Detail-preserving depth estimation from a single image based on modified fully convolutional residual network and gradient network

Huihui Xu¹ · Nan Liu¹

Received: 14 May 2021 / Accepted: 18 November 2021

Published online: 29 November 2021

© The Author(s) 2021

Abstract

Predicting a convincing depth map from a monocular single image is a daunting task in the field of computer vision. In this paper, we propose a novel detail-preserving depth estimation (DPDE) algorithm based on a modified fully convolutional residual network and gradient network. Specifically, we first introduce a new deep network that combines the fully convolutional residual network (FCRN) and a U-shaped architecture to generate the global depth map. Meanwhile, an efficient feature similarity-based loss term is introduced for training this network better. Then, we devise a gradient network to generate the local details of the scene based on gradient information. Finally, an optimization-based fusion scheme is proposed to integrate the depth and depth gradients to generate a reliable depth map with better details. Three benchmark RGBD datasets are evaluated from the perspective of qualitative and quantitative, the experimental results show that the designed depth prediction algorithm is superior to several classic depth prediction approaches and can reconstruct plausible depth maps.

Article Highlights

- We changed the classic network and loss function to obtain the global 3D depth information of the scene.
- A depth gradient acquisition scheme is designed to generate the local details of the scene.
- We can obtain a plausible depth map with better depth details through our developed depth and depth gradients fusion strategy.

Keywords Depth prediction · Fully convolutional residual network · Gradient network · Depth gradient

1 Introduction

Reconstructing the 3D structure of a scene from a single RGB image has been receiving widespread attention owing to the comprehensive applications in the subject of image processing and computer vision. Precise depth information helps to understand the scene's internal

geometric structure better, which could significantly improve the performance on various existing vision tasks, such as autonomous driving, augmented reality, 3D reconstruction, semantic labeling, and pose estimation [1–3]. Traditional active depth generation methods typically use hardware devices, such as lasers and depth sensors, to simultaneously capture the RGB images and

✉ Huihui Xu, xuhuihui@mail.sdu.edu.cn | ¹School of Computer Science and Technology, Shandong Jianzhu University, No. 1000 Fengming Road, Jinan 250101, People's Republic of China.

corresponding depth maps. Nevertheless, the disadvantages of low depth resolution, short distance perception, and high system cost seriously hinder the wide application of such methods [4]. In contrast, using computational schemes to infer depth from a single RGB image provides a reasonable and low-cost way to generate the depth map.

Last few years, many researchers have explored multiple algorithms for single image depth recovery. Traditional approaches mainly focus on the use of monocular depth cues [5, 6]. However, these approaches are not suitable for all actual scenarios. In parallel, some algorithms leverage geometric assumptions [7] and additional information [8, 9] to recover the depth information have been proposed. However, these methods cannot furnish detailed depth information, and additional clues are not available.

More recently, many researchers are focusing on machine learning methods to generate reasonable depth information. Recent learning-based depth prediction algorithms from a single RGB image can be classified into non-parametric learning algorithms and parametric learning algorithms. The first category casts depth recovery as a non-parametric learning process. Non-parametric approaches can transfer depth to a single input image by leveraging a wide-ranging RGBD dataset efficiently [10–16]. For an input image, similar candidate images are first retrieved from the RGB-D database by utilizing the k-nearest neighbors algorithm. Then, the dense pixel-pixel relationship between the test image and the candidate images is established. The corresponding depths of candidate images are warped and fused by leveraging an optimization algorithm for generating the final depth. Nevertheless, when the input image has a scene structure that is not similar to the training dataset, the non-parametric approaches cannot generate convincing depth maps.

Another category is the parametric learning method. As a groundbreaking work, Saxena et al. [17, 18] learned a Markov random field (MRF) for exploring the relationship between the color image and the depth information. In parallel, many scholars also proposed other parametric learning algorithms, for instance, conditional random field (CRF) [8] and structured forest [19]. Lately, deep learning-based depth prediction methods have emerged, which has influential feature learning ability and can recover accuracy scene depth information. Meanwhile, the leverage of convolutional neural networks (CNNs) has vastly enhanced the performance of depth prediction algorithms [20–24]. With a deeper architecture, the CNN-based methods can infer the global layout of scene reliability. Nevertheless, the reconstructed depth maps lack finer depth details after many times of convolution and max-pooling, and researchers demanded these details in many computer vision tasks.

To fill the research gap mentioned above, we present an efficient framework for single image depth prediction based on modified fully convolutional residual network and gradient network in this study. The idea of this paper is to benefit from the capacity of CNN structure, but avoid the lack of depth details. Since gradient information is not very sensitive to the scene features of training data and can provide subtle details of the scene, we incorporate depth gradient information into the depth recovery process. In particular, first, an efficient deep network that combines the fully convolutional residual network (FCRN) [23] and U-net architectures [25] for recovering the global structure of a scene is proposed. Nonetheless, the depth details are imperfect, since the intermediate features are not integrated into the network. Since the depth gradients are very sparse, we designed a depth gradient generation network for capturing the local information of the depth map. In the end, an energy equation minimization scheme is proposed for integrating the complementary depth information and depth gradient information to infer a convincing depth map.

