Noname manuscript No.
(will be inserted by the editor)

Joint Image Denoising and Disparity Estimation via Stereo Structure PCA and Noise-tolerant Cost

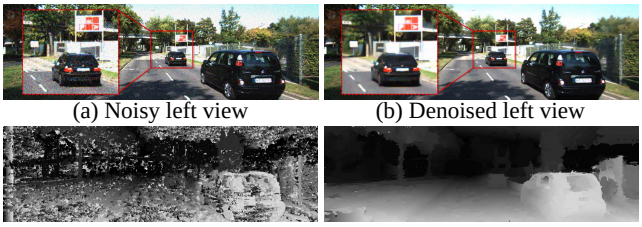
Jianbo Jiao[1] · **Qingxiong Yang**[2] · **Shengfeng He**[3] · **Shuhang Gu**[4] · **Lei Zhang**[4] · **Rynson W.H. Lau**[1]

Received: date / Accepted: date

Abstract Stereo cameras are now commonly available on cars and mobile phones. However, the captured images may suffer from low image quality under noisy conditions, producing inaccurate disparity. In this paper, we aim at jointly restoring a clean image pair and estimating the corresponding disparity. To this end, we propose a new joint framework that iteratively optimizes these two different tasks in a multiscale fashion. First, structure information between the stereo pair is utilized to denoise the images using a non-local means strategy. Second, a new noise-tolerant cost function is proposed for noisy stereo matching. These two terms are integrated into a multiscale framework in which cross-scale information is leveraged to further improve both denoising and stereo matching. Extensive experiments on datasets captured from indoor, outdoor, and low-light conditions show that the proposed method achieves superior performance than the state-of-the-art image denoising and disparity estimation methods. While it outperforms multi-image denoising methods by about 2dB on average, it achieves a 50% error reduction over radiometric-change-robust stereo matching on the challenging KITTI dataset.

Jianbo Jiao (Jambol.Jiao@my.cityu.edu.hk)
Qingxiong Yang (liiton.research@gmail.com)
Shengfeng He (shengfeng_he@yahoo.com)
Shuhang Gu (cssgu@comp.polyu.edu.hk)
Lei Zhang (cslzhang@comp.polyu.edu.hk)
Rynson W.H. Lau (Rynson.Lau@cityu.edu.hk)

1 Department of Computer Science, City University of Hong Kong, Hong Kong, China
2 School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui, China
3 School of Computer Science and Engineering, South China University of Technology, Guangzhou, China
4 Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China



(a) Noisy left view (b) Denoised left view
(c) Traditional disparity (d) Proposed disparity

Fig. 1 Given (a) a pair of noisy stereo images, (c) traditional disparity methods typically fail in noisy conditions. The proposed method jointly estimates (b) the denoised images and (d) the disparity map.

Keywords Stereo matching · image denoising · disparity estimation · non-local means.

1 Introduction

Although there has been a significant progress on digital imaging technologies in the last decade, the captured images may still suffer from noise corruption due to the properties of the camera sensors, especially for the low power CMOS image sensors [15]. Image denoising is one of the most important problems in vision-related works and has been studied for decades. Although a remarkable performance has been achieved by existing methods, most of them focus on single image denoising. Due to the limited information that we can obtain from a single input image, even with a sophisticated denoising algorithm may still fail to handle all challenging scenarios (fine structure, low-light, etc.). On the other hand, consumer-level stereo cameras are becoming popularly available on cars and mobile phones, which also suffer from the noise corruption problem. Hence, if images captured by these cameras are used for stereo depth estimation, there may be large disparity errors. Unfortunately, conventional stereo methods do not work well

in noisy conditions, as shown in Fig. 1(c), and works that consider noisy stereo matching are limited. In this paper, we attempt to address these two problems. Given a pair of stereo images captured in noisy conditions (Fig. 1(a)), we propose a joint framework to iteratively optimize image denoising (Fig. 1(b)) and stereo matching (Fig. 1(d)).

