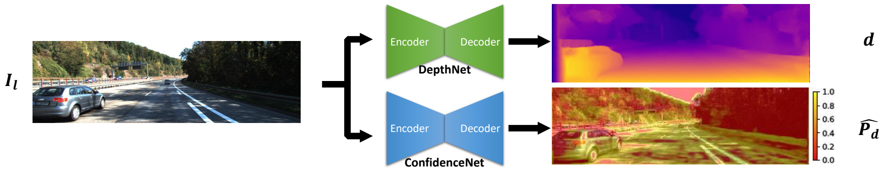


Figure 1: Our proposed framework can simultaneously estimate depth and the confidence of estimated depth.

by themselves from ”looking” to stereo image pairs. To be more specific, we construct a patch-based loss function that leverages the epipolar constraint [3] of stereo vision to minimize the depth prediction error from the context of a single image for each training iteration. Our approach does not require the ground truth depth for supervised training. Instead, we derive the implicit function of estimating depth from monocular images by the epipolar constraint of the stereo image pair, which is very easy to acquire compared with the ground truth depth that can only be obtained from LiDAR measurements. Therefore, our method can be regarded as self-supervised learning.

Compared with previous work [4] [5] [6] addressing the same problem, we propose a novel patch-based depth learning strategy, inspired by the classic patch matching algorithms for finding the best-matched patches between the left and right images. We use the Zero-Mean Normalized Cross Correlation (ZNCC) to measure the normalized similarities between these patches. A fully-differential patch-based ZNCC cost function is implemented to guide the depth synthesis process for more accurate and robust results. Visual assessment shows that our approach can produce more accurate and reliable depth estimations in both texture-rich and texture-less areas due to the enlargement of matching field from a pixel to a patch (see Figure 5). Empirical evaluations on KITTI dataset demonstrate the effectiveness of our approach and produce a state-of-the-art performance in monocular depth estimation task.

Our second contribution is that we train a parallel DCNN to evaluate the performance of the monocular depth estimation which can output a 0 to 1 confidence map. The parallel DCNN is also trained in a self-supervised manner thanks to our ZNCC similarity measurement function. As ZNCC is a normalized measure of similarity, which can be approximated as the confidence of the depth estimation, we take the ZNCC loss to self-supervise the parallel DCNN (ConfidenceNet) during training so that we can estimate the confidence of the depth estimated from the first DCNN (DepthNet) during testing mode as shown in Figure 1. A confidence map is extremely useful for the monocular depth estimation task trained in an unsupervised manner, as the learned epipolar constraint only works well when there are clear corresponding pixels between the image pairs; it will fail and produce uncertain depth when occlusion and specularity exist in the images. Our confidence map can give a real-time assessment of the reliability of the predicted depth, which can then be further integrated into many applications such as monocular dense reconstruction [7], SLAM-based depth fusion [8], and many tasks need crucial accurate and confidence such as the monocular endoscopic surgery and the perception task for self-driving.

2. Related Work

2.1. Stereo Depth Estimation

The problem of stereo images depth estimation has been well studied for a long time [9] [10]. With the theory of epipolar constraint, accessing depth from stereo images can be regarded as a well-posed problem when ignoring the occlusions and depth discontinuities. Many stereo vision algorithms managed to achieve comparable results to ground truth depth acquired from depth sensors [11] [12].

2.2. Monocular Depth Estimation

In contrast, estimating depth from monocular images is an ill-posed problem that is inherently ambiguous [13], and many research efforts have been devoted to the problem of monocular image depth estimation. One of the classic methods is Shape from Shading (SFS) [14], which is based on the gradual variation of shading as a cue to estimate the shape and depth. However, SFS has a strict prior assumption of Lambertian reflectance, uniform color and texture, and fixed light source direction, which are not applicable to most of the images in the real world. Saxena *et al* [15][16][17][18] used Markov Random Field (MRF) incorporated with multiscale image features to learn monocular cues in a supervised manner. However, the hand-craft local features used in these approaches limit the expressive power of supervised learning, and lack a global contextual understanding of the scene for learning consistent depth.

2.3. DCNNs based Monocular Depth Learning

More recently, DCNNs [13] [19] are introduced to solve the challenge of monocular depth estimation problem, and has pushed the state-of-the-art forward in this area. Building on the success of this approach, several improvements have been made by incorporating probabilistic models such as Conditional Random Fields (CRFs)[20] [21] [22] [23] [24], advanced network structures such as Resnet [25], fully convolutional Resnet [26], two-streamed networks [27], multi-task joint training [28] [19] [29] [30] [31] and novel loss functions such as sparse semi-supervision [32] [33], relative depth [34] [35] and depth as classification [26]. Impressive as these works are, ground-truth depth data are still needed for the supervision of training these DCNNs. Recently,

2.4. Unsupervised Monocular Depth Learning

Driven by DCNNs, view synthesis technology [36] has proven to be effective on synthesizing new views by sampling pixels from existing views [37] [38], which enables novel frameworks of unsupervised learning of monocular depth from stereo pairs, e.g., Deep3D [39], Garg *et al* [4]. The works by Godard *et al* [5] and Zhou *et al* [6] advanced the networks by incorporating left-right consistency and pose estimations. Further improvements including introducing Visual Odometry (VO) or Multi-View Stereo (MVS) to learn depth from monocular videos [40] [41] [42] [43]. However, a common weakness of these approaches is the use of pixel-wised photometric loss (L1-norm) to construct loss functions to guide the back-propagation process. Gradients are derived from the pixel intensity difference [6], which will lead to ambiguous gradients in texture-less areas and also in the regions that contain the mixture of thin structures and texture-less areas. Although multi-scale and smoothness loss functions are used to prevent such issue [4] [5] [6], the results are still not desirable and gradients are still likely to converge to local minimums due to the ambiguous pixel-wise loss. As shown in Figure 5, in a common speed limitation board area from the KITTI dataset, the direct pixel-wise photometric loss will lead to many local minimums shown in the right curve chart. While as the left curve chart shows the result of using our proposed patch-based ZNCC loss, the loss is more smooth and likely to converge to the global minimum in the epipolar line. And the experiment result (the last row in Figure 5) shows our proposed method can effectively generate accurate depth in complex regions.