To sum up, the contribution of this paper is three-fold.

- We modified the classic network and loss function to generate the global 3D depth information of the scene.
- We designed a depth gradient acquisition scheme to generate the local details of the scene.
- Plausible depth map with better depth details can be recovered through our developed depth and depth gradients fusion strategy.

Sect. 2 describes the related works for single image depth prediction. Sect. 3 introduces the presented depth prediction algorithm. Sect. 4 summarizes all experimental results. Finally, Sect. 5 concludes this work.

2 Related work

Last few years, various methods have been exploiting how to infer depth from a single RGB image. Previous algorithms to depth prediction mainly focus on the use of monocular depth cues, for instance, defocus, saliency, atmospheric scattering, and occlusion [5, 6]. However, these kinds of methods are not suitable for all scenarios. In parallel, some schemes have been developed for estimating the depth map by utilizing geometric presumption [7]. These models can efficiently recover the global structure of a scene, but cannot obtain the desired result when the structure is complex. Meanwhile, depth prediction can be implemented by employing additional ancillary information, such as surface normal [8] and repetitive texture [9] if auxiliary cues are accessible. But, regrettably, such

supplemental details are not always available in universal scenarios.

More recently, machine learning-based depth prediction methods have been receiving researchers' great attention. We classified these approaches into two groups: one is the non-parametric learning methods and the other is the parametric learning methods.

For non-parametric learning approaches, Konrad et al. [10] put forward the first pioneering work in this field. They used the histograms of oriented gradients (HOG) descriptor to select candidate images with the photometric content that best matched the test image. Later, the median operator was leveraged to fuse candidate depths to obtain the initial depth, which was enhanced by using efficient filtering. To enhance the quality of the depth map, Karsch et al. [11] deformed the corresponding depths of candidate images based on the pixel-pixel correspondence relationship. These deformed depths were integrated by solving nonlinear optimization issues for estimating the scene depth. To reduce the dependency on similar datasets, Chios et al. [12] designed an efficient non-parametric learning method, which is named the depth analogy method. They extracted the depth gradient and used the depth from the gradient algorithm for recovering the scene depth. Meanwhile, Herrera et al. [13] leveraged features based on Local Binary Patterns to retrieve candidate images. They recovered the final depth by integrating the depths of these candidate images adaptively. These methods could be impractical for large-scale databases as they carry out a time-consuming retrieve to select the matching candidates from the training dataset. To work out this difficulty, Herrera et al. [14] presented a graduated retrieve scheme based on a clustering algorithm. More recently, Mohaghegh et al. [15] combined the global and local features of the scene and devised a modified stacked generalization framework to learn depth values of image patches. Liu et al. [16] introduced an efficient monocular depth estimation strategy via an improved segmentation scheme and a consistency constraint. However, if the candidate images and the input image have different depth values in the regions with similar appearance, the problem of depth ambiguity will arise.

For parametric learning methods, this study tried to establish the mapping model between color images and depth values based on supervised learning. Lately, deep learning algorithm has been resoundingly applied in single image depth prediction. A depth recovery scheme was first introduced by Eigen et al. [20] via utilizing CNNs. They developed a two-scale architecture (a coarse-scale network and a fine-scale network) to obtain accurate depth information. Based on this study, a multi-task learning framework including depth estimation, semantic labeling, and surface normal was presented by Eigen et al.

[21]. Moreover, a deep convolutional neural field framework was introduced by Liu et al. [22], they combined the strength of deep CNN and continuous CRF in a unified CNN framework. The residual network (Resnet) has effectively worked out the gradient vanishing issue when the network layer becomes deeper and deeper, and this network has also been used for single image depth prediction. For instance, Laina et al. [23] adopt a fully convolutional residual network (FCRN) as the encoder and four up-projection blocks as the decoder to perform up-sampling to generate a final depth map with higher resolution. An efficient convolutional neural field framework was introduced by Harsanyi et al. [24], who combined the strength of the residual network and U-nets in a deep framework. As visual attention plays an essential role in vision tasks, Chen et al. [26] designed a supervised self-attention model and utilized it to adaptively learn the task-specific similarities between different pixels to model the continuous context information. Tu et al. [27] designed an efficient monocular depth estimation model (MDE) for precise depth sensing on edge devices. They employed a reinforcement learning algorithm and automatically prune redundant channels of MDE by finding a relatively optimal pruning policy. Song et al. [28] et al. proposed a simple but effective scheme by incorporating the Laplacian pyramid into the decoder architecture. Specifically, encoded features were fed into different streams for decoding depth residuals. Ye et al. [29] proposed DPNet for high-quality monocular depth estimation. They designed an efficient non-local spatial attention module and a spatial branch (SB) to preserve spatial information.