The proposed method leverages the external information brought by the stereo image pair and the internal information extracted from across different scales of the image pair for denoising and disparity estimation. First, inspired by the non-local means method [3], which denoises an image by grouping similar patches within it, we group similar patches from both input stereo images and across different scales of the images so as to strengthen the matching confidence for image denoising. Second, we propose a noise-tolerant matching cost for disparity estimation. The denoised image pair is used to update the data cost and improve the accuracy of the estimated disparity. These two tasks (image denoising and disparity estimation) are iteratively optimized in a multiscale framework. The effectiveness of the proposed framework is validated by extensive experimental evaluations. Results show that the proposed method outperforms the state-of-the-art image denoising methods and disparity estimation methods, quantitatively and qualitatively. The proposed method achieves better performance than the state-of-the-art multi-image denoising methods by about 2dB and the conventional stereo matching methods under the radiometric-change condition by about 50%. In addition, we demonstrate other applications of our framework, including image refocusing and view synthesis under noisy conditions.

The main contributions of this paper include:

- We propose a new similarity measurement, which takes advantages of stereo structure information and principal component analysis (SS-PCA), for noisy stereo patch grouping.
- We propose a matching cost that integrates informative edge and subspace information (PCIE) for noisy stereo matching.
- We propose a multiscale framework to jointly denoise images and estimate disparity, by introducing "needle group" and cross-scale stereo matching.

The rest of the paper is organized as follows. We first summarize related works on image denoising and stereo matching in Section 2. We then present our joint framework in Section 3 and experimental results in Section 4. Finally, we discuss depth-aware image processing applications in Section 5 and briefly conclude the paper in Section 6.

2 Related Work

As a comprehensive review of existing works on image denoising and stereo matching is beyond the scope of this paper, we only discuss closest related works to the proposed method on two main aspects: (1) non-single image denoising, i.e., denoising with external information or using multiple images, and (2) stereo matching in noisy conditions.

2.1 Non-single Image Denoising

Image denoising is an ill-posed problem in computer vision and image processing, and has been extensively studied for decades. The aim of image denoising is to recover a clean image from a noisy one. Generally, image denoising techniques can be categorized into *single image based methods* and *external prior based methods*. We refer the readers to [4] and [41] for a review of single image denoising. The major problem of single image based methods is that they have almost reached their bound [25]. As a consequence, a lot of methods proposed in recent years rely on additional information beyond the noisy image itself [5,47,57,7,29,49,50,34,8,28,6,44]. These external prior based methods involve *external learning* or *multiple input images*, which share similar objectives to the proposed method.

External Learning Based Denoising: It has been shown that the theoretical minimum mean square error (MSE) of denoising can be achieved when using a large enough external dataset [25,26]. The expected patch log likelihood (EPLL) method [57] outperforms other generic prior methods, by training a Gaussian Mixture prior using over 50,000 patches sampled from the external training set, while the patch group based prior denoising (PGPD) method [47] shows a better grouping performance, by learning nonlocal self-similarity (NSS) using millions of external patch groups extracted from clean images. In additional to training on generic datasets, specific datasets are also used for application-oriented denoising. Targeted image denoising (TID) [29] uses a targeted external database (text images, human faces, etc.) to learn an optimal filter for denoising, while [49,50] apply image retrieval to construct the targeted database and combine it with the internal information for denoising. Other methods that combine internal and external information for denoising include [34,8,28]. In recent years, neural networks are also used for image denoising by training on a huge external dataset [6,44].

Multiple Image Denoising: In [5], the original NLM is extended to denoise video sequence by treating the video as a union of multiple images. In [31], the state-of-the-art single image denoising method BM3D [9] is extended into 4D space (called BM4D) to deal with volumetric data like MRI images. In [14], an adaptive spatial-spectral dictionary learning method is proposed for hyperspectral image (HSI) denoising. It is considered as multiple image denoising, as each band is a separate image. In [12], a low-rank tensor approximation method is proposed to denoise multi-frame (usually dozens of frames) data like multispectral images