2.5. Novelty Compared to Previous Work

We propose a novel multi-scale patch-based cost function that adopts the ZNCC as a similarity function to explicitly enlarge the matching field and increase the matching robustness. From another point of view, our proposed patch-based cost function implicitly integrates the classic Patch Matching (PM) algorithm as a minimization problem in our loss function. Although Garg *et al* [4] have discussed a straightforward idea of using the stereo matching algorithm as a pre-processing method to generate "quasi ground-truth" depth for training, their result is not desirable due to the poor quality of "quasi ground-truth". Similarly, Guo *et al* [44] proposed a more advanced method by training a proxy stereo network from synthetic, then fine tuned it on real data, and finally used it to train a monocular network. Due to the good quality of the fine tuned stereo network, the distilled monocular network can achieve good results. In contrast, Luo *et al* [45] also proposed a similar framework that firstly use a DCNN to synthesize stereo pairs from single images, and then use conventional stereo matching to get depth for monocular depth training. Essentially different from these works which separate the stereo matching with monocular depth learning, we treat the stereo matching as a minimization problem and implement a fully differential Patch-Matching algorithm as a cost function that is seamlessly integrated into our neural network. As the loss of the PM cost function can be passed through the whole network during a backward propagation, our network can produce more robust and consistent depth by large-scale self-supervised training, which will not be limited by the performance of off-the-shelf stereo matching algorithms.

Another novelty of our work is the confidence map. As monocular depth estimation itself is an ill-posed problem, although learning-based approaches

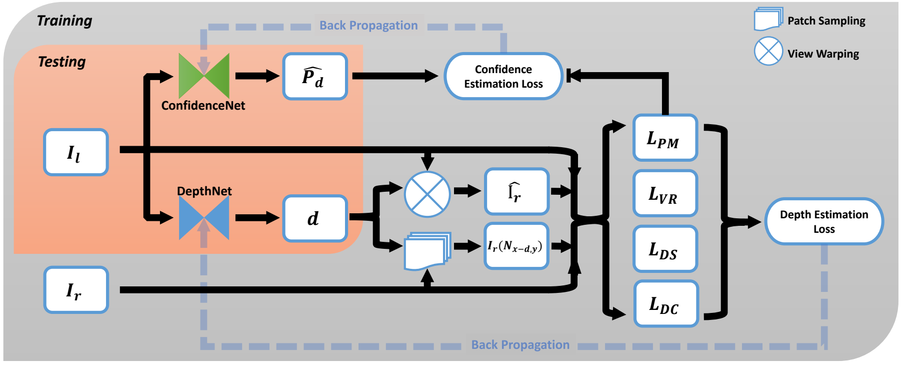


Figure 2: Framework for proposed self-supervised monocular depth learning and confidence estimating networks.

achieve comparable results to stereo depth estimation, there are still many unavoidable mistakes in the predicted depth map. For the first time, our method is able to provide a pixel-wise confidence of the predicted depth by using a parallel DCNN to capture and learn the confidence during training. The confidence map will greatly improve the usability of deploying monocular depth estimation into many practical tasks.

3. Method

3.1. Framework Overview

Figure 2 illustrates the entire framework for our self-supervised monocular depth learning and confidence estimation networks. Since the ground-truth depth D_{gt} is absent for supervised training, we treat the monocular depth estimation as a problem of image synthesis error minimization during training. Specifically, during training, we use the left images I_l of the stereo pairs to synthesize per-pixel depth D using an encoder-decoder network $D = F_{depth}(I_l, \theta)$, which is converted into disparities maps d by the Equation 2. The disparities map d is then used to guide the stereo view reconstruction $\hat{I}_r = F_{warp}(I_l, d)$ and the sampling of patches $N_{x-d,y} = F_{sample}(I_r, d)$. After that, the loss function L_{total} is calculated based on Patch Matching Loss L_{PM}, View Reconstruction

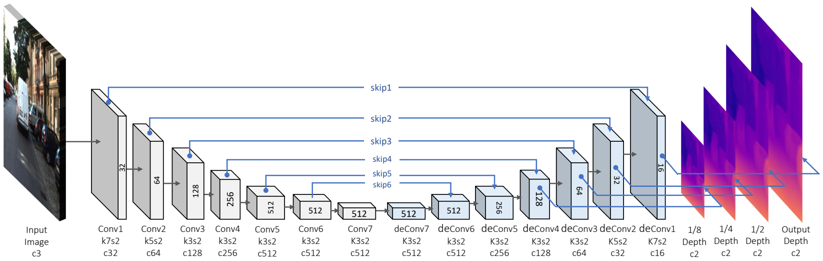


Figure 3: Depth synthesis network structure. "k" is the kernel size, "s" for the stride, "c" for the channel number. For simplicity, we do not draw the conv layers after each conv and deconv layer, which have the same kernel and channel size as previous layers but with stride 1.

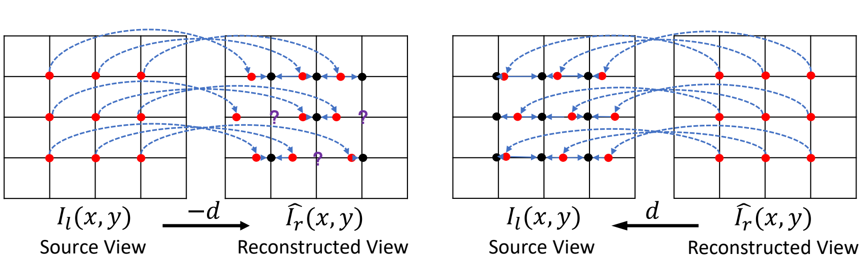
Loss L_{VR}, Disparity Smoothness Loss L_{DS}, and Disparity Consistency Loss L_{DC}. As these processes are differentiable, back propagation can be used to update the parameters θ of our depth learning network to minimize the total loss L_{total}.

$$\begin{aligned}\frac{\partial L_{total}}{\partial \theta} =& \frac{\partial L_{PM} + \partial L_{VR} + \partial L_{DS} + \partial L_{DC}}{\partial F_{warp}(I_l, d) + \partial F_{sample}(I_r, d)} \times \frac{\partial F_{warp}(I_l, d) + \partial F_{sample}(I_r, d)}{\partial d} \\ & \times \frac{\partial d}{\partial D} \times \frac{\partial D}{\partial F_{depth}(I_l, \theta)} \times \frac{\partial F_{depth}(I_l, \theta)}{\partial \theta}\end{aligned} \tag{1}$$

Since our patch-based ZNCC loss map $L_{PM}(x, y)$ represents the normalized inverted similarity between each pixel of the I_l and I_r, it can be approximated as the inverted confidence of the depth estimation result. We use the $L_{PM}(x, y)$ to self-supervise the training of a second encoder-decoder network – ConfidenceNet to generate the confidence $\hat{P}_d$ of the per-pixel depth estimation of our DepthNet.