In this study, to enhance the quality of depth maps, an efficient single image depth prediction framework based on a modified fully convolutional residual network and gradient network is developed. An improved FCRN network-based depth estimation model is designed for reconstructing the overall depth structure. Then, we develop a gradient network-based depth gradient generation scheme to enhance the global depth map. As we combine gradient information outputs with the modified FCRN for enhancement, the final estimated depth contains both global information and local details.

3 Proposed algorithm

To fill the research gap noted above, a high-quality depth generation algorithm based on a modified fully convolutional residual network and gradient network is presented. The framework makes full use of the fully convolutional residual network, U-net architecture, and gradient information of the scene. An overview of our DPDE scheme is shown in Fig. 1. Input to the algorithm is an RGB image,

Fig. 1 Overview of our presented network architecture

its output is a corresponding depth map estimated by the algorithm. In particular, first, an improved framework that integrates FCN and U-shaped architecture is introduced for predicting the depths. For single image depth prediction, the structure of FCN has been proved to have powerful estimation ability, which was further boosted by combining it with a U-shaped architecture. Since gradient information can provide subtle details of the scene, depth gradient information is incorporated into the depth estimation process. Then, a depth gradients extraction scheme is designed for generating the depth gradients along with horizontal and vertical directions. Considering the sparseness of the depth gradients and gradient information is not very sensitive to the scene features of training data, a gradient network-based method is leveraged to perform the gradient generation process. Finally, a depth and depth gradient fusion scheme is designed to enhance the global depth map locally to produce a reasonable depth map with finer details via the optimization method.

3.1 Network architecture

The proposed DPDE algorithm contains two different prediction frameworks. i.e., the deep learning-based depth prediction framework and transfer learning-based depth gradients prediction framework. This section describes the structure of a deep learning-based depth prediction network. Specifically, an improved depth prediction model which combines a fully convolutional residual network (FCN) [23] and U-shaped architecture [25] is proposed.

As the backbone of the entire depth prediction framework, a fully convolutional residual network (FCRN) was first introduced by Laina et.al [23]. The FCN model in [23] comprises two components: encoding and decoding. Specifically, ResNet50 is utilized as the basic architecture of the encoding part, and several repeated up-projection modules constitute the decoding part. The structure of ResNet-50 [30] can make the FCN network deeper and can avoid the problem of vanishing gradients. Therefore, FCN has a large receptive field. In the training process, the entire convolutional network will generate billions of parameters and occupy dozens of GB of memory. To ease this situation, FCN leverages the up-projection modules, which contain fewer parameters. The structure of FCN has been proved to have a powerful estimation ability for recovering the scene depth, nonetheless, the local details of the depth maps are imperfect as the intermediate features are not integrated into the framework. Thus, FCN was further modified by combining it with a U-shaped architecture.

Enlightened by the U-shape structure, we enhance the inter-connectivity of the model by bringing in lateral connections into the network flow rather than deepening the model by adding the network layers. These side-to-side links are established by connecting features of the same dimension between the Res-Block parts of the encoding and Up-Block parts of the decoding. Then, an additional 1×1 convolution is inserted after each depth concatenation to maintain the number of the up-projection modules' input channels constant. Figure 2 demonstrates the detailed architecture of the modified

Fig. 2 Structure diagram of the deep learning-based depth prediction method

network. Thanks to the lateral connections mentioned above, the network is capable to explore extra details in the previous features.

Horizontal and vertical gradients of the depth map contain information about meaningful depth differences in the local structure, which can be used to improve predicted depth maps for better details. Compared with the multi-scale deep network, the structure of the gradient network is relatively simple, which composes of 10 convolution layers with the 3×3 convolution kernel. Owing to the depth gradient containing positive and negative values, we use ReLU for the first nine convolution layer and do not leverage it after the convolution of the last layer. Given a single input image, the horizontal depth gradient and vertical depth gradient of the image can be extracted through the gradient network. The following experiments will prove that our gradient network can obtain reasonable local depth gradients and can enhance the depth quality.

3.2 Loss function

The loss function used in the training process takes into account the difference between the ground truth depth map d^{gt} and the predicted depth d . The definition of the loss function will affect the overall depth estimation performance.

For training our modified FCRN network, the loss L between d and d^{gt} is defined as the weighted sum of two loss functions:

$$L(d, d^{gt}) = \lambda_B L_B(d, d^{gt}) + \lambda_F L_{FSIM}(d, d^{gt}) \quad (1)$$

The first loss term L_B is the reverse Huber loss [31], which was applied as the loss function during training by combining the regularly utilized ℓ_1 loss and ℓ_2 loss, which was first proposed by Laina et al. [23]. Through leveraging ℓ_2 loss, BerHu loss can give higher weight for pixels higher residuals. In parallel, owing to the utilize of ℓ_1 loss in the training process, the BerHu loss enables slighter residuals to have a greater impact on the gradients. For estimated depth maps d and the ground truth depth maps d^{gt} , the BerHu loss $L_B(d, d^{gt})$ is defined as follows:

$$L_B(d, d^{gt}) = \begin{cases} |d - d^{gt}| & |d - d^{gt}| \leq c \\ \frac{(d - d^{gt})^2 + c^2}{2c} & |d - d^{gt}| > c \end{cases} \quad (2)$$

where $c = \frac{1}{5} \cdot \max_i |d_i - d_i^{gt}|$ indicates the index of each pixel in each depth image d . From Eq. (2) we can see that the BerHu loss is considered as the ℓ_1 norm when $|d - d^{gt}| \in [-c, c]$ and is equal to ℓ_2 when $|d - d^{gt}| > c$. The form of BerHu loss is advantageous due to the continuity and differentiability at the switch point c .