or MRIs. In [22], a total of 124 images of the same scene are captured to denoise a specific mountain view. Some specially designed input [21,42], such as a near-infrared image or dark-flashed image, is also used for denoising in specific scenarios. TID [29], which is proposed for a targeted external dataset, also shows to be able to denoise multiview inputs. In [55,30], correspondences computed from multiple views (up to 25) are used for denoising. They first compute the correspondences between the target view and other views. Based on the correspondences, other views can be considered for denoising the target view. The similar patches are selected according to the initial correspondences and the Euclidean distance. However, it is difficult to estimate correspondences in noisy situations, and easily causes many outliers, which seriously affect similar patch grouping. In addition, only using the Euclidean distance to measure the similarity is not robust for patch selection (see Fig. 2).

When denoising a video sequence, optical flow can be used to gather information across nearby frames. Methods like [36,27] restore a noise-free video by motion estimated from the optical flow. They use several frames in the forward and backward directions to get patches for denoising and have achieved good performance. [27] performs denoising with 11 consecutive frames, and then combines the denoised pixels at each frame to get the denoised result of the target pixel in the current frame. However, the supporting patches are searched only based on inner-frame similarity and denoising is performed separately on each frame. Hence, it does not take full advantage of the multi-frame property. In addition, the optical flow is computed separately before denoising the noisy sequence. [36] proposes a variational formulation that simultaneously solves optical flow and sequence restoration. [36] modifies an optical flow algorithm by adding a fidelity term to penalize the deviation from the given noisy sequence. However, the similarity is measured by L1 and L2 metrics on the original data, which is shown to be not accurate as in Fig. 2. A large number of iterations (600) is also needed in each optical flow. Both above methods require many frames as support.

Summary: Most of the above non-single image denoising methods show excellent results and outperform the single image methods. While non-single image denoising can be considered as a current trend, these existing methods require either an external training set of clean images or a specially designed multiple image input. In contrast, the proposed method is only based on two noisy input images.

2.2 Noisy Stereo Matching

Stereo matching is another important problem in computer vision. It aims to find dense correspondences between two input views captured from the same scene but at different viewpoints [39]. The output of this process is a disparity map, which represents the depth information of the scene and is important to many computer vision applications like robotics and view synthesis. Almost all of the stereo matching algorithms rely on the intensity-consistency assumption, i.e., assuming that two corresponding points in the left and right views have the same intensity. However, in real scenes, a captured stereo pair may easy be corrupted by noise [1], and very few works consider this practical but difficult noisy stereo matching problem.

There have been studies on how to handle radiometric changes in different views [20,2,19,51,18,17,10,23,46,24]. In [20], an evaluation of different cost functions on various radiometric changes is presented for stereo matching, including Absolute Difference (AD), Birchfield and Tomasi (BT)[2], BT with mean filtering, BT with Laplacian of Gaussian (LOG) [19], Rank and Census transform [51], Normalized Cross-Correlation (NCC), and Hierarchical Mutual Information (HMI) [18]. It shows that HMI is the best performer to noisy situation. After [20], research on radiometric changes for stereo matching emerges. Based on NCC, [17] proposes an adaptive normalized cross-correlation (ANCC) cost measure, which is proved to be robust to lighting and illumination changes. In [46], a measure robust to affine illumination changes is proposed. Other methods [10,23] taking advantages of the gradient information are also shown to be effective on radiometric changes. An adaptively weighted descriptor [24] that combines image content attributes, including gradient information, is proposed to handle images with radiometric changes. However, all of the aforementioned methods focus on radiometric (illumination or exposure) changes, in which the gradient information or other structure information can be used for prediction even though image intensity is not consistent between the two views. In contrast, the gradient information for an image pair corrupted by noise is not reliable.

To the best of our knowledge, there are no existing methods designed specifically for noisy stereo matching except [16]. In [16], a new data cost combining the non-local means and perceptually modified Hausdorff distance (PMHD) [37] is proposed to generate the disparity map. A global optimization is used for stereo matching. However, its performance drops at high noise levels and for outdoor scenes (Section 4). Instead of directly extending NLM to two views as in [16], we propose a new patch grouping strategy leveraging the structure information between two views and measuring the similarity in a PCA-optimized dimension. Our evaluations show that the proposed strategy is more robust than directly using L2-norm [16]. In addition, the proposed method utilizes the cross-scale information to increase the accuracy of patch grouping for denoising and expands the support regions for disparity estimation.