3.2. Depth Synthesis Network

The core part of our framework is the depth synthesis and generation. Our goal is to learn an implicit function F_{depth} that estimates a per-pixel depth from a single input image. Inspired by the architectures of FlowNet [46], DispNet [47] and the network of Godard *et al* [5] and Zhou *et al* [6], we employ a VGG-like



(a) Foward mapping

(b) Backward mapping

Figure 4: The difference between forward mapping and backward mapping.

fully convolutional neural network architecture [48] in order to generate per-pixel depth from a single image. Our encoder-decoder model is illustrated in Figure 3. The input image is encoded by 7 conv layers with stride 2 each followed by a conv layer with stride 1, which efficiently compress the input image into a feature tensor with $1/2^7$ original size and 512 channels. Then, the feature tensor is up-sampled by 7 deConv layers with stride 2 each followed by a conv layer with stride 1, which decode the feature tensor into a full original size depth. Following the method in [46], 6 skip connections are implemented for preserving high-level information to ensure the high quality per-pixel prediction after up-sampling. Multi-scale depth images are outputted and used for further steps to constraint the network for a coarse-to-fine up-sampling.

3.3. Warping-based Stereo View Reconstruction

View warping is an enabling technology for self-supervised learning framework [4] [5] [6]. Given the per-pixel disparity map estimated from a single image in the previous step, the target view of the stereo pairs can be reconstructed by the epipolar relationship in stereo vision. According to the epipolar constraint: the projection of a pixel x_l on the right camera plane x_r must be contained in the epipolar line. For calibrated stereo pairs discussed in this paper, x_l and x_r must be in the same row y, and the disparity d describes the horizontal displacement of the corresponding pixels x_l and x_r . Through the stereo triangulation, we can get that

$$D_{xy} = \frac{bf}{d} \Rightarrow d = x_l - x_r = \frac{bf}{D_{xy}} \tag{2}$$

where D_{xy} is the depth estimated in the pixel at (x, y), b and f are the camera baseline and focal distance. By the relationship discussed in the above equation, the target view in a stereo pair can be reconstructed given the source view and the corresponding depth (estimated through our depth synthesis network).

However, the direct mapping from one known view to the other view (forward mapping) will result in holes in the target image that are not differentiable. Therefore, we use the inverse mapping: for each pixel in the target view, by picking points from the source to reconstruct the target view guided by the d. Thus, a complete and differentiable target view can be generated. Then the bilinear sampling [49] is used to get the interpolated pixel value from the source view.

3.4. Disparity-guided Patch Sampling

Inspired by the stereo view reconstruction described above, we propose a novel patch sampling process guided by the estimated disparity from our DepthNet. $N_{x,y}$ is defined as a patch with window size n, centered at the coordinate (x, y). We sample patches on each pixel in the left image $\{x, y \in I_l | I_l(N_{x,y})\}$, and the corresponding patches shifted by disparity values d of each pixel in the right image, $\{x, y \in I_r | I_r(N_{x-d,y})\}$. According to Equation 2, if d is correct, then we have $I_l(N_{x,y}) = I_r(N_{x-d,y})$. And this relationship will be used to construct the patch matching loss. These sampled patches are computed and stored vectorized so that can be deployed parallelly on GPU for accelerated computation.

The patch sampling size is very important and can affect the final performance of similarity measurement. However, there is no optimal patch size and the performance varies greatly across different images and local details. When small patch size is used, little information will be captured, and the similarity comparison robustness will be decreased. If we use a large patch size, computational complexity will be greatly increased and also cannot recover accurate depth at stereo occlusion and depth discontinuous. Therefore, we use a multi-scale patch sampling scheme and sample a combination of 4 different patch sizes in an image to fully exploit the effects of different patch sizes. We will discuss the choice of patch sizes in Section 4.1.3.

3.5. Loss Function Construction

We define a loss function L_{total} with multiple strategies to effectively train our networks for accurate, smooth and realistic depth.

$$L_{total} = \omega_p L_{PM} + \omega_v L_{VR} + \omega_d L_{DS} + \omega_c L_{DC} \tag{3}$$

where from left to right is: Patch Matching Loss, View Reconstruction Loss, Disparity Smoothness Loss and Disparity Consistency Loss. ω is the corresponding weights to balance the effects of gradients back propagation. Each loss function will be explained in details below:

3.5.1. Patch Matching Loss

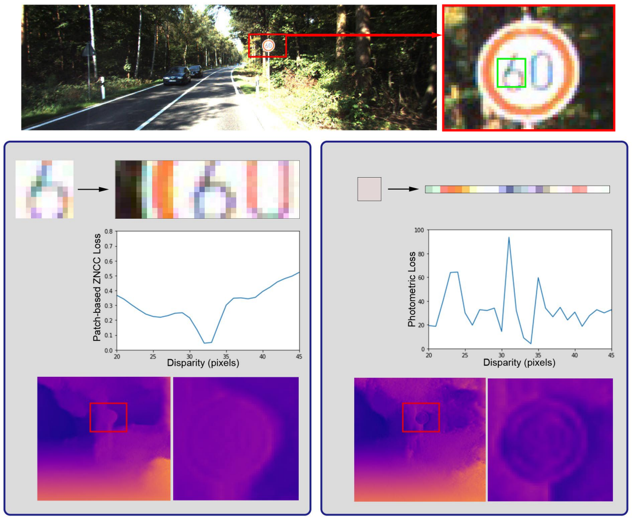
Inspired by patch matching algorithm that by finding the best-matched patches in the left and right image to get correct disparities. We propose a patch matching loss that maximize the similarities (minimize the differences) of patches in left image $I_l(N_{x,y})$ and the shifted patches in right image $I_r(N_{x-d,y})$ to get correct disparities. Here, the ZNCC measure of similarity is used to compute a normalized similarity between the patches $I_l(N_{x,y})$ and $I_r(N_{x-d,y})$:

$$C_{ZNCC}\left(I_l(N_{x,y}), I_r(N_{x-d,y})\right) = \frac{\sum_{i,j \in N_{x,y}} \left(I_l(i,j) - \bar{I}_l(N_{x,y})\right) \cdot \left(I_r(i-d,j) - \bar{I}_r(N_{x-d,y})\right)}{\sqrt{\sum_{i,j \in N_{x,y}} \left(I_l(i,j) - \bar{I}_l(N_{x,y})\right)^2 \cdot \sum_{i,j \in N_{x,y}} \left(I_r(i-d,j) - \bar{I}_r(N_{x-d,y})\right)^2}} \tag{4}$$

where $\bar{I}\left(N_{x,y}\right) = \frac{1}{n} \sum_{x,y \in N_{x,y}} I\left(x,y\right)$ is the mean intensity of the patch $N_{x,y}$ centered at the coordinate (x, y).

The ZNCC returns a similarity ranging from $[-1, 1]$. We first normalize it into $[0, 1]$ then invert it to get the patch matching loss:

$$L_{PM} = \sum_{x,y} 1 - \frac{1 + C_{ZNCC}\left(I_l(N_{x,y}), I_r(N_{x-d,y})\right)}{2} \tag{5}$$



Proposed Patch-based ZNCC Loss

Photometric Loss used in previous works

Figure 5: Comparison of our proposed patch-based ZNCC loss (left image) with the photometric loss used in previous works (right image) to demonstrate that a patch naturally encodes more information than a single pixel and our loss function is more smooth and convex than other methods, therefore is more likely to converge to global minimum in the epipolar line.

Our patch matching loss is computed at all 4 patch sizes to cover both small structures and large areas. There are several advantages of using our patch-based ZNCC loss to regularize the depth synthesis:

(1) Our patch matching loss uses patches for measurement that involve larger regions than the direct pixel-wise photometric loss used in previous work, which is more robust and can achieve sub-pixel accuracy. Figure 5 demonstrates the effect of our patch-based ZNCC loss. We charted the values of our patch-based ZNCC loss and the photometric loss against the disparity value of a pixel located at the center of the image patch ”6”. It is obvious that by using our proposed patch-based ZNCC loss, the loss is more smooth and likely to converge to the global minimum. Whereas the direct pixel-wise photometric loss will lead to many local minimums shown in the right curve chart.

(2) Compared to other similarity measures such as Sum of Absolute Differences (SAD), Census, and Normalized Cross Correlation (NCC), ZNCC is

especially robust against Gaussian noise and variation between the compared patches, which can help to recover more accurate depth in our self-supervised framework.

(3) As a zero-mean normalized similarity measurement function, our patch-based ZNCC loss can provide a similar value ranging from $[-1, 1]$. After normalized to $[0, 1]$ as shown in Equation 5, it can be regarded as the confidence of the generated depth at each pixel, which can be further used to self-supervise the training of our confidence network.

3.5.2. View Reconstruction Loss

We use the view reconstruction loss as a second supervision on the depth synthesis. Guided by the synthesized depth, the right views can be reconstructed by collecting pixels from left images. The view reconstruction loss is defined as the L1 loss between the reconstructed view $\hat{I}_r$ and the original view I_r:

$$L_{VR} = \sum_{xy} \left| I_r(x, y) - \hat{I}_r(x, y) \right| \tag{6}$$

Compared to the patch matching loss, the view reconstruction L1 loss is more sensitive to small structures and depth discontinuities and can provide more detailed depth information.

3.5.3. Disparity Smoothness Loss

We use a disparity smoothness term to regularize our network to produce more smooth depth. Similar to [4] [5] [6], we use the sum of the L1 norm of the disparity gradients along the x and y directions as a smoothness factor. The edge-aware terms are used to reduce the penalty on edges where depth discontinuities usually happen, which can prevent over-smoothing.

$$L_{DS} = \frac{1}{XY} \sum_{x,y} \left| \frac{\partial d(x, y)}{\partial x} \right| e^{-\| \frac{\partial I(x,y)}{\partial x} \|} + \left| \frac{\partial d(x, y)}{\partial y} \right| e^{-\| \frac{\partial I(x,y)}{\partial y} \|} \tag{7}$$

3.5.4. Disparity Consistency Loss

The left-right disparity consistency loss proposed in [5] has achieved a great improvement for monocular depth generation. Here, we adopt this loss function into our framework. The left and right image disparities are both generated, and the difference of left disparity map and the reconstructed left disparity map from right disparity is computed and minimized. This loss will ensure the left and right disparities coherence.

$$L_{DC} = \frac{1}{XY} \sum_{x,y} |d_l(x, y) - d_r(x - d_l(x, y), y)| \tag{8}$$

3.6. Confidence Estimation Network

One of the advantages of our proposed patch matching loss is that a normalize similarity measurement can be generated for each pixel at the training time. With the well-known epipolar constraint, the per-pixel confidence of the estimated depth can be approximated as the normalized similarity measurement of the left patches and the corresponding patches in the right image.

$$P_d(x, y) \approx C_{Normalized}(I_l(N_{x,y}), I_r(N_{x-d,y})) = (1 - L_{PM}(x, y)) \tag{9}$$

Here, we propose to use another encoder-decoder network to learn the confidence map generated by our depth estimation network during training, so that the confidence map can be preserved and generated during the testing time. We tried to train the confidence and depth in one network like [28] [19] [29] [30], but the multi-task training would reduce the depth estimation performance. Therefore, we use a parallel encoder-decoder network to learn the confidence supervised by the per-pixel ZNCC loss of our depth estimation network. The loss of our ConfidencNet is shown below:

$$L_{ConfidenceNet} = \sum_{x,y} \left| (1 - L_{PM}(x, y)) - \hat{P}_d(x, y) \right| \tag{10}$$

where $\hat{P}_d(x, y)$ is the generated confidence map, $L_{PM}(x, y)$ is the patch matching loss from our depth estimation network described in above sections. The static copy is used here to prevent the gradients propagating back to the depth estimation network. The $1 - L_{PM}(x, y)$ operation inverts the loss to confidence, and L1 loss is used to access the confidence estimation error.