The second term L_{FSIM} utilizes feature similarity, which is a commonly leveraged metric for the task of image reconstruction. Since the maximum and the minimum values of FSIM [32] are 1 and 0, respectively, the FSIM loss is defined as follows:

$$L_{FSIM}(d, d^{gt}) = \frac{1 - FSIM(d, d^{gt})}{2} \quad (3)$$

For learning the gradient network to generate the horizontal and vertical depth gradients, we apply a new modified ℓ_1 loss term from the point of the logarithm. The pixel-level difference between the estimated depth gradients $\mathbf{G}$ and the ground truth depth gradients ∇d^{gt} is defined as follows:

$$L_g = \ln \left(\left\| \mathbf{G}_p - \nabla d_p^{gt} \right\|_1 + 1.0 \right) \quad (4)$$

where $\mathbf{G}_p$ represent the estimated depth gradients at a pixel p .

By utilizing the deep learning-based depth estimation method, we can generate the global scene depth map. Nevertheless, the reconstructed depth maps lack finer depth details after many times of convolution and max-pooling. Based on our work, depth gradient can reflect the detailed information of the scene, thus a gradient network-based depth gradient estimation framework is proposed for extracting meaningful horizontal and vertical gradient information in the next section.

3.3 Depth and depth gradients fusion

As optimization strategies have been proved to be effective in improving the details of the output map [11]. A combination scheme based on the optimization method for fusing the depth and gradient information is proposed. The optimal depth can be estimated based on the following minimization:

$$\begin{aligned} \hat{d} = \arg \min \sum_{p=1}^n \phi(d^p - d_{est}^p) \\ + \omega \sum_{p=1}^n \left[\phi(\nabla_x d^p - G_x^p) + \phi(\nabla_y d^p - G_y^p) \right] \end{aligned} \quad (5)$$

where d_{est} is the predicted depth map based on multi-scale deep network, G_x and G_y are the recovered gradients in x and y directions. ∇_x and ∇_y are x y gradient operators on final depth d . For designing the distance measure, $\phi(x)$ is used as a replacement of the ℓ_1 norm:

$$\phi(x) = \sqrt{x^2 + \varepsilon}, \varepsilon = 10^{-4} \quad (6)$$

The final depth can be generated by leveraging iteratively re-weighted least squares (IRLS) for minimizing our loss function. IRLS is a better choice to solve unconstrained, nonlinear minimization issues. IRLS works by approximating the objective and solving the problem

via minimizing the square residual; It is repeated until convergence.

4 Experimental results

4.1 Benchmark dataset

We conducted depth prediction experiments on indoor and outdoor datasets. Specifically, the proposed DPDE algorithm and other state-of-the-art algorithms are assessed on three RGBD datasets.

As an outdoor dataset, the Make3D dataset [18] is comprised of 534 color images and 534 ground truth depths. We utilize 400 images for training and 134 images for the test. For ensuring the consistency of comparison results, a bilinear interpolation algorithm is used to adjust all the images to the 460×345 pixels.

As an indoor dataset, the NYUv2 dataset [33] is comprised of 1449 color images and 1449 ground truth depths. We utilize 795 images for training and 654 images for the test. For reducing the computational complexity, all images are rescaled to 240×320 pixels via bilinear interpolation.

As an indoor testing dataset, the Middlebury stereo dataSet (MID) [34] is comprised of 35 pairs of color images and ground truth depths. Note that the disparity map is provided in the MID dataset. All original color images and disparity maps are rescaled to 380×430 pixels. We used a simple division to convert the Middlebury disparity maps into the ground truth depth map and then compared it with the predicted Middlebury depths.

4.2 Implementation details

We perform all experiments on a PC platform with Intel Core i7-10,710 CPU 1.61 GHz and 16 GB RAM and Nvidia RTX2080 8 GPU. For the deep learning-based depth generation branch, we train the model on basis of the TensorFlow framework and initialize the network utilizing the ResNet-50 parameters pre-trained on ImageNet. Then the model weights are finetuned with its Berhu loss function for 30 epochs. In our study, a stochastic gradient descent optimizer is used, and the value of momentum is set to 0.9. The learning rate is initialized to 10⁻² for all layers and halved after 10 epochs. In the last 10 epochs, it is reduced to 10⁻³. For transfer learning-based depth gradients generation branch. The parameters λ_b and λ_f were set to 1 and 1, respectively. For the depth gradients extraction branch, random initialization with Gaussian distributions was used for the gradient network. For the

depth and gradients fusion branch, the parameter ω is set to 10.

