J. Vis. Commun. Image R. 57 (2018) 234–242

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Deep intensity guidance based compression artifacts reduction for depth map ☆

Xu Wang [a,b,*], Pingping Zhang [a,b], Yun Zhang [d], Lin Ma [e], Sam Kwong [c], Jianmin Jiang [a,b]

[a] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[b] National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China
[c] Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
[d] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[e] Tencent AI Lab, Shenzhen, China

ARTICLE INFO

Article history:
Received 31 May 2018
Revised 30 September 2018
Accepted 5 November 2018
Available online 7 November 2018

Keywords:
Convolutional neural network
Compression artifacts reduction
JPEG compression
Depth map

ABSTRACT

In this paper, we propose an deep intensity guidance based compression artifacts reduction model (denoted as DIG-Net) for depth map. The proposed DIG-Net model can learn an end-to-end mapping from the color image and distorted depth map to the uncompressed depth map. To eliminate undesired artifacts such as discontinuities around object boundary, the proposed model is with three branches, which extracts the high frequency information from color image and depth maps as priors. Based on the modified edge preserving loss function, the deep multi-scale guidance information are learned and fused in the model to make the edge of depth map sharper. Experimental results show the effectiveness and superiority of our proposed model compared with the state-of-the-art methods.

1. Introduction

With the development of sensing technologies such as Microsoft Kinect and laser radio, it is easy to capture both the color and distance information of objects in 3D scene, which is also called RGB-D data. Based on the mapping function of imaging system, the depth map records the physical distance information from the surface of the objects to the camera in terms of 8-bits gray level. For each pixel in the depth map, the larger gray level indicates that the objects is closer to the camera. Since the depth map contains explicit geometry information of 3D scenes [1], the RGB-D data formats can be applied in many fields, such as saliency detection[2], 3D reconstruction [3], virtual reality [4] and autonomous driving [5].

Due to the limitation of bandwidth, the color image and depth map need to be compressed and transmitted to the client end for further processing, which cause compression artifacts. The inaccurate depth information will lead to the failure of the depth map based applications. Common image compression artifacts include blurring, blockiness and ring artifacts. These co-existed artifacts are inevitably imposed to image, especially for low bit-rate compression condition. Different from the color image, degradation on the depth map will lead to the error in image rendering and 3D scene reconstruction, even destroy the structure. Thus, compression artifacts reduction for depth map is important for the success of 3D application.

Recently, convolutional neural network (CNN) based models have shown excellent performance on the compression artifacts reduction task [6]. Each stage of CNN model is composed of a filter bank with followed rectified linear unit (ReLu). The whole model is trained to obtain the optimal filter coefficients and parameters together. Since the CNN based model can directly learning an mapping between the degraded image and its original version, the co-existed artifacts can be suppressed and eliminated together. However, existing CNN based model are mainly designed for color image, the performance of these models on depth map are limited since the characteristic of depth map is significantly different.

According to the best of our knowledge, there are little published CNN based works that aimed to reduce the compression artifacts of depth map. Existing works on depth recovery such as depth inpainting and super-resolution shows that color image can be used as prior to guide the depth enhancement. Previously, we proposed a model (denoted as IG-Net [7]) by using the color information to guide the filtering processing. To eliminate undesired artifacts such as discontinuities around object boundary and blurring, a model structure with three branches was proposed, where two ancillary branches were designed to extract the high-

☆ This article is part of the Special Issue on REV 5.

* Corresponding author.
E-mail addresses: wangxu@szu.edu.cn (X. Wang), yun.zhang@siat.ac.cn (Y. Zhang), cssamk@szu.edu.cn (S. Kwong), jianmin.jiang@szu.edu.cn (J. Jiang).

https://doi.org/10.1016/j.jvcir.2018.11.008
1047-3203/

frequency information from color image and depth map, respectively. The multi-scale guidance information were learned from color image and depth map to strength the edge information of the restored depth map. Although the IG-Net model can achieve significantly performance improvement on artifacts removal for depth map, there are still some open issues need to be well addressed such as over-smoothing and computational complexity. In this paper, we proposed a deep intensity guided CNN (denoted as DIG-Net) model by further investigating and analyzing the characteristics of depth map based on IG-Net, which avoids introducing false boundary and strengths the edge information of the restored depth map. The main contributions are listed as follows:

- To increase the capacity of whole model, the feature maps from different branches are fused through the concatenation instead of the element-wise addition operation. Consequently, the input feature numbers of feature enhancement and mapping layer in M-branch are increased by three times, which improve the performance.
- To protect the edge information in depth map, the loss function of model training is modified by introducing an edge similarity term, which significantly speed up the convergence of training stage and improve the performance.
- The improved DIG-Net contains little convolutional layers and the number of filters, thus the theoretical time complexity of the proposed DIG-Net is only 50% of IG-Net. During the inference stage, the actual running time consumption of DIG-Net is only 83% of IG-Net in average.
- The optimal setting of hyper-parameters such as weighting factor λ of loss function, filter numbers are determined in the experimental part. Meanwhile, the contribution of each component and the influence of compressed color image are also discussed.

The rest of this paper is organized as follows. Section 2 overviews the related works. The detailed descriptions of proposed model are provided in Section 3. Then the implementation details such as model training and optimization, performance comparison between the state-of-the-art models are summarized in Section 4. Finally, the conclusion is given in Section 5.

2. Related works

Due to the bandwidth limitation, the artifacts imposed on image will become obviously as the quantization error of image compression increased. For example, JPEG compression will cause block artifacts and blurness around the edge. Since we focus on the compression artifacts removal for depth map with the guidance of color image, related works about the state-of-the-arts on compression artifacts reduction and joint image filtering are discussed as follows.

2.1. JPEG compression artifacts reduction

The basic concept of JPEG compression artifacts reduction is to suppress or eliminate the quantization noises from the original signals. According the difference on image modeling methodology, existing works are summarized as follows:

Deblocking oriented method regards quantization noise as additive white Gaussian noise, and recover the signal similar to the image denoising via pre-defined prior model, such as smoothness, sparsity and Gaussian processes [8]. For example, spatial domain filtering is widely investigated for decades [9] to remove the blockness artifacts. To improve the compression performance, adaptive deblocking filter is embedded in the reconstruction stage of video codecs [10,11]. Luo et al. [12] proposed an adaptive approach to reduce the block-to-block discontinuities in both the spatial and discrete cosine transform (DCT) domains. Singh et al. [13] modeled the blockness as a 2D step functions between two neighboring blocks in DCT domain. Different filters are applied to smooth/non-smooth regions, respectively. Pointwise shape-adaptive DCT (SA-DCT) [14] can be computed on a support of arbitrary shape, then reconstructed edges are clean and have no ringing artifact. Recently, Yang et al. [15] proposed a cross-view multilateral filtering scheme, which significantly improves the quality of compressed depth maps. However, the quantization operations in DCT transform based encoder is non-linear, thus the performance is limited due to the inaccurate modeling of compression artifacts.