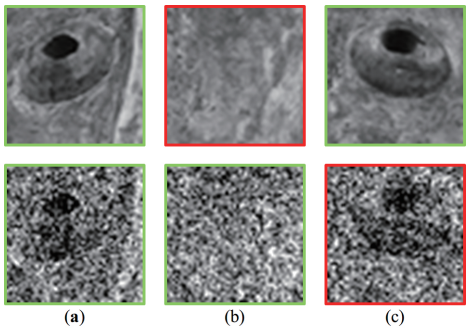


Fig. 2 (a), (b) and (c) are three patches extracted from an image. The patches in the first row are from the clean version of the image, while those in the second row are from the noisy version. In each row, patches with a green border are considered as similar based on Euclidean distance while the patch with red border is dissimilar to the others.

2.3 Non-local Means Filtering

As the proposed framework is mainly inspired by the idea of non-local means (NLM) filtering, we briefly review the non-local means algorithms [3] here. Given a clean image I, the noisy version of I at pixel p can be defined as: $I_n(p) = I(p) + n(p)$, where $n(p)$ is the noise at pixel p. In the denoising problem, the actual observation is the noisy version $I_n(p)$, while the true clean image pixel $I(p)$ is not available in practice. As the noise distribution is supposed to have a zero mean, the average of a sufficiently large number of observations should approach to the clean value. NLM is designed based on this idea by weighted averaging all the pixels non-locally. In its basic form, the weight for averaging is defined as the similarity between the neighborhoods of two pixels. Let N_i denotes the neighborhood centered at pixel i with a window size of $r \times r$, and the Euclidean distance is used to calculate the similarity between two neighborhoods of pixel i and j as $\left\|I_n(N_i) - I_n(N_j)\right\|_2^2$. Then, the NLM algorithm defines the estimate/denoised value for pixel i as:

$$\hat{I}(i) = \sum_{j \in S} \frac{1}{Z(i)} e^{-\frac{\left\|I_n(N_i) - I_n(N_j)\right\|_2^2}{h^2}} \cdot I_n(j), \tag{1}$$

where S is the set of pixels used for averaging. In the original NLM algorithm, S covers the whole image, i.e., a global averaging, but for computational efficiency, S is usually set to a fixed-size search window centered at pixel i. h is an averaging parameter controlling the degree of averaging. $Z(i)$ is the normalizing term summing up all the weights in S.

In general, the NLM algorithm can perform very well due to the redundancy in images. However, the core step, which is also the most difficult problem, of NLM is to find similar patches within the search window. While the Euclidean distance has been shown to be effective in measuring the similarity of two patches in a noise-free situation, it may not well capture the true similarity for noisy images. Fig. 2 shows an example. In the first row, patch (a) is more similar to patch (c) than patch (b). The Euclidean distances for the clean version (first row) are 1780 between (a) and (b), and 1713 between (a) and (c). For noisy version (second row), the Euclidean distances are 3585 between (a) and (b), and 3595 between (a) and (c). This shows that the results for the clean version and the noisy version are contradictory.

3 Proposed Method

In this section, we first give an overview of the proposed method, and then discuss the ideal performance in the noisy stereo situation. Finally, we describe the framework for joint image denoising and disparity estimation.

3.1 Overview

Fig. 3 shows the proposed framework. In order to utilize the cross-scale information, a multiscale pyramid pair is constructed from the noisy input stereo pair. At each scale, we take full advantage of the stereo information for denoising. The structure information among the stereo images is leveraged to group similar patches from two views, and the similarity is measured in a low-dimension through PCA projection. The computed stereo structure and PCAs are combined as SS-PCA for denoising (see Section 3.3).