Instead of using the same encoder-decoder network structure as our DepthNet, we employ a simpler structure by only using first 5 conv-layer and last 5 deconv-layer without skip layers as described in Figure 3 for two reasons:

(1) To reduce memory usage and training time, as training two neural networks at the same time is very computationally expensive. The second network can be replaced by a deeper and more complex encoder-decoder network to produce sharper and more accurate confidence, but the main purpose of our work is to prove that our self-supervised monocular depth learning and confidence estimation framework is feasible and helpful for depth prediction, hence we choose to use a simple network structure as the proof of concept.

(2) We intend to use a simpler network with fewer weights to prevent overfitting to noises and to learn more generic confidence – high confidence in texture-rich areas, low confidence in texture-less, blurry and occluded areas, which is what we design this confidence net for.

4. Experiments

In this section, we evaluate our framework and compare the results with prior approaches both quantitatively and qualitatively on KITTI dataset. We use the rectified stereo image pairs for training our networks. For testing time, we use the left image to generate depth, and the corresponding sparse LIDAR data is served as the ground truth for benchmarking.

4.1. Implementation Details

Our networks are implemented in Tensorflow and trained on a workstation with a single Nvidia Titan X GPU (12G Memory). Our models take around 60 hours to train for 50 epochs. When in testing mode, our networks can output depth and confidence map at around 20 frames per second.

4.1.1. Hyper Parameters

All input images are scaled to 512x256 with a batch size of 4. Adam Optimizer is used with $\beta_1 = 0.9$, $\beta_1 = 0.999$, and initial learning rate $\lambda = 0.0001$ that decays after half of the training process. The weights to construct our total loss function for depth estimation network are $w_p = 0.5$,$w_v = 1$,$w_d = 0.1$,$w_c = 1$.

4.1.2. Data Augmentation

The same data augmentation approach in [5] is used to randomly flip the image and change the gamma, brightness, and color shifts to increase the network robustness and prevent over-fitting.

4.1.3. Multi-scale Implementation

We employ a multi-scale strategy to ensure a coarse-to-fine up-sampling. As can be seen from Figure 3, 4 depth scales are outputted with $1/8, 1/4, 1/2$ and a full resolution. All of our loss functions are computed for each of these 4 scales, and for each of left and right images/disparities. We take the means of these loss functions as the final loss.

4.1.4. Patch Size

By applying different patch sizes on different image scales, we can get very large equivalent patch sizes with less computation. For patch size choices, based on our empirical test, we use $n = 5, 5, 7, 9$ pixels for our patch-based ZNCC loss on 4 different scales, which is equivalent $n = 5, 10, 28, 72$ pixels' windows on full resolution images.

4.2. Training dataset

To be able to compare with the state-of-the-art monocular depth learning approaches, we trained and evaluated our networks using two different train/test splits: *Godard* and *Eigen*.

4.2.1. Godard Split

We use the same train/test sets that Godard *et al* [5] proposed in their work. 200 high quality disparity images in 28 scenes provided by the official KITTI training set are served as the ground truth for benchmarking. For the rest of 33 scenes with a total of 30,159 images, 29,000 images are picked for training and the remaining 1,159 images for testing.

4.2.2. Eigen Split

For fair comparison with more previous works, we also use the test split proposed by Eigen *et al* [13] that has been widely evaluated by the works of Garg *et al* [4], Liu *et al* [23], Zhou *et al* [6] and Godard *et al* [5]. This test split contains 697 images of 29 scenes. The rest of 32 scenes contain 23,488 images, in which 22,600 are used for training and the remaining for testing, similar to [4] and [5].

4.3. Results

4.3.1. Quantitative Evaluation

Evaluation Metrics. To access the quantitative performance of our proposed depth prediction network and compare with previous works, we evaluate each method using several error and accuracy metrics from [13] [5] [4] [6]. The error metrics we use include Absolute Relative Difference (AbsRel), Squared Relative Difference (SqRel), Root Mean Square Error (RMSE) and Root Mean Squared Logarithmic Error (RMSElog). The accuracy metrics [4] [23] that we use are the percentages of estimated depth d_{est} that subject to

$$max(\frac{d_{est}}{d_{gt}}, \frac{d_{gt}}{d_{est}}) = \delta < threshold \tag{11}$$

Results on KITTI dataset. The evaluation results on the KITTI dataset are reported in Table 1. We use different combinations of train/test splits (E for Eigen, G for Godard) and cap distances (80m and 50m) to compare with different works. For Eigen *et al* [13], Liu *et al* [23], Zhou *et al* [6] and Godard *et al* [5] , the Eigen split with 80m cap distance are used. For Garg *et al* [4], Zhou *et al* [6] and Godard *et al* [5], the Eigen split with 50m cap distance are used. We also report our result on Godard split with 80m cap. For the ablation study of the ZNCC loss, we have implemented a patch-based Sum of Absolute Differences (SAD) loss that is a common and basic similarity measurement used for stereo matching algorithm to replace the ZNCC loss and keep the same multi-level patch setting. The results for the multi-level patch-based SAD loss are reported as ours-SAD, which shows that our dedicated multi-level patch-based loss with

Table 1: Comparison with state-of-the-art methods on KITTI dataset.