Dequantization based method is also called soft decoding based method. It treat the dequantization process as ill-posed optimization problem and reconstructs each block by selecting appropriate coefficient values under the constraint of both indexed quantization bins and signal priors. For example, Zakhor et al. [16] proposed the projection on convex sets (POCS) algorithm to minimize the reconstruction errors. To find suitable and effective priors, Liu et al. [8] proposed a soft decoding algorithm by exploiting the random walk graph Laplacian based smoothness prior. It consists of three types of prior, including Laplacian prior for DCT coefficients, sparsity prior and graph-signal smoothness prior for image patches. Since modeling of compression artifact considers the quantization operation, soft decoding based method significantly improves the reconstruction performance compared to the deblocking oriented method. However, the time complexity of inference stage is very huge due to the iterative optimization of solving inverse problems.

Deep CNN based method has shown great success in handling image restoration tasks [17,18], which can automatically learning both the filter banks and the corresponding combination of weights from the large scale training dataset. The convolutional layers can capture the abstraction of image contents while eliminating corruptions. Dong et al. [6] proposed a artifacts reduction convolutional neural network (AR-CNN) model with four convolutional layers, which is effective in reducing blocking artifacts while preserving edge and sharp details. To reduce the computation complexity and meet the requirement of real-time image processing, a compact and efficient network [19] was proposed for seamless attenuation of different compression artifacts by introducing layer decomposition and joint using large-stride convolutional and deconvolutional layers. In [20], the networks module with eight layers in a single step and in relatively short time was proposed by combining residual learning, skip architecture, and symmetrical weight initialization techniques. CAS-CNN [21] is a model with hierarchical skip connections and multi-scale loss function. However, above mentioned models are focused on the compression artifact removal for color image. Although these models can be directly applied to depth map, the performance are limited since they did not make full use of the guidance information from the corresponding color image. Thus, we need to modified the loss function by considering the edge similarity between the restored depth map and the ground-truth depth map.

2.2. Joint image filtering

Joint image filtering is widely used to transfer the salient structural details from the guidance image to the degraded image, to fill the missing pixels or suppress noise. The depth map represents the depth information of the scene, and their characteristics are different from the color image. Existing works on depth recovery showed that there are local/non-local similarly between the depth map and the corresponding color image [22]. Yang et al. [23] pro-

posed a color-guided autoregressive (AR) model which predicts the pixels of depth map based on both the local correlation of degraded depth map and non-local similarity form the corresponding color image. Inspired by these, IG-Net proposed depth map artifacts reduction model with three branches, which is guided by the high-frequency information from depth map and its corresponding color image, respectively. However, there is no regularized term in the loss function of IG-Net, the structure details such as object boundary are not protected well. Thus, the loss function need to be modified to preserve the edge details for the restored depth map.

3. Proposed model

In this section, the concept of proposed DIG-Net model is discussed in details. Suppose the compressed depth map, its corresponding intensity image (luminance component of color image) and ground truth of depth map are denoted as X, Y and D, respectively, then the goal of compression artifacts reduction model is to learn an end-to-end mapping from X and Y to D. As shown in Fig. 1, the architecture of the DIG-Net network contains three channels, denoted as Y-branch, M-branch and D-branch, respectively. The Y- and D-branches are designed to extract high frequency information of the intensity image and depth map, respectively. The M-branch concatenates and fuses the feature maps extracted from the Y-branch and D-branch to achieve compression artifacts reduction.

3.1. Spectral decomposition for intensity image and depth map

Actually, the physic meaning of intensity image and depth map are significantly different. The depth map record the distance information, whereas the pixel value of corresponding color image indicates the intensity of luminance and color information. The color image contains mixed information such as texture, intensity and edges. The salient structures that are consistent with both guidance and target images can reduce the uncertainty and speedup the convergence of training stage. Conversely, the erroneous structural patterns may misguide and slow down the training stage, even cause non-convergence.

Based on the observations, we find that the edge of the depth image corresponds to the edge of the color image, since the pixels belong to the same object usually have the same depth. Different from the corresponding intensity image, depth maps are smooth in the object region and sharp around the object boundary. To reduce the interrupt of low frequency information, the concept of spectral decomposition inspired by [22,24] is employed to extract the high frequency information from color image and depth map in Y- and D-branches, respectively. Meanwhile, using the high-frequency information of intensity image and depth map can speed up the training stage and guarantee the convergence of network training. This operation for Y- and D-branches can be expressed as:

$$F_Y^0 = Y - W_Y^0 * Y \tag{1}$$

$$F_D^0 = X - W_D^0 * X \tag{2}$$

where $*$ denotes the convolution operation. W_0^s is the filter with size $f_0^s \times f_0^s$ for Y- and D-branches, where the superscript $s \in \{Y, D\}$. Since we aim to obtain low frequency component, then W_0^s is fixed as mean filter for its simplicity. Fig. 2 illustrates the high frequency components extracted from intensity image and depth map, respectively. It is observed that the high frequency components of intensity image contains rich structure information of depth map. Although it still contains erroneous structure patterns, the influence will be suppressed by the following feature extraction and enhancement operations of Y branch.

3.2. Deep intensity guidance

To suppress and eliminate the noise introduced by compression, our proposed model contains five convolutional layers in M-branch as shown in Fig. 1. Each of these layer is designed for a specific function, including feature extraction, feature enhancement, mapping and reconstruction. The Y-branch and D-branch contains three convolutional layers, including the high frequency extraction, feature extraction and feature enhancement layers. To exploit the guidance information from Y- and D-branches, it is necessary to combine and merge the feature maps of all the branches. In our previous work [7], the feature maps from each branch are fused through the element-wise summation operation. However, this type of operation directly merge the local structure details of