We then construct a cost volume for each scale based on a new matching cost, referred to as PCIE (principal components of intensity and informative edges). Both the informative edges and the principal components of patch intensity are designed to be robust to noise and intensity changes (see Section 3.4). The disparity is estimated by aggregating the intra-scale and cross-scale costs. Meanwhile, denoising is also performed on both intra-scale and cross-scale only at the finest scale. In addition to the patches selected by SS-PCA, a "needle-group" is added for cross-scale patch grouping (see Section 3.5). The cross-scale operations improve the quality of both denoising and disparity estimation.

After that, we have the initial result of the denoised images and disparity map. As our SS-PCA depends on the disparity map, and the denoised image pair is beneficial to patch grouping, we feed the denoised image pair and disparity map back to the framework to iteratively improve the quality of both denoising and stereo matching.

3.2 Ideal Method for Stereo Scenario

Before discussing the proposed method in detail, we present an ideal method that gives the ideal performance in a noisy stereo condition. The ideal denoiser directly uses the clean image for similar patch selection, while the ideal stereo

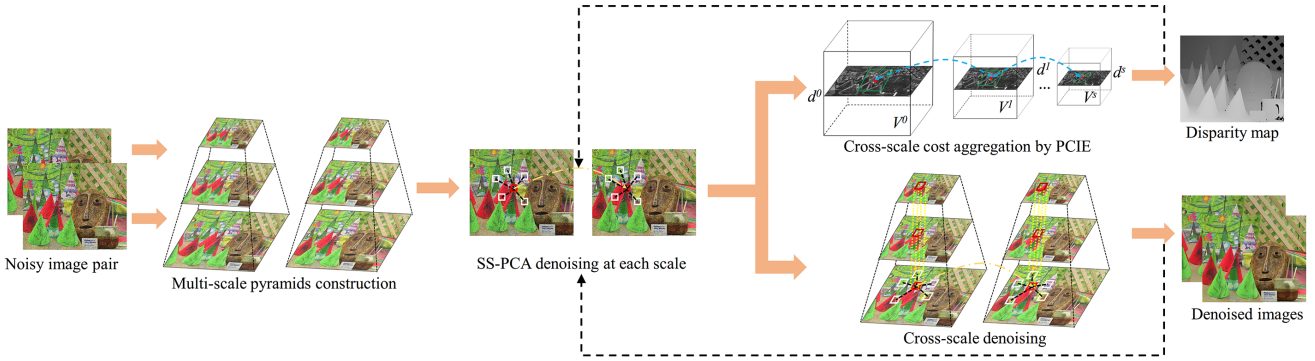


Fig. 3 Overview of the proposed multiscale joint framework.

matcher uses the ground truth disparity map as a guidance on the clean stereo images to obtain the disparity map.

Unlike single image denoising, a stereo image pair provides an additional view for denoising, and the patch selection process can be extended to two views. Since we process the left and right views in a similar way, the following discussion will be based on the left image. Given a stereo image pair I_L and I_R, we denote the neighborhood (patch) surrounding pixel i as N_i, and j as the center pixel of N_j. If N_i is similar to N_j in the left image, their corresponding patches N_{i-d_i} and N_{j-d_j} in the right image should also be similar to each other, where d_i represents the disparity at pixel i. Based on this assumption, the similarity measurement for patch selection is defined as:

$$\left\|I_L(N_i) - I_L(N_j)\right\|_2^2 + \left\|I_R(N_{i-d_i}) - I_R(N_{j-d_j})\right\|_2^2, \tag{2}$$

where d_i and d_j are the ground truth disparities at pixels i and j, respectively. Applying this similarity measurement in the NLM algorithm, we get the ideal denoiser. We refer to this as the "ideal" case because the image pair used for similar patch selection are clean images, and the disparity used in Eq. 2 is the ground truth disparity, which is not available in real applications. On the other hand, the matching cost for stereo matching is usually aggregated by a filtering technique, e.g., BoxFilter (Fig. 4 (g-h)). Ideally, the matching cost should be aggregated on the same ground truth disparity. Thus, given the ground truth disparity as a guidance, the best performance of a specific matching cost can be achieved. It should be noted that only the patch grouping is performed on the clean images and the denoising process is still performed on the noisy images, while the ground truth disparity is only used as a guidance. Example results of the ideal method are shown in Fig. 4. The ideal denoiser (Fig. 4 (e)) performs much better than the state-of-the-art methods. Note that noise can significantly influence the matching result, even for the ideal case.