Method	Super-vision	Split	Cap	Error (Lower better)					Accuracy (Higher better)		
				AbsRel	SqRel	RMSE	RMSElog	D1-all	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen *et al* [13]	Yes	E	80	0.203	1.548	6.307	0.282	-	0.702	0.890	0.958
Liu *et al* [23]	Yes	E	80	0.201	1.584	6.471	0.273	-	0.680	0.898	0.967
Zhou *et al* [6]	No	E	80	0.208	1.768	6.856	0.283	-	0.678	0.885	0.957
Godard *et al* [5]	No	E	80	0.148	1.344	5.927	0.247	-	0.803	0.922	0.964
ours-SAD	No	E	80	0.147	1.302	5.901	0.246	-	0.805	0.922	0.964
ours-ZNCC	**No**	**E**	**80**	**0.145**	**1.267**	**5.786**	**0.244**	**-**	**0.811**	**0.925**	**0.965**
Garg *et al* [4]	No	E	50	0.169	1.080	5.104	0.273	-	0.740	0.904	0.962
Zhou *et al* [6]	No	E	50	0.201	1.391	5.181	0.264	-	0.696	0.900	0.966
Godard *et al* [5]	No	E	50	0.140	0.976	4.471	0.232	-	0.818	0.931	0.969
ours-SAD	No	E	50	0.140	0.959	4.463	0.232	-	0.821	0.931	0.969
ours-ZNCC	**No**	**E**	**50**	**0.138**	**0.937**	**4.399**	**0.231**	**-**	**0.825**	**0.933**	**0.969**
Godard *et al* [5]	No	G	80	0.124	1.388	6.125	0.217	30.272	0.841	0.936	0.975
ours-SAD	No	G	80	0.121	1.358	6.073	0.215	29.937	0.842	0.936	0.976
ours-ZNCC	**No**	**G**	**80**	**0.117**	**1.202**	**5.953**	**0.210**	**29.612**	**0.845**	**0.938**	**0.976**

SAD similarity measurement can already improve the benchmark results, but more improvements came with our proposed multi-level patch-based loss using the advanced ZNCC similarity measurement (reported as ours-ZNCC), which achieved the state-of-the-art results for monocular depth estimation problem on KITTI dataset.

Results on Make3D dataset. To further access the generalization ability of our proposed methods and compare with other methods, we also evaluate our trained networks on Make3D dataset [18]. For supervised methods [50] [21] [25], they are trained using ground truth depth data from the Make3D training set. For unsupervised methods [6] [5] and ours, are trained on KITTI + Cityscapes datasets without the presence of any image from Make3D dataset. For evaluation, we measure the error metrics (AbsRel, SqRel, RMSE and RMSElog) using the test image with ground truth from Make3D dataset. As can be seen from Table 4.3.1, although our method scored similar results to Zhou *et al* [6] regarding relative errors, for the RMSE, our methods outperform all of the state-of-the-art unsupervised methods.

Table 2: Comparison with state-of-the-art methods on Make3D dataset [18].

Method	Supervision	Cap	Error (Lower better)			
			AbsRel	SqRel	RMSE	RMSElog
Karsch *et al*[50]	Yes	70	0.428	5.079	8.389	0.149
Liu *et al*[21]	Yes	70	0.475	6.562	10.05	0.165
Laina *et al* [25]	Yes	70	0.204	1.840	5.683	0.084
Zhou *et al* [6]	No	70	0.383	5.321	10.47	0.478
Godard *et al* [5]	No	70	0.544	10.94	11.76	0.193
Ours	**No**	**70**	**0.393**	**5.714**	**8.908**	**0.186**

Compared among unsupervised methods, our method produced better results regarding RMSE (RMSE and RMSElog) and at large cap distance (70m and 80m), and not significantly improve the relative error metrics (AbsRel, SqRel) at small cap distance (50m). This is totally what we expect as our multi-scale patch-based loss function performs better results when the distances of left-right corresponding pixels are large (meaning the pixel is at large distance), which the pixel-based loss function will prone to fail.

4.3.2. Qualitative Evaluation

The qualitative comparison to some of the related methods on KITTI dataset is shown in Figure 6. While our network structure is similar to that of Godard *et al*[5], both generate clear and accurate depth than other works. We also provide a detailed comparison with the results of Godard *et al*[5] in the lower part of Figure 6. Our network can generate more accurate depth in complex regions with thin structures and texture-less areas such as the pillars and traffic signs. This verified the theory we explained in Figure 5 that our patch-based loss function is more robust and easier to converge to the global minimum in complex regions.

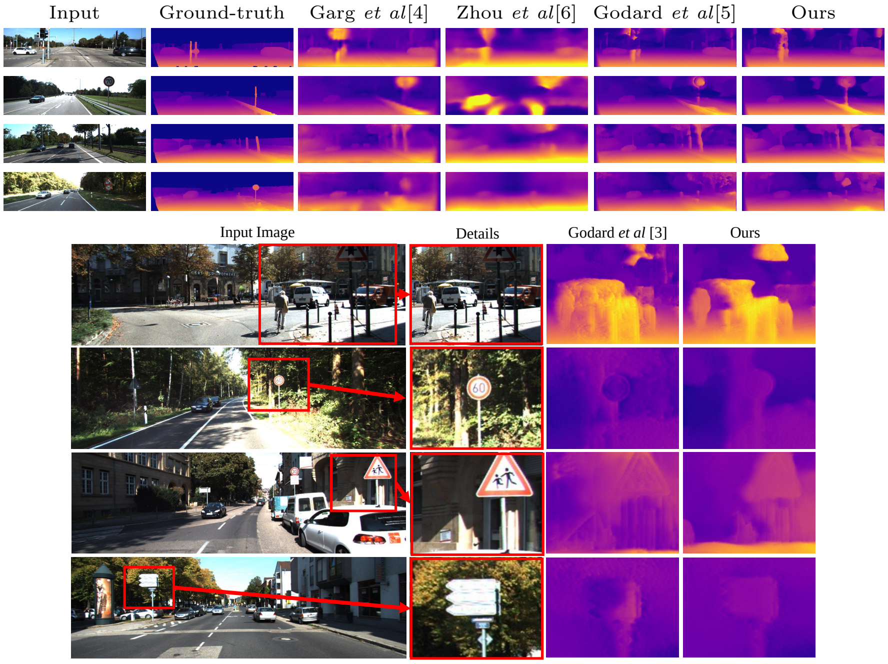


Figure 6: Upper part: comparison of monocular depth estimation on KITTI dataset between Garg *et al*[4], Zhou *et al*[6], Godard *et al*[5], and ours. Lower part: comparison of details with Godard *et al*[5]. All of the results are generated using authors' provided pre-trainned model. The ground-truth depth map is interpolated from sparse point map only for visualization.