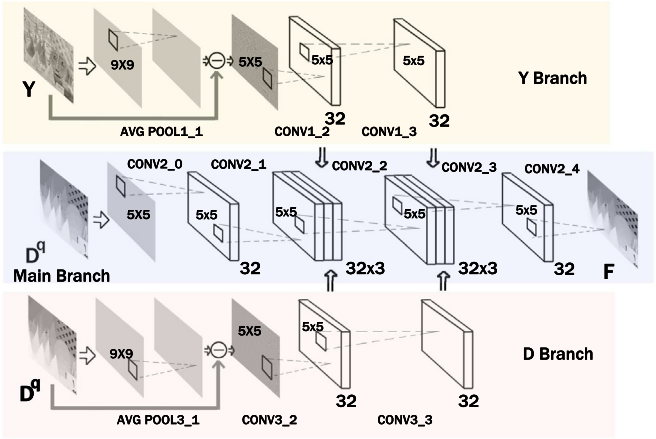
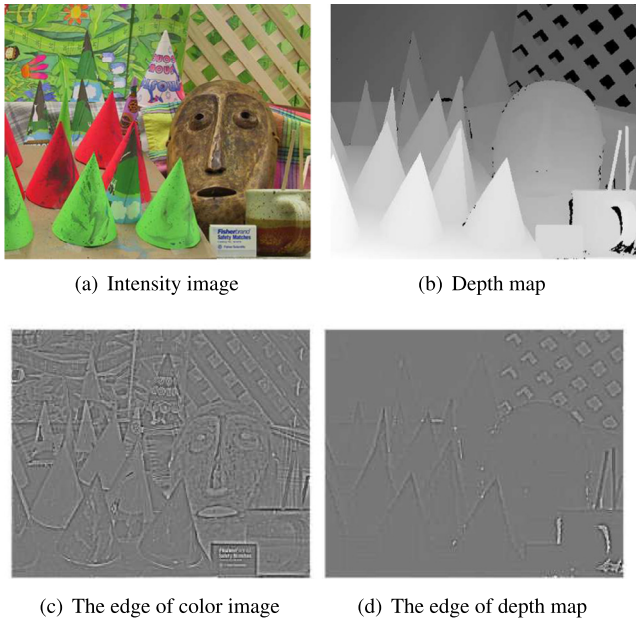


Fig. 1. The architecture of proposed DIG-Net. The overall network contains three branches, named as Y-, D- and M-branch. The feature maps of Y-, D- and M-branches are combined as the input of next layer through the concatenation operation.



(a) Intensity image (b) Depth map

(c) The edge of color image (d) The edge of depth map

Fig. 2. The intensity image, depth map, high frequency components of intensity image and depth map for test image **"Cones"**, respectively. For better visualization, the value range of coefficients are rescaled into 0–255.

Y- and D-branches to the M-branch, which also introduce erroneous local details to the feature maps of M-branch. Thus, it is hard to train the model since the network needs to reduce the influence of erroneous information introduced from guidance information. In this paper, the output feature maps of feature extraction (or feature enhancement) layers from all three branches are concatenated as the input of feature enhancement (or mapping) layer in M-branch. Consequently, the capacity of whole model is significantly increased since the number of input feature maps of feature enhancement (or mapping) layer in M-branch is increased by three times. Final, the whole model can automatically learn the useful information from guidance branches in different stages. The detailed mathematical formula of model can be expressed as follows:

$$F_M^0 = max\left(0, W_M^0 * X + B_M^0\right) \tag{3}$$

$$F_M^1 = max\left(0, W_M^1 * X + B_M^1\right) \tag{4}$$

$$F_M^2 = max\left(0, W_M^2 * \left[F_M^1, F_Y^1, F_D^1\right] + B_M^2\right) \tag{5}$$

$$F_M^3 = max\left(0, W_M^3 * \left[F_M^2, F_Y^2, F_D^2\right] + B_M^2\right) \tag{6}$$

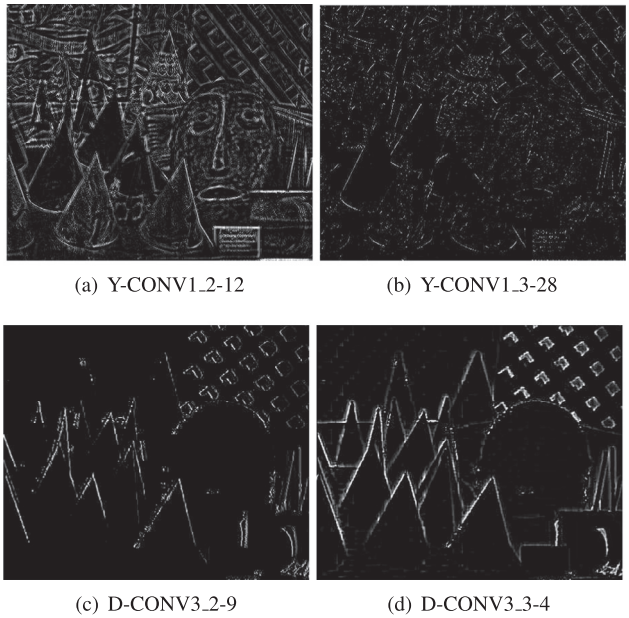
$$F = max\left(0, W_M^4 * F_M^3 + B_M^4\right) \tag{7}$$

where

$$F_Y^j = max\left(0, W_Y^j * F_Y^j + B_Y^j\right),\ j \in 1, 2, \tag{8}$$

$$F_D^j = max\left(0, W_D^j * F_D^j + B_D^j\right),\ j \in 1, 2 \tag{9}$$

W_i^s is the filter with size $n_{i-1}^s \times f_i^s \times f_i^s \times n_i^s$ for Y, M and D branches, where the superscript $s \in \{Y, M, D\}$. n_{i-1}^s and n_i^s are the number of feature maps of input and output, respectively. B_i^s is a n_i^s-dimensional bias vector. For the top layer, it is scalar. $\left[F_M^i, F_Y^i, F_D^i\right]$ refers to the concatenation of the feature maps produced by layers of each branch. During the implementation, the multiple inputs of F_M^2 and F_M^3 are concatenated into a single tensor. Denote the overall network architecture as F and the model parameters as $\theta = \{W, B\}$,



(a) Y-CONV1_2-12 (b) Y-CONV1_3-28

(c) D-CONV3_2-9 (d) D-CONV3_3-4

Fig. 3. Selected example feature maps from the Y-Branch and D-Branch for test image **"Cones"**, respectively. For better visualization, the value range of coefficients are rescaled into 0–255.

then the final output $\widehat{D} = F(Y, X; \theta)$ is the restored depth map of the same size as the input compressed depth map X. For easy understanding, Fig. 3 provides some visualization results of random selected feature maps from Y- and D-branches, respectively. It is observed that both of two guidance branches provide useful local structure details to the M-branch.