4.3 Performance evaluation

We performed the qualitative and quantitative evaluations in this section for evaluating the performance of the designed scheme.

4.3.1 Qualitative evaluation

For the outdoor scene, the qualitative evaluation between our DPDE method and other depth reconstruction methods is supplied: the Depth Transfer (DT) method [11], the Structured Forest (SF) method [19], the Deep

Convolutional Neural Field (DCNF) algorithm [22] and the FCRN algorithm [23]. The qualitative evaluation results for test images in the Make3D dataset are provided in Fig. 3. It can be seen from the figure that all the methods used for comparison can estimate reasonable depth maps, but the depths predicted by the DT [11] method seem too smooth, as shown in the third column of Fig. 3. The depth values in the sky area of the depth map predicted by the SF [19] method are not accurate as in the fourth column in Fig. 4. On the whole, the performance of the DCNF [22] and FCRN [23] methods outperform that of the DT [11] and SF [19] methods, but the local details of the depth maps estimated by DCNF [22] and FCRN [23] methods are not very clear in some areas in the fifth column and the sixth column in Fig. 3. Due to the introduction of the gradient

Fig. 3 Experimental results on Make3D dataset **a** the input image **b** the ground truth **c** DT [11] **d** SF [19] **e** DCNF [22] **f** FCRN [23] and **g** the proposed method

Fig. 4 Experimental results on NYUv2 dataset. First row: the input image, second row: the ground truth, third row: DT [11], fourth row: DA [12], fifth row: DCNF [22], sixth row: FCRN [23] and last row: the proposed method

extraction process, depth, and gradient fusion scheme, our DPDE method can show the scene details better. The depth outputs predicted by our DPDE approach seem more convincing than other results from the point of the whole structure and local details, which can be confirmed in the last column in Fig. 3.

For the indoor scene, our results are compared with other classic approaches [11, 12, 22, 23]. Figure 4 shows

the depth evaluation results on the NYU dataset. From the figures, we can see that the structures of the predicted depths are most similar to the structures of the ground truth depths. On the whole, the performance of deep learning-based depth estimation methods [22, 23] is better than that of non-parametric learning-based methods [11, 12]. Deep learning-based methods can obtain reasonable depth maps, but some areas in the depth map

Fig. 5 Experimental results on MID dataset: First row: the input image, second row: the ground truth, third row: DT [11], fourth row: DA [12], fifth row: DCNF [22], sixth row: FCRN [23] and last row: the proposed method

estimated by DCNF [22] and FCRN [23] methods seem indistinct. Due to the introduction of a multi-scale feature extraction module, feature similarity-based loss term, and gradient network, the proposed method can obtain

plausible depth outputs with better details, which can be verified in the last row in Fig. 4.

Figure 5 shows the test results on the MID dataset, which uses NYU as the training set. As demonstrated, the

depth ordering of the depth maps predicted by our DPDE algorithm seems more reasonable. Compared with the proposed DPDE method, the depth outputs predicted by the methods [11, 12, 22, 23] cannot keep the depth boundaries sharp. On the whole, the results in NYUv2 are better than the MID as the images in the MID test set have a low depth distribution correlation with the images in the NYUv2 dataset.

Furthermore, we provide the 3D reconstruction results in Fig. 6 and normalize the depth values to [0,255] for visualization. Aside from sharp depth discontinuities, the 3D reconstruction results from our method present realistic scenes compared to DCNF [22] and FCRN [23]. The effectiveness of the proposed DPDE algorithm can be confirmed by the reconstruction results. The reconstruction results from different views are shown in Fig. 7. The effectiveness of the presented DPDE algorithm can be proved by these visual results.

4.3.2 Quantitative evaluation

The similarity between the estimated depth maps d and the ground truth depths d^{gt} can be measured by the following evaluation indicators:

$$\text{Root mean squared error (RMS)}: \sqrt{\frac{1}{T} \sum_{i=1}^T (d_i - d_i^{gt})^2}.$$

$$\text{Root mean squared error (RMS (log))}: \sqrt{\frac{1}{T} \sum_{i=1}^T (\text{Log} d_i - \text{Log} d_i^{gt})^2}.$$

$$\text{Mean relative error (REL)}: \frac{1}{T} \sum_{i=1}^T \frac{\|d_i - d_i^{gt}\|_1}{d_i^{gt}}.$$

$$\text{Mean log 10 error (log 10)}: \frac{1}{T} \sum_{i=1}^T \|\log_{10} d_i - \log_{10} d_i^{gt}\|_1.$$

Threshold: th , i.e., the percentage of such that $\delta = \max\left(\frac{d^{gt}}{d}, \frac{d}{d^{gt}}\right) < th$.

The results of the presented DPDE method together with those of existing methods are shown in Table 1. Our approach outperformed the algorithms presented in [4, 11, 14, 15] on all basic measures. It is observed that our proposed method obtains the best performance on Rel and RMS, while the method in [35] achieved the best RMS and log10 values. In addition, our method provided the second-best performance for log10. Compared with the case where gradient information is not used, the proposed DPDE method with gradient stream improves log10 and Rel values by 0.005 and 0.002.