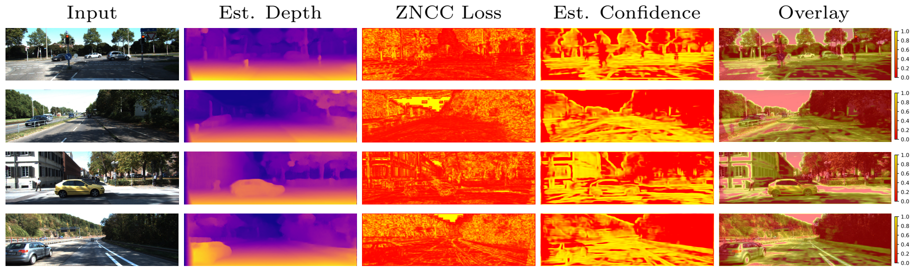


Figure 7: Confidence estimation results. A colorbar from red to yellow is used to represent 0 to 1.

4.3.3. Confidence Map Evaluation

We show the confidence estimation results in Figure 7. A colorbar from red to yellow is used to represent 0 to 1. We can see that the estimated confidence can nicely represent the inverted ZNCC loss but less noisy due to the small network we use to prevent over-fitting. The overlaid confidence on input image shows that our ConfidenceNet has learned to generate confidence from contextual information. For example, in texture-less areas (sky, building), dark areas (trees under shadow), occluded areas (around thin structures) and reflective areas (car window), the estimated confidence is usually very low, while the texture-rich areas and edges usually have high confidence.

5. Discussion

In this paper, we have presented a novel self-supervised framework for monocular depth learning and confidence estimation. We incorporate the patch matching theory into a fully differential DCNN and achieve self-supervised training of both depth and the confidence of depth. Our proposed loss function exploits the epipolar constraint of stereo vision and also provides a normalized similarity that is further used to supervise the confidence estimation. Our method not only outperforms the state-of-the-art results on the KITTI benchmark evaluation, but also for the first time, we are able to simultaneously generate depth from monocular images and estimate the confidence of the generated depth. This is a step change for monocular depth estimation as it significantly increases the feasibility of using monocular depth estimation into many practical applications such as autonomous driving and monocular endoscopic surgery [7], where the accuracy of estimated depth is crucial.

Why Our ConfidenceNet Works? As there is certain limitation of unsupervised monocular depth learning from stereo pairs (ambiguous depth estimation in texture-less area, reflection, etc.). Our ConfidenceNet is supervised by the per-pixel ZNCC loss of our depth estimation network (which can be regarded as the confidence of current depth), it explicitly learns the regions where our depth estimation network performs well and badly. But on a deeper level, our ConfidenceNet actually implicitly learns the inherent defect of the patch matching algorithm – it would fail on texture-less regions and performs badly near stereo view occlusions, reflections and blurred areas. Therefore, after sufficient training steps, our ConfidenceNet can capture and memory where the DepthNet would perform good or bad, and give an estimation of the confidence of our DepthNet, although they are two different networks.

In Future Work. We will continue optimizing our model and explore the possibility of using adaptive window size for patch sampling to decrease the training time and increase accuracy in small structures.

6. Acknowledgments

This research was supported/partially supported by the EU H2020 Programme, H2020-MSCA-RISE-2018:iGame project 82387.

References

[1] B. J. Dunkin, C. Flowers, 3d in the minimally invasive surgery (mis) operating room: Cameras and displays in the evolution of mis, in: Imaging and Visualization in The Modern Operating Room, Springer, 2015, pp. 145–155.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.
URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[3] Z. Zhang, Determining the epipolar geometry and its uncertainty: A review, International Journal of Computer Vision 27 (2) (1998) 161–195. doi:10.1023/A:1007941100561.
URL https://doi.org/10.1023/A:1007941100561

[4] R. Garg, V. K. B.G., G. Carneiro, I. Reid, Unsupervised cnn for single view depth estimation: Geometry to the rescue, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 740–756.

[5] C. Godard, O. M. Aodha, G. J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6602–6611. doi:10.1109/CVPR.2017.699.

[6] T. Zhou, M. Brown, N. Snavely, D. G. Lowe, Unsupervised learning of depth and ego-motion from video, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6612–6619. doi:10.1109/CVPR.2017.700.

[7] L. Chen, W. Tang, N. W. John, T. R. Wan, J. J. Zhang, Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality, Computer Methods and Programs in Biomedicine 158 (2018) 135 – 146. doi:https://doi.org/10.1016/j.cmpb.2018.02.006.
URL http://www.sciencedirect.com/science/article/pii/S0169260717301694

[8] K. Tateno, F. Tombari, I. Laina, N. Navab, Cnn-slam: Real-time dense monocular slam with learned depth prediction, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6565–6574. doi:10.1109/CVPR.2017.695.

[9] S. T. Barnard, M. A. Fischler, Computational stereo, ACM Comput. Surv. 14 (4) (1982) 553–572. doi:10.1145/356893.356896.
URL http://doi.acm.org/10.1145/356893.356896

[10] D. Scharstein, R. Szeliski, R. Zabih, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, in: Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), 2001, pp. 131–140. doi:10.1109/SMBV.2001.988771.

[11] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 328–341. doi:10.1109/TPAMI.2007.1166.

[12] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, End-to-end learning of geometry and context for deep stereo regression, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 66–75. doi:10.1109/ICCV.2017.17.

[13] D. Eigen, C. Puhrsch, R. Fergus, 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 - Volume 2, NIPS'14,

MIT Press, Cambridge, MA, USA, 2014, pp. 2366–2374.
URL http://dl.acm.org/citation.cfm?id=2969033.2969091

[14] R. Zhang, P.-S. Tsai, J. E. Cryer, M. Shah, Shape-from-shading: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (8) (1999) 690–706. doi:10.1109/34.784284.

[15] A. Saxena, S. H. Chung, A. Y. Ng, Learning depth from single monocular images, in: Y. Weiss, B. Schölkopf, J. C. Platt (Eds.), Advances in Neural Information Processing Systems 18, MIT Press, 2006, pp. 1161–1168.
URL http://papers.nips.cc/paper/2921-learning-depth-from-single-monocular-images.pdf

[16] A. Saxena, J. Schulte, A. Y. Ng, Depth estimation using monocular and stereo cues, in: Proceedings of the 20th International Joint Conference on Artifical Intelligence, IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, pp. 2197–2203.
URL http://dl.acm.org/citation.cfm?id=1625275.1625630

[17] A. Saxena, S. H. Chung, A. Y. Ng, 3-d depth reconstruction from a single still image, International Journal of Computer Vision 76 (1) (2008) 53–69. doi:10.1007/s11263-007-0071-y.
URL https://doi.org/10.1007/s11263-007-0071-y

[18] A. Saxena, M. Sun, A. Y. Ng, Make3d: Learning 3D scene structure from a single still image, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (5) (2009) 824–840. doi:10.1109/TPAMI.2008.132.