3.3. Loss function

Different from traditional image restoration problem for compression artifacts reduction, the depth map is not used for viewing. Thus, the perceptual oriented optimization criterion may not suitable for the model training of depth maps. As mentioned above, the quantization error of JPEG compressed depth map blurred the object boundary and introduce false edge or texture on smooth regions. In our initial model IG-Net [7], the root-mean-square deviation (RMSD) is employed as loss function to measure the signal error. However, this type of L2 loss function cannot well protect the structural details of depth map since the guidance information from Y- and D-branches are not fully exploited.

In this paper, the loss function of model training takes both the signal error and edge similarity into consideration. Given N training samples, the overall network is trained to determine the optimal model parameters by minimizing the loss function as follows:

$$L(\theta) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}||\widehat{D}_n - D_n||^2} + \lambda\sqrt{\frac{1}{N}\sum_{n=1}^{N}||S\left(\widehat{D}_n\right) - S(D_n)||^2} \tag{10}$$

$S(\cdot)$ is the measurement of edge strength in terms of gradient magnitude on the restored depth map $\widehat{D}$ and ground truth D, respectively. Given the input image patch I, the mathematical form of $S(I)$ is defined as

$$S(I) = |I * G_h| + |I * G_v|, \tag{11}$$

G_h and G_v are the horizontal and vertical gradient extraction operators respectively. In this paper, the classical Sobel operations are employed to obtain the horizontal and vertical gradient from the input image path I, which is provided as follows:

$$G_h = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{12}$$

and

$$G_v = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{13}$$

The first term in Eq. (10) is used to measure the signal error between the reconstructed output and ground truth. The second term is used to measure the edge similarity, which try to limit the margin of error on edge strength as small as possible. This is reasonable since the compression artifacts are not equally distributed in the spatial domain. The area with strong texture or edge may suffer more degradation due to the quantization strategy, which allocate more bits to the area with smooth details. Since there are strong correlation between depth map and color image in terms of structure details, then the network will be forced to learn the optimal filters and parameters of Y- and D-branches during the train stage. The parameter λ in Eg. (10) is used to balance the contribution of each term. In the experimental results section, we will further discuss the influence of adjusting λ on the restoration performance and the speed of convergence in training stage.

4. Experimental results

4.1. Dataset preparation

To demonstrate the performance of proposed model, experiments are conducted on both synthetic and real RGB-D datasets. For synthetic datasets, 105 RGB-D datasets from Middlebury Stereo dataset, [25–28], MPI Sintel depth dataset [29] and Multi Modal Stereo dataset [30] are used for training. The remaining 12 RGB-D datasets, including "Alley", "Cave", "Market", "Room", "Tsukuba", "Venus", "Ambush", "Adirondack", "Art", "Motorcycle", "Wood" and "Cones", are used for testing. Detailed information are provided in Table 1.

For real dataset, the common used NYU-Depth V2 dataset [31] is used for performance evaluation, which contains 1449 densely labeled pairs of aligned RGB and depth images captured by Microsoft Kinect device. The first 1000 pairs are used for training, the remaining 449 pairs are used for testing.

Similar to the performance comparison protocols adopted in ARCNN [6], the depth map is compressed by the MATLAB JPEG encoder with different quality setting where $q \in \{10, 20, 30, 40\}$ (from low quality to high quality). The corresponding color image in training set is keep uncompressed. Detailed information about the compression setting for color and depth images are provided in Table 2.

4.2. Implementation details

Instead of directly using whole large-size images for training, sub-images are generated by dividing each depth map into a regular grid (32×32) of small overlapping patches ($stride = 20 \times 20$). The proposed model is trained and tested in tensorflow platform on a GPU server with two 14-core Intel Xeon E5-2690v4 CPUs, 256-GB DRAM, and eight Nvidia Titan X cards. The optimization is conducted by the Adam method with a batch size of 64. The learning rate is initially set to 10^{-4}. All the filters in convolution layers are randomly initialized from a zero-mean Gaussian distribution with standard deviation 10^{-3}. Since depth image is not used for viewing, the criterion of performance evaluation in terms of peak signal to noise ratio (PSNR) is employed in the experiment. For simplify, we denote the trained model as DIG-Net-q for the training set with quality setting q of depth image. The test set is denoted as TestSet-q where the test depth images are compressed with quality setting q. In this paper, the performance of model DIG-Net-q will be evaluated on TestSet-q.

Table 1
Number of images in training/testing dataset.

Dataset	Training set	Testing set
Middlebury DataSet	47	7
MPI sintel depth dataset	36	4
Multi-modal stereo dataset	22	1
Total	105	12

Table 2
Compression setting for color/depth images in training/testing set.

Model	Training set		Test set	
	Color	Depth	Color	Depth
DIG-Net-10	Original	10	Original	10
DIG-Net-20	Original	20	Original	20
DIG-Net-30	Original	30	Original	30
DIG-Net-40	Original	40	Original	40

4.3. Model and performance trade-offs

To make our proposed model flexible, model setting such as weight factors of loss function, number of filters are not fixed in the experiment. In the following parts, we will investigate the influence of model settings on the restoration performance, where the experiments are conducted on synthetic datasets in terms of Peak Signal-to-Noise Ratio (PSNR). Besides, the contributions of each module on the performance are also discussed.

4.3.1. Loss function parameters optimization

The contribution of constraint on edge strength is controlled by the parameter λ. The restoration performance may be influenced by the value of λ. In this part, we varies the value of λ to evaluate the importance of the edge similarity term to the final performance of overall network. The numbers of filter in each convolutional layers are set to 32. The evaluation results of model DIG-Net-40 on TestSet-40 are summarized in Table 3. It is observed that when the regularized term is enabled ($\lambda > 0$), the performance is improved. Apparently, if λ is very small (e.g., $\lambda = 0.01$), the contribution of regularized item is not significant. When $\lambda = 0.1$, the proposed model can achieve the best performance.

4.3.2. Filter numbers

In our model, the number of filters are set as equal for all the convolutional layers. To evaluate the influence of filter numbers on the performance, we fix some hyper-parameters, such as $\lambda = 0.1$, and train the model with different filter numbers. The evaluation results of model DIG-Net-10 on TestSet-10 and model DIG-Net-40 on TestSet-40 in terms of PSNR and inference time are summarized in Table 4. Since the inference time is independent of image quality, only the inference time of model DIG-Net-10 on TestSet-10 is reported. Due to the limitation of GPU memory capacity, the maximum number of filters is set to 48. It is observed that increasing filter numbers will improve the restoration performance but with increased computation complexity. However, the gain on reconstruction quality is limited when the number of filters larger than 32, whereas the overhead of computational complexity is significantly increased. To make a balance between reconstruction quality and inference time, we set the number of filters to 32 as default.