The comparison results on the NYU dataset are shown in Table 2. It is evident that our DPDE method is slightly better for some cases compared with the FCRN [23] method. The performance of RMS, log10, Rel, $\delta < 1.25$, $\delta < 1.252$ and $\delta < 1.253$ are improved by about 1.9%, 5.5%, 3%, 3%, 1.2% and 0.3% than FCRN [23], respectively. The RMS values of the presented algorithm with gradient stream increase by 0.638, 0.558, 0.345, 0.038, 0.007, 0.183, 0.478, 0.024, 0.259,

Fig. 6 The 3D reconstruction results on the NYU v2 dataset: **a** the ground truth, **b** DCNF [22], **c** FCRN [23] and **d** the proposed method

Fig. 7 The 3D reconstruction results from different views: **a** the ground truth, **b** the proposed method (original view) and **c** the proposed method (alternative view)

Table 1 Comparison results on Make3D dataset

Method	RMS	log10	Rel
DT[11]	15.1	0.148	0.361
Herrera et al.[14]	14	NR	NR
DCNF[22]	8.6	0.119	0.314
FCRN[23]	4.46	0.072	0.176
Mohaghegh et al. [15]	13.43	0.145	0.407
Liu[16]	13.8	0.138	0.346
MS-CRF [35]	4.38	0.065	0.184
Li et al.[8]	7.19	0.092	0.278
DEPT[4]	16.7	0.172	0.421
Our DPDE method (without gradient stream)	4.40	0.070	0.171
Our DPDE method	4.38	0.067	0.169

Bold indicates the best and second best results

0.262, 0.011, 0.007, 0.017 and 0.301 compared to the methods in [8, 11, 14–16, 20, 22, 23, 27, 35–39] and our DPDE method without gradient stream respectively. The proposed DPDE method attains best RMS and log10 values while the MS-CRF [35] method gets the lower scores for Rel

and log10 metrics. Our integration of local gradients enhances the performance on all metrics.

Table 3 shows the quantitative evaluation results on the MID dataset. To assist the evaluation, we chose three other indicators: structural similarity (SSIM) [41] and gradient similarity (GSM) [42], where a higher value indicates better performance for SSIM and GSM. As can be seen from the table, the method in [43] generates the best performance on RMS and Log10, while our DPDE method gets the lower Rel, SSIM, and GSM values. Thanks to the introduction of a U-shaped structure, the depth gradient extraction scheme, and the fusion strategy, our DPDE algorithm is superior to other methods for all metrics and reaches 0.684, 0.068, 0.206, 0.942, and 0.988 for RMS, log10, Rel, SSIM, GSM, and MGE metrics, respectively.

4.4 Ablation study

In this section, the larger and more challenging NYU v2 dataset is used to conduct our ablation study. The whole depth prediction processes are shown in the Fig. 8. Figure. 8a and b represent the input image and the ground truth depths respectively. Fig. 8c and d represent the FCRN and modified FCRN results respectively. Fig. 8e is the local

Table 2 Comparison results on NYUv2 dataset

Method	RMS	log10	Rel	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
DT[11]	1.20	0.131	0.350	NR	NR	NR
Herrera et al.[14]	NR	0.137	0.352	NR	NR	NR
Mohagheghet al. [15]	1.12	0.144	0.395	NR	NR	NR
Eigen [20]	0.907	NR	0.215	0.611	0.887	0.971
Carvalho et al. [36]	0.600	0.059	0.135	0.819	0.957	0.987
Moukari et al. [37]	0.569	0.057	0.133	0.830	0.966	0.993
Wang et al. [38]	0.745	NR	0.220	0.605	0.890	0.970
Liu[16]	1.04	0.126	0.353	NR	NR	NR
MS-CRF [35]	0.586	0.052	0.121	0.811	0.954	0.987
Li et al.[8]	0.821	0.094	0.232	0.621	0.886	0.968
DCNF[22]	0.824	0.095	0.230	0.614	0.883	0.971
FCRN[23]	0.573	0.055	0.127	0.811	0.953	0.988
Tu et al. [27]	0.579	NR	0.165	0.772	0.943	0.983
Ma et al. [39]	0.863	0.257	0.212	NR	NR	NR
Lo et al. [40]	0.498	0.058	0.136	0.823	0.960	0.990
Our DPDE method (without gradient stream)	0.569	0.053	0.125	0.833	0.960	0.989
Our DPDE method	0.562	0.052	0.123	0.836	0.964	0.991

Bold indicates the best and second best results

Table 3 Comparison results on MID dataset

Method	RMS	log10	Rel	SSIM	GSM
DT[11]	1.34	0.131	0.418	0.730	0.9742
DA[12]	1.32	0.152	0.524	0.734	0.9753
DCNF[22]	0.879	0.095	0.330	0.804	0.9763
FCRN[23]	0.720	0.076	0.239	0.898	0.9780
TL[43]	0.584	0.063	0.224	NR	NR
Hu et al. [44]	0.697	0.064	0.230	NR	NR
Our DPDE (without gradient stream)	0.703	0.070	0.223	0.901	0.9822
Our DPDE method	0.684	0.068	0.206	0.942	0.9880