[19] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2650–2658. doi:10.1109/ICCV.2015.304.

[20] B. Li, C. Shen, Y. Dai, A. van den Hengel, M. He, Depth and surface normal estimation from monocular images using regression on deep

features and hierarchical crfs, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1119–1127. doi:10.1109/CVPR.2015.7298715.

[21] M. Liu, M. Salzmann, X. He, Discrete-continuous depth estimation from a single image, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 716–723. doi:10.1109/CVPR.2014.97.

[22] Y. Hua, H. Tian, Depth estimation with convolutional conditional random field network, Neurocomputing 214 (2016) 546–554. doi:10.1016/j.neucom.2016.06.029.

[23] F. Liu, C. Shen, G. Lin, I. Reid, Learning depth from single monocular images using deep convolutional neural fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (10) (2016) 2024–2039. doi:10.1109/TPAMI.2015.2505283.

[24] D. Xu, E. Ricci, W. Ouyang, X. Wang, N. Sebe, Multi-scale continuous crfs as sequential deep networks for monocular depth estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 161–169. doi:10.1109/CVPR.2017.25.

[25] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, in: 3D Vision (3DV), 2016 Fourth International Conference on, 2016, pp. 239–248. doi:10.1109/3DV.2016.32.

[26] Y. Cao, Z. Wu, C. Shen, Estimating depth from monocular images as classification using deep fully convolutional residual networks, IEEE Transactions on Circuits and Systems for Video Technology PP (99) (2017) 1. doi:10.1109/TCSVT.2017.2740321.

[27] J. Li, R. Klein, A. Yao, A two-streamed network for estimating fine-scaled depth maps from single rgb images, in: 2017 IEEE International

Conference on Computer Vision (ICCV), 2017, pp. 3392–3400. doi:10.1109/ICCV.2017.365.

[28] L. Ladický, J. Shi, M. Pollefeys, Pulling things out of perspective, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), CVPR '14, IEEE Computer Society, Washington, DC, USA, 2014, pp. 89–96. doi:10.1109/CVPR.2014.19.
URL http://dx.doi.org/10.1109/CVPR.2014.19

[29] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. Yuille, Towards unified depth and semantic prediction from a single image, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2800–2809. doi:10.1109/CVPR.2015.7298897.

[30] A. Mousavian, H. Pirsiavash, J. Košecká, Joint semantic segmentation and depth estimation with deep convolutional networks, in: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE, 2016, pp. 611–619.

[31] H. Yan, S. Zhang, Y. Zhang, L. Zhang, Monocular depth estimation with guidance of surface normal map, Neurocomputing 280 (2018) 86–100. doi:10.1016/j.neucom.2017.08.074.

[32] Y. Kuznietsov, J. Stckler, B. Leibe, Semi-supervised deep learning for monocular depth map prediction, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2215–2223. doi:10.1109/CVPR.2017.238.

[33] F. Ma, S. Karaman, Sparse-to-dense: Depth prediction from sparse depth samples and a single image, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 1–8. doi:10.1109/ICRA.2018.8460184.

[34] D. Zoran, P. Isola, D. Krishnan, W. T. Freeman, Learning ordinal relationships for mid-level vision, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 388–396. doi:10.1109/ICCV.2015.52.

[35] W. Chen, Z. Fu, D. Yang, J. Deng, Single-image depth perception in the wild, in: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29, Curran Associates, Inc., 2016, pp. 730–738.
URL http://papers.nips.cc/paper/6489-single-image-depth-perception-in-the-wild.pdf

[36] A. Fitzgibbon, Y. Wexler, A. Zisserman, Image-based rendering using image-based priors, in: 2003 IEEE International Conference on Computer Vision (ICCV), 2003, pp. 1176–1183 vol.2. doi:10.1109/ICCV.2003.1238625.

[37] T. Zhou, S. Tulsiani, W. Sun, J. Malik, A. A. Efros, View synthesis by appearance flow, in: European Conference on Computer Vision, 2016.

[38] J. Flynn, I. Neulander, J. Philbin, N. Snavely, Deep stereo: Learning to predict new views from the world's imagery, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5515–5524. doi:10.1109/CVPR.2016.595.

[39] J. Xie, R. Girshick, A. Farhadi, Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, 2016, pp. 842–857.

[40] C. Wang, J. Miguel Buenaposada, R. Zhu, S. Lucey, Learning depth from monocular videos using direct methods, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[41] R. Li, S. Wang, Z. Long, D. Gu, Undeepvo: Monocular visual odometry through unsupervised deep learning, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 7286–7291.

[42] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, I. Reid, Unsupervised learning of monocular depth estimation and visual odometry with

deep feature reconstruction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 340–349.

[43] Z. Li, N. Snavely, Megadepth: Learning single-view depth prediction from internet photos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2041–2050.

[44] X. Guo, H. Li, S. Yi, J. Ren, X. Wang, Learning monocular depth by distilling cross-domain stereo networks, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 506–523.

[45] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, L. Lin, Single view stereo matching, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[46] A. Dosovitskiy, J. T. Springenberg, M. Tatarchenko, T. Brox, Learning to generate chairs, tables and cars with convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4) (2017) 692–705. doi:10.1109/TPAMI.2016.2567384.

[47] N. Mayer, E. Ilg, P. Husser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4040–4048. doi:10.1109/CVPR.2016.438.

[48] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4) (2017) 640–651. doi:10.1109/TPAMI.2016.2572683.

[49] M. Jaderberg, K. Simonyan, A. Zisserman, k. kavukcuoglu, Spatial transformer networks, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 2017–2025.

URL http://papers.nips.cc/paper/5854-spatial-transformer-networks.pdf

[50] K. Karsch, C. Liu, S. B. Kang, Depth transfer: Depth extraction from video using non-parametric sampling, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (11) (2014) 2144–2158. doi:10.1109/TPAMI.2014.2316835.