4.3.3. Validation of deep intensity guidance module

One main contribution of proposed model is using the color image as guidance information for the restoration process. To evaluate the performance gain of deep intensity guidance (Y-branch) module on the reconstruction quality, we train the DIG-Net model without Y branch (denoted as Proposed-nY). Table 5 summarized the performance comparison results. It is observed that the deep intensity guidance module can significantly improve the performance.

To guarantee the quality of guidance information, we train the model by using the uncompressed color image. However, it is unpractical to ensure the high quality of color image due to the compression requirement. To investigate the influence of color image quality on the inference performance, we evaluate the pro-

Table 3
The results of different λ setting in loss function in terms of average PSNR (dB) for the whole test dataset.

λ	0	0.01	0.1	1
PSNR (dB)	45.10	45.11	**45.30**	45.16

posed model DIG-Net-q on the test images where both depth map and color image are compressed with quality factor q. The results (denoted as Proposed-cY) is provided in Table 5. It is observed when the color images are with low quality ($q = 10$), the performance gain is influenced and decreased. This is reasonable since the JPEG compressed color image with low quality will cause blockness artifacts, which introduce erroneous structure to misguide the restoration process. When the quality of color images is acceptable, there are no significantly difference on the average performance.

4.4. Comparison with state-of-the-arts

To demonstrate the performance of our proposed model, we compare DIG-Net with the state-of-the-arts algorithms, including Shen [32], NLR [33], Ham [34], SA-DCT [14], Liu [8], AR-CNN [6], RED-Net [35] and IG-Net [7]. The first three models are targeted for image restoration task. SA-DCT [14] and Liu [8] are designed for compression artifacts reduction. AR-CNN, RED-Net and IG-Net are CNN based models, which have better performance than the first five models. All the training and test stage of AR-CNN, RED-Net, IG-Net and proposed DIG-Net models are same. The parameters of these models are set as default. It should be noted that the filter numbers in IG-Net are 64 except the final reconstruction layer. Compared to IG-Net, the extended DIG-Net contains little convolutional layer and learnable parameters.

Quantitative results evaluated on test set with different quality levels are shown in Tables 5 and 6 for synthetic and real depth datasets, respectively. For synthetic datasets, it is observed that our proposed model can achieve the best reconstruction quality in terms of PSNR, and outperforms other models on all scale of JPEG compression qualities. In other words, the proposed DIG-Net model can significantly reduce the artifacts of JPEG compressed depth map. Existing CNN models are not efficiency on removing compression artifacts of depth map since these models did not exploit the correlation between color and depth images. For real dataset, the proposed DIG-Net model achieves the best reconstruction quality for lower quality factors. For the conditions $q = 30$ and $q = 40$, the performances of IG-Net and proposed DIG-Net are similar.

For better visualization, we provide some examples from synthetic and real datasets as shown in Figs. 4 and 5, respectively. It is observed that our proposed model could produce sharper edges with much less blocking artifacts compared with other models. Since the CNN based models such as AR-CNN and RED-Net have no edge guidance and edge constraints, they cannot protect the edges very well. The SA-DCT model over-smoothed the edges and suppressed the edge strength. Ham model jointly leverage structural information of guidance and input images, which can protect some structure information, but it does not have a good result in PSNR evaluation. In contrast, the results generated by our proposed model are smoother, sharper and more accurate with respect to the ground truth for both synthetic and real depth datasets.

Currently, CNN based architectures need to be retrained for different quality factor. It is not very friendly for practical application since these pre-trained models will occupy many storage resources. It is necessary to improve the adaptivity of the CNN models for real scenes with different quality level. In the future, we will try to model and predict the non-linear characteristics of quantization operation, which maybe helpful for solving this problem.

Table 5
Performance comparison of compression artifacts reduction algorithms in terms of PSNR (dB) on synthetic datasets.

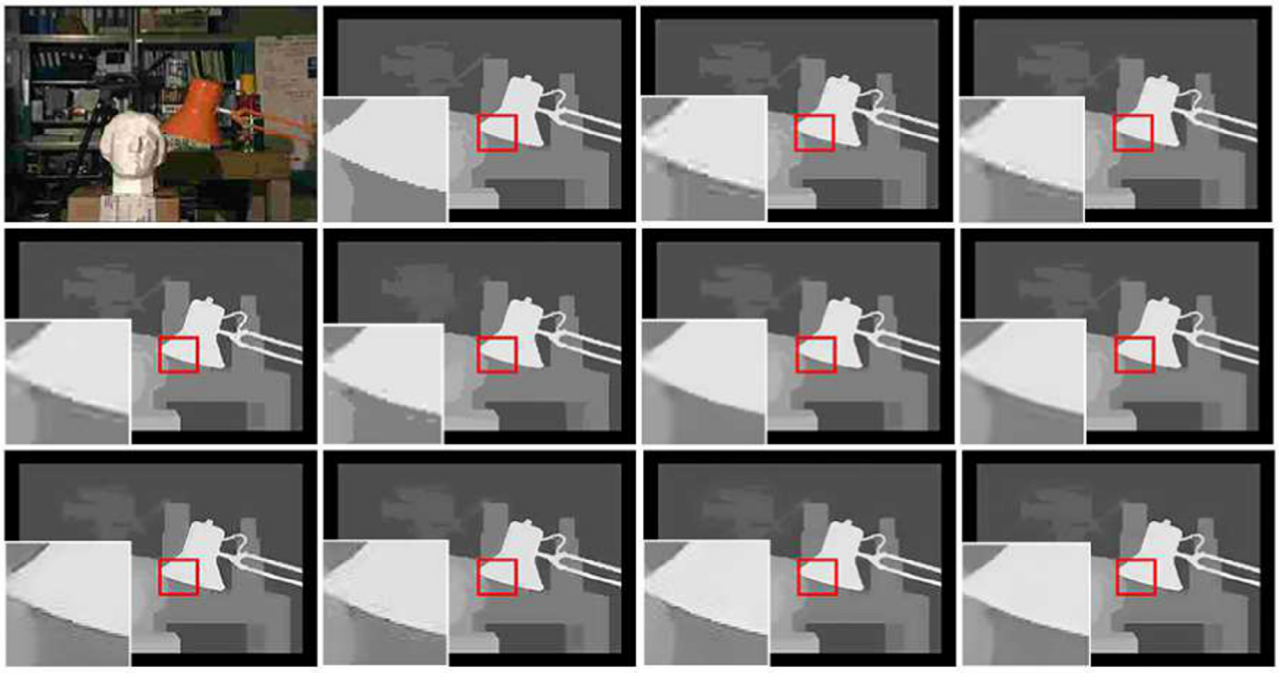
Model	$q = 10$	$q = 20$	$q = 30$	$q = 40$
JPEG	32.30	34.71	36.17	37.20
Shen [32]	32.78	35.26	36.76	37.80
NLR [33]	32.69	35.30	36.89	38.03
Ham [34]	33.91	36.30	37.84	38.94
SA-DCT [14]	33.40	35.81	37.37	38.59
Liu [8]	33.79	36.44	38.23	40.01
ARCNN [6]	34.51	37.63	39.35	40.46
RED-Net [35]	34.76	39.68	42.45	44.07
IG-Net [7]	35.71	39.98	42.88	44.50
Proposed-noY	35.63	39.98	42.98	45.04
Proposed-cY	35.75	40.19	43.26	45.30
Proposed	**35.88**	**40.19**	**43.27**	**45.30**