Bold indicates the best and second best results

depth gradient and Fig. 8f is the DPDE result. As shown in Fig. 8f, when integrating with each module in turn, the scene structures are clearer and more protected better. The final depths have a high similarity to the ground truth depths and have finer details. The reason is that the global depth map reflects the overall structure while the local depth gradient can represent the subtle details of a scene. A reasonable depth map with better details can be estimated via integrating the global and local information based on the optimization approach. Therefore, the

means of extracting the depth gradient information and combining it with global depth is effective compared to only recovering the global depth.

We divide the proposed network framework into three parts and verify the effectiveness of the proposed DPDE method by gradually adding it. The quantitative comparison results of the ablation experiment are shown in Table 4. It can be seen from the table that the FCRN alone can get a more reasonable depth map, but the RMS index is higher. After modifying the FCRN by adding the U-shaped structure, the overall performance of the network has been improved. When the depth gradient module is added, the depth quality is improved significantly. For example, RMS is increased from 0.569 to 0.562, and Rel is increased by 0.002. Therefore, a complete network framework can get better estimation results.

Table 5 shows the comparison of the number of parameters contained in different models. The parameter of the FCRN model is, and the parameter of the deep gradient network is, which is one of 76 of the FCRN model parameters. From the perspective of improving the performance, the addition of the deep gradient network is reasonable, which further verifies the effectiveness of the proposed DPDE method.

Fig. 8 Depth prediction results: **a** the input image **b** the ground truth **c** FCRN result **d** modified FCRN result **e** depth gradients and **f** final result

Table 4 Quantitative results of different variants of our method

Method	RMS	log10	Rel
Baseline(FCRN)	0.569	0.053	0.125
Modified FCRN	0.567	0.053	0.124
Our full model	0.562	0.052	0.123

Table 5 Comparison of the number of parameters

Method	Number of parameters
Modified FCRN	2550×10^4
Gradient network	33.4×10^4
Ratio(Gradient network/Modified FCRN)	1/76

5 Conclusions

To reconstruct reasonable depth information from a single-color image, a novel detail preserving depth estimation (DPDE) approach that utilizes a fully convolutional residual network and gradient information was proposed. Practically, the global structure of a scene was recovered via learning the improved FCRN network. For preserving scene details, a gradient network-based depth gradient generation scheme is proposed to obtain the horizontal and vertical depth gradients. Then, an optimization scheme is proposed to integrate the local depth gradients and the global depth to generate a final depth map. Experimental results on indoor and outdoor datasets indicate that the presented depth prediction method can recover convincing depth information with better details and outperforms several classic approaches. In the future,

we plan to utilize the semantic information and surface normal information of the scene to further improve the quality of the depth map.

Acknowledgements The authors would like to thank the anonymous reviewers for their constructive comments. This research was supported in part by the Natural Science Foundation of China (No.61902221).

Funding This research was supported in part by the Natural Science Foundation of China (No.61902221).

Declarations

Conflict of interest on behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Kán P, Kaufmann H (2020) Correction to deeplight: light source estimation for augmented reality using deep learning. *Vis Comput* 36(1):229
- Fu K, Peng J, He Q, Zhang H (2021) Single image 3D object reconstruction based on deep learning: a review. *Multimed Tools Appl* 80(1):463–498

3. Yu L, Fan G (2021) DrsNet: Dual-resolution semantic segmentation with rare class-oriented superpixel prior. *Multimed Tools Appl* 80(2):1687–1706
4. Qin H, Li X, Wang Y et al (2016) Depth estimation by parameter transfer with a lightweight model for single still images. *IEEE T Circ Syst Vid* 27(4):748–759
5. Tang C, Hou C, Song Z (2015) Depth recovery and refinement from a single image using defocus cues. *J Mod Optic* 62(6):441–448
6. Yang Y, Hu X, Wu N et al (2017) A depth map generation algorithm based on saliency detection for 2D to 3D conversion. *3D Res* 8(3):1–11
7. Fouhey DF, Gupta A, Hebert M (2014) Unfolding an indoor origami world. In: *Proceedings of the European conference on computer vision*, Cham, pp 687–702
8. Li B, Shen C, Dai Y, et al (2015) Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, USA, pp 1119–1127
9. Wu C, Frahm JM, Pollefeys M (2011) Repetition-based dense single-view reconstruction. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Colorado Springs, pp 3113–3120
10. Konrad J, Wang M, Ishwar P (2012) 2d-to-3d image conversion by learning depth from examples. In: *Proceedings of the computer society conference on computer vision and pattern recognition workshops*, Rhode Island, pp 16–22
11. Karsch K, Liu C, Kang SB (2014) Depth transfer: depth extraction from video using non-parametric sampling. *IEEE T Pattern Anal* 36(11):2144–2158
12. Choi S, Min D, Ham B et al (2015) Depth analogy: data-driven approach for single image depth estimation using gradient samples. *IEEE T Image Process* 24(12):5953–5966
13. Herrera JL, Del-Bianco CR, García N (2014) Learning 3D structure from 2D images using LBP features. In: *Proceedings of the IEEE International conference on image processing*, Paris, France, 2022–2025
14. Herrera JL, Del-Bianco CR, García N (2018) Automatic depth extraction from 2D images using a cluster-based learning framework. *IEEE T Image Process* 27(7):3288–3299
15. Mohaghegh H, Karimi N, Soroushmehr SMR et al (2018) Aggregation of rich depth-aware features in a modified stacked generalization model for single image depth estimation. *IEEE T Circ Syst Vid* 29(3):683–697
16. Liu H, Lei D, Zhu Q et al (2021) Single-image depth estimation by refined segmentation and consistency reconstruction. *Signal Process-Image* 90:116048
17. Saxena A, Chung SH, Ng AY (2005) Learning depth from single monocular images. In: *Advances in neural information processing systems*, british columbia, Canada, pp 1161–1168
18. Saxena A, Sun M, Ng AY (2008) Make3d: learning 3d scene structure from a single still image. *IEEE T Pattern Anal* 31(5):824–840
19. Fang S, Jin R, Cao Y (2016) Fast depth estimation from single image using structured forest. In *IEEE International conference on image processing*. 4022–4026
20. Eigen D, Puhrsch C, Fergus R (2014) Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*
21. Eigen D, Fergus R (2015) Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *Proceedings of the IEEE International conference on computer vision*, Santiago, Chile, pp 2650–2658
22. Liu F, Shen C, Lin G et al (2015) Learning depth from single monocular images using deep convolutional neural fields. *IEEE T Pattern Anal* 38(10):2024–2039
23. Laina I, Rupprecht C, Belagiannis V, et al (2016) Deeper depth prediction with fully convolutional residual networks. In: *Proceedings of the international conference on 3D vision*, California, USA, pp 239–248
24. Harsányi K, Kiss A, Majdik A et al (2018) A hybrid CNN approach for single image depth estimation: A case study. *International conference on multimedia and network information system*. Springer, Cham, pp 372–381
25. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: *Proceedings of the international conference on medical image computing and computer-assisted intervention*, Springer, Cham, pp 234–241
26. Chen Y, Zhao H, Hu Z et al (2021) Attention-based context aggregation network for monocular depth estimation. *Int J Mach Learn Cybern* 12(6):1583–1596
27. Tu X, Xu C, Liu S et al (2021) Efficient monocular depth estimation for edge devices in internet of things. *IEEE Trans Industr Inf* 17(4):2821–2832
28. Song M, Lim S, Kim W (2021) Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE transactions on circuits and systems for video technology*.
29. Ye X, Chen S, Xu R (2020) DpNet: Detail-preserving network for high quality monocular depth estimation. *Pattern Recognition* 109:107578
30. He K, Zhang X, Shaoqing R, Jian S (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, USA, pp. 770–778.
31. Zwald L, Lambert-Lacroix S (2012) The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*
32. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*
33. Silberman N, Hoiem D, Kohli P et al (2012) Indoor segmentation and support inference from rgb-d images. *European Conference on computer vision*. Springer, Berlin, Heidelberg, pp 746–760
34. Scharstein D, Pal C (2007) Learning conditional random fields for stereo. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1–8
35. Xu D, Ricci E, Ouyang W, et al (2017) Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 5354–5362
36. Carvalho M, Le Saux B, Trounev-Peloux P, et al (2018) On regression losses for deep depth estimation. In: *Proceedings of the IEEE international conference on image processing*, Athens, Greece, pp 2915–2919
37. Moukari M, Picard S, Simon L, et al (2018) Deep multi-scale architectures for monocular depth estimation. In *Proceedings of the IEEE international conference on image processing*, Athens, Greece, pp 2940–2944
38. Wang P, Shen X, Lin Z, et al (2015) Towards unified depth and semantic prediction from a single image. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Boston, USA, pp 2800–2809
39. Ma Z, Niu Y, Hu J (2020) Deep multi-scale convolutional neural network method for depth estimation from a single image. In *Chinese control and decision conference (CCDC)*. IEEE, pp3984–3988

40. Lo W Y, Chiu C T, Luo J Y (2020) Depth estimation from single image through Multi-Path-Multi-Rate diverse feature extractor. In IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1613–1617
41. Liu A, Lin W, Narwaria M (2011) Image quality assessment based on gradient similarity. IEEE T Image Process 21(4):1500–1512
42. Wang Z, Bovik AC, Sheikh HR et al (2004) Image quality assessment: from error visibility to structural similarity. IEEE T Image Process 13(4):600–612
43. Alhashim I, Wonka P (2018) High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941
44. Hu J, Ozay M, Zhang Y, et al (2019) Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries. In IEEE winter conference on applications of computer vision (WACV). IEEE, pp1043–1051

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.