Table 6
Performance comparison of compression artifacts reduction algorithms in terms of PSNR (dB) on real datasets.

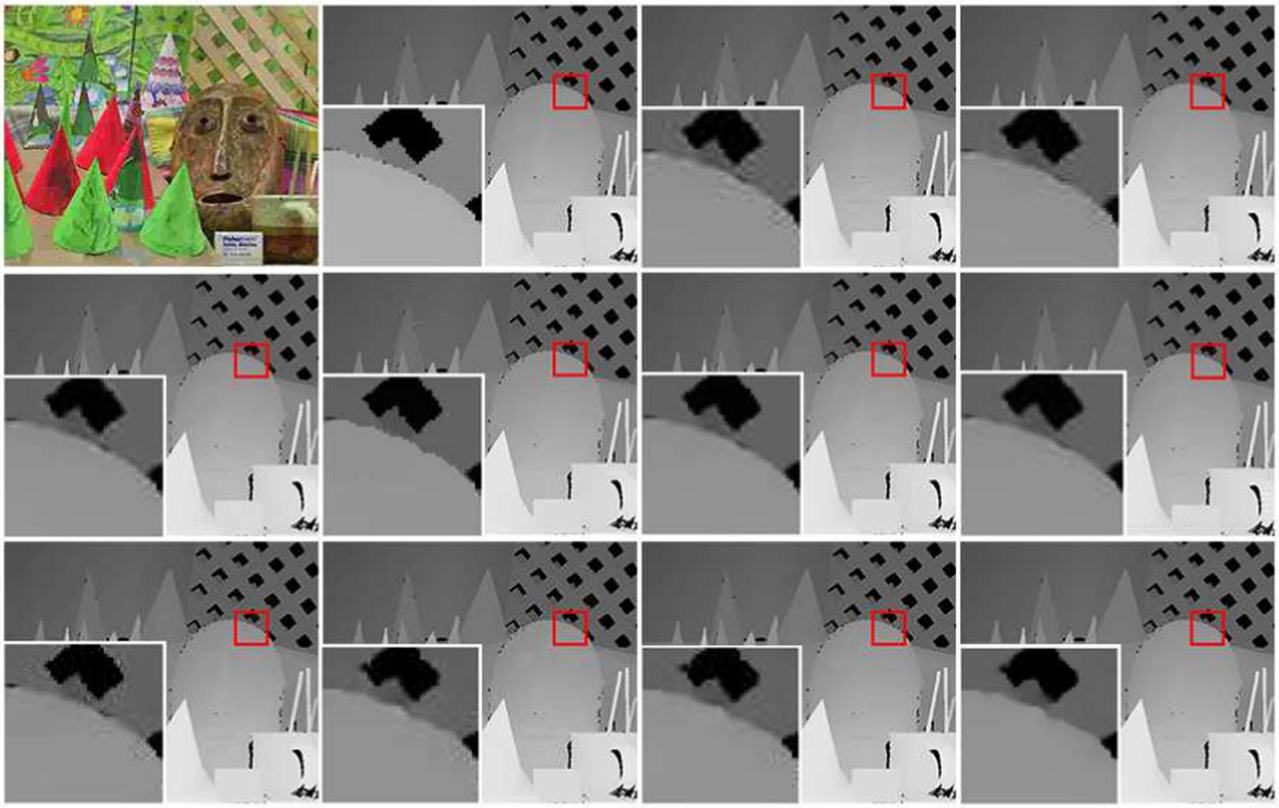
Model	$q = 10$	$q = 20$	$q = 30$	$q = 40$
JPEG	36.52	40.56	42.54	43.84
Shen [32]	37.46	41.09	42.41	43.13
NLR [33]	37.26	41.89	43.82	44.93
Ham [34]	38.64	42.28	43.31	43.89
SA-DCT [14]	38.68	42.74	44.42	45.39
Liu [8]	36.40	36.95	37.15	37.20
ARCNN [6]	39.69	43.34	45.00	46.06
RED-Net [35]	39.09	43.75	45.30	46.28
IG-Net [7]	40.58	43.93	45.96	**47.02**
Proposed	**40.71**	**44.24**	**45.96**	47.00

Table 4
Summarized evaluation results of different setting on filter numbers in terms of PSNR (dB) and inference time (ms) for the whole test dataset.

Test image	Resolution	Filter No. = 8			Filter No. = 16			Filter No. = 32			Filter No. = 48		
		q = 40	q = 10	Time	q = 40	q = 10	Time	q = 40	q = 10	Time	q = 40	q = 10	Time
Alley	1024x436	40.53	36.09	30	44.22	36.97	41	45.50	38.04	62	45.24	38.04	107
Cave	1024x436	43.13	35.94	28	43.53	36.29	38	45.22	37.30	63	45.28	36.73	107
Market	1024x436	39.83	33.94	29	41.08	34.46	38	41.82	35.22	65	41.61	35.44	109
Room	640x480	40.57	32.01	24	43.59	32.27	30	45.49	32.80	48	45.56	32.56	80
Tsukuba	384x288	39.46	33.63	15	41.91	34.56	21	43.56	35.06	24	43.25	35.32	36
Venus	434x383	46.85	39.33	16	48.51	40.12	23	48.26	40.92	30	49.74	40.71	48
Ambush	1024x436	43.08	37.11	29	45.33	37.54	41	46.12	38.58	65	45.86	38.39	109
Adirondack	2880x1988	42.85	34.12	221	45.02	34.46	365	46.68	34.82	657	46.82	34.79	1257
Art	1390x1110	40.33	32.80	71	43.27	33.11	109	45.32	33.67	184	45.61	33.75	349
Motorcycle	2964x2000	39.63	30.98	234	41.57	31.17	378	43.05	31.56	689	43.21	31.71	2087
Wood	1306x1110	48.39	38.97	67	49.36	39.44	101	50.67	39.85	173	50.77	39.76	329
Cones	450x375	38.76	31.84	18	41.09	32.29	25	41.87	32.73	33	42.00	32.68	49
Average		41.95	34.73	65	44.04	35.23	101	45.30	35.88	174	45.41	35.82	389



(a) Color Image (“Tsukuba”), Ground truth , JPEG Compression with q=10: 31.03dB, Shen: 31.59dB, NLR: 31.41dB, Ham: 33.15dB, SA-DCT: 33.09dB, Liu: 32.52dB, AR-CNN: 33.71dB, RED-Net: 34.41dB, IG-Net: 35.02dB, DIG-Net: 35.06dB.



(b) Color Image (“Cones”), Ground truth, JPEG Compression with q=40: 33.79dB, Shen: 34.22dB, NLR: 34.47dB, Ham: 35.97dB, SA-DCT: 35.76dB, Liu: 35.45dB, AR-CNN: 37.65dB, RED-Net: 41.06dB, IG-Net: 40.98dB, DIG-Net: 41.87dB.

Fig. 4. Qualitative comparisons of compression artifact reduction models. From left to right and up to down, the compressed depth maps are restored by Shen, NLR, Ham, SA-DCT, Liu, AR-CNN, RED-Net, IG-Net and DIG-Net.

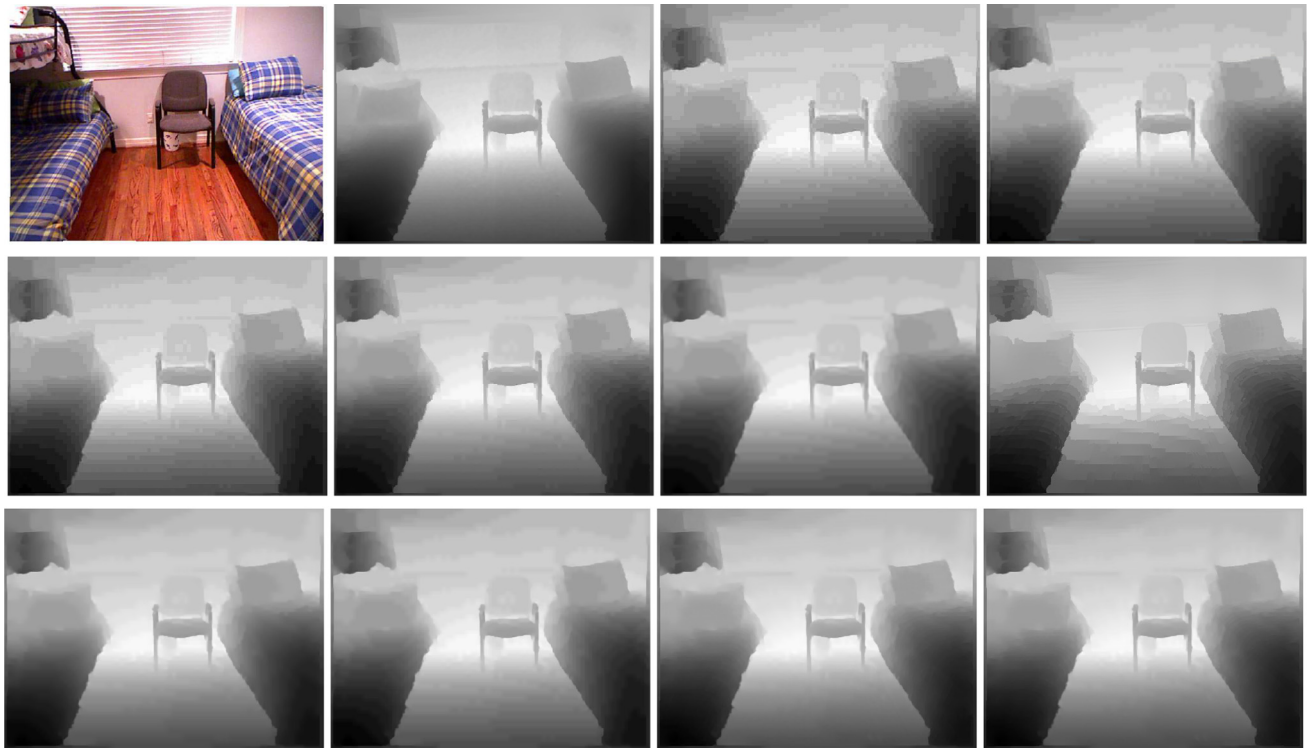


Fig. 5. Qualitative comparisons of compression artifact reduction models. From left to right and up to down, Color image and Ground truth ("1056-th sample of NYU-Depth V2 dataset"), JPEG Compression with q = 10: 36.34 dB, Shen: 37.68 dB, NLR: 37.19 dB, Ham: 39.00 dB, SA-DCT: 38.89 dB, Liu: 35.92 dB, AR-CNN: 40.07 dB, RED-Net: 39.36 dB, IG-Net: 40.83 dB, DIG-Net: 40.97 dB.

5. Conclusion

This paper proposed an compressed artifacts reduction model for depth map, which is guided by the high-frequency information from depth map and its corresponding color image, respectively. To strength the edge information of restored depth map, our proposed DIG-Net model consists of three branches, including Y branch, D branch and main branch. Besides, the loss function is modified to make the restored depth map sharper. Compared with the state-of-the-arts models, our proposed model can achieve the best performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 31670553, 61871270, 61501299, 61672443 and 61620106008, in part by the Guangdong Nature Science Foundation under Grant 2016A030310058, in part by the Shenzhen Emerging Industries of the Strategic Basic Research Project under Grants JCYJ20160226191842793, in part by the Natural Science Foundation of SZU (Grant No. 827000144), and in part by the Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University.

References

[1] K. Muller, P. Merkle, T. Wiegand, 3-D video representation using depth maps, Proc. IEEE 99 (4) (2011) 643–656, https://doi.org/10.1109/JPROC.2010.2091090.
[2] Y. Yang, B. Li, P. Li, Q. Liu, A two-stage clustering based 3D visual saliency model for dynamic scenarios, IEEE Trans. Multimedia (2018) 1, https://doi.org/10.1109/TMM.2018.2867742.
[3] Q.-Y. Zhou, V. Koltun, Color map optimization for 3D reconstruction with consumer depth cameras, ACM Trans. Graph. 33 (4) (2014) 155:1–155:10, https://doi.org/10.1109/TIP.2018.2867740.
[4] J. Thatte, J. Boin, H. Lakshman, G. Wetzstein, B. Girod, Depth augmented stereo panorama for cinematic virtual reality with focus cues, in: 2016 IEEE International Conference on Image Processing (ICIP), 2016, pp. 1569–1573, https://doi.org/10.1109/ICIP.2016.7532622.
[5] M. Meilland, A.I. Comport, P. Rives, Dense omnidirectional RGB-D mapping of large-scale outdoor environments for real-time localization and autonomous navigation, J. Field Robot. 32 (4) (2015) 474–503, https://doi.org/10.1002/rob.21531.
[6] C. Dong, Y. Deng, C. Change Loy, X. Tang, Compression artifacts reduction by a deep convolutional network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 576–584.
[7] P. Zhang, X. Wang, Y. Zhang, L. Ma, J. Jiang, S. Kwong, Compression artifacts reduction for depth map by deep intensity guidance, in: B. Zeng, Q. Huang, A. El Saddik, H. Li, S. Jiang, X. Fan (Eds.), Advances in Multimedia Information Processing – PCM 2017, Springer International Publishing, Cham, 2018, pp. 863–872.
[8] X. Liu, G. Cheung, X. Wu, D. Zhao, Random walk graph Laplacian-based smoothness prior for soft decoding of JPEG images, IEEE Trans. Image Process. 26 (2) (2017) 509–524, https://doi.org/10.1109/TIP.2016.2627807.
[9] J.S.L. Howard, C. Reeve, Reduction of blocking effects in image coding, Opt. Eng. 23 (1984), https://doi.org/10.1117/12.7973248, 23–23–4.
[10] P. List, A. Joch, J. Lainema, G. Bjontegaard, M. Karczewicz, Adaptive deblocking filter, IEEE Trans. Circ. Syst. Video Technol. 13 (7) (2003) 614–619, https://doi.org/10.1109/TCSVT.2003.815175.
[11] C. Wang, J. Zhou, S. Liu, Adaptive non-local means filter for image deblocking, Signal Process.: Image Commun. 28 (5) (2013) 522–530, https://doi.org/10.1016/j.image.2013.01.006.
[12] Y. Luo, R.K. Ward, Removing the blocking artifacts of block-based DCT compressed images, IEEE Trans. Image Process. 12 (7) (2003) 838–842, https://doi.org/10.1109/TIP.2003.814252.
[13] S. Singh, V. Kumar, H. Verma, Reduction of blocking artifacts in JPEG compressed images, Digital Signal Process. 17 (1) (2007) 225–243, https://doi.org/10.1016/j.dsp.2005.08.003.
[14] A. Foi, V. Katkovnik, K. Egiazarian, Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images, IEEE Trans. Image Process. 16 (5) (2007) 1395–1411.
[15] Y. Yang, Q. Liu, X. He, Z. Liu, Cross-view multi-lateral filter for compressed multi-view depth video, IEEE Trans. Image Process. (2018) 1, https://doi.org/10.1109/TIP.2018.2867740.
[16] A. Zakhor, Iterative procedures for reduction of blocking effects in transform image coding, IEEE Trans. Circ. Syst. Video Technol. 2 (1) (1992) 91–95, https://doi.org/10.1109/76.134377.
[17] J. Guo, H. Chao, One-to-many network for visually pleasing compression artifacts reduction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3038–3047.
[18] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, T.S. Huang, D3: deep dual-domain based fast restoration of jpeg-compressed images, in: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, pp. 2764–2772.
[19] C. Dong, C.C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: Proceedings of European Conference on Computer Vision, 2016, pp. 391–407.
[20] P. Svoboda, M. Hradis, D. Barina, P. Zemcik, Compression artifacts removal using convolutional neural networks, J. WSCG 24 (2) (2016) 63–72.
[21] L. Cavigelli, P. Hager, L. Benini, CAS-CNN: a deep convolutional neural network for image compression artifact suppression, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 752–759.
[22] T.-W. Hui, C.C. Loy, X. Tang, Depth map super-resolution by deep multi-scale guidance, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), European Conference on Computer Vision, Springer International Publishing, Cham, 2016, pp. 353–369.
[23] J. Yang, X. Ye, K. Li, C. Hou, Y. Wang, Color-guided depth recovery from RGB-D data using an adaptive autoregressive model, IEEE Trans. Image Process. 23 (8) (2014) 3443–3458, https://doi.org/10.1109/TIP.2014.2329776.
[24] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, M. Yang, Gated fusion network for single image dehazing, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[25] D. Scharstein, R. Szeliski, High-accuracy stereo depth maps using structured light, in: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings, 2003.
[26] D. Scharstein, C. Pal, Learning conditional random fields for stereo, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8, https://doi.org/10.1109/CVPR.2007.383191.
[27] H. Hirschmuller, D. Scharstein, Evaluation of cost functions for stereo matching, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8, https://doi.org/10.1109/CVPR.2007.383248.
[28] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, P. Westling, High-resolution stereo datasets with subpixel-accurate ground truth, in: X. Jiang, J. Hornegger, R. Koch (Eds.), German Conference on Pattern Recognition, Springer International Publishing, Cham, 2014, pp. 31–42.
[29] D.J. Butler, J. Wulff, G.B. Stanley, M.J. Black, A naturalistic open source movie for optical flow evaluation, in: A. Fitzgibbon et al. (Eds.), European Conf. on Computer Vision (ECCV), Part IV, LNCS, vol. 7577, Springer-Verlag, 2012, pp. 611–625.
[30] M. Yaman, S. Kalkan, An iterative adaptive multi-modal stereo-vision method using mutual information, J. Vis. Commun. Image Represent. 26 (2015) 115–131, https://doi.org/10.1016/j.jvcir.2014.11.010.
[31] N. Silberman, P. Kohli, D. Hoiem, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, 2012, pp. 746–760.
[32] X. Shen, C. Zhou, L. Xu, J. Jia, Mutual-structure for joint filtering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3406–3414.
[33] W. Dong, G. Shi, X. Li, Y. Ma, F. Huang, Compressive sensing via nonlocal low-rank regularization, IEEE Trans. Image Process. 23 (8) (2014) 3618–3632, https://doi.org/10.1109/TIP.2014.2329449.
[34] B. Ham, M. Cho, J. Ponce, Robust image filtering using joint static and dynamic guidance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4823–4831.
[35] X. Mao, C. Shen, Y. Yang, Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates Inc., USA, 2016, pp. 2810–2818, https://doi.org/10.1109/TIP.2014.2329449.