The ideal method described above is clearly impossible to achieve in practice. However, it serves as a benchmark and presents a direction for improvement. It also demonstrates that a good performance can be achieved once a good similarity measurement is available for denoising, and a cleaner image pair as well as a good cost function are used for disparity estimation.

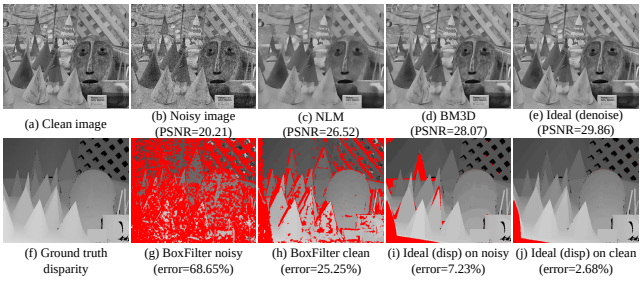


Fig. 4 Sample results of stereo image denoising (first row) and disparity estimation (second row). Results aggregated by the BoxFilter are also included for comparison (g-h). Red regions indicate errors. (e) shows the ideal performance of stereo image denoising, while (i-j) show the ideal disparity estimation performances on noisy and clean image pair.

3.3 Stereo Structure PCA for Denoising

3.3.1 Patch Grouping

In practice, disparity calculated from a noisy image pair is not accurate, as shown in Fig. 4. Hence, the ideal method cannot be directly used here. As shown in Fig. 5(b), both the reference patch (red box) and the neighbor patches (white boxes) in the left view are mapped to the right view according to the initial disparity (shown in Fig. 5(a)). This is referred to as "all-map". Some denoising methods[55, 30] based on multiple images use this kind of mapping to search for similar patches. They use the similarity measurement strategy in Eq. 2 and are based only on the initial estimated disparity map. However, because of a large number of outliers in the initial disparity, many neighbor patches are mapped to wrong locations. As the aim of introducing another view is to increase the patch grouping accuracy, such incorrect mappings produce worse grouping. To address this

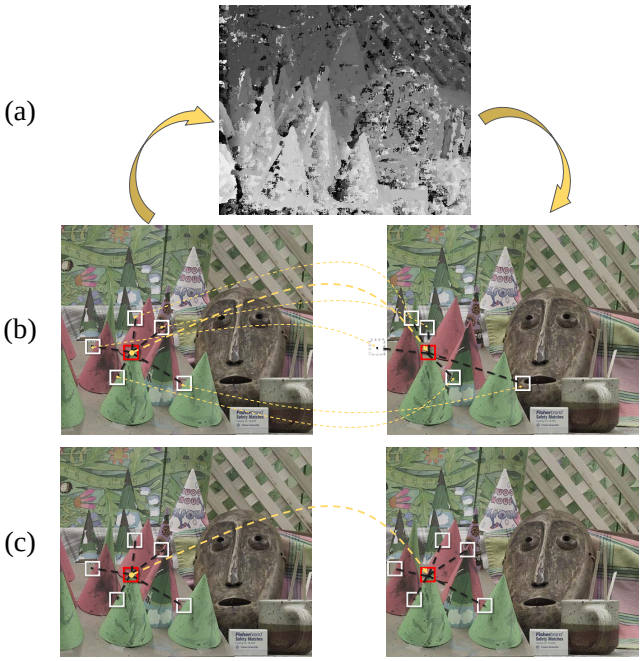


Fig. 5 Patch matching on stereo images. (a) the initial disparity map. (b) the reference patch (red box) and the neighbor patches (white boxes) on the left view are mapped to the right view according to (a). (c) the patch pattern on the left view is mapped to the right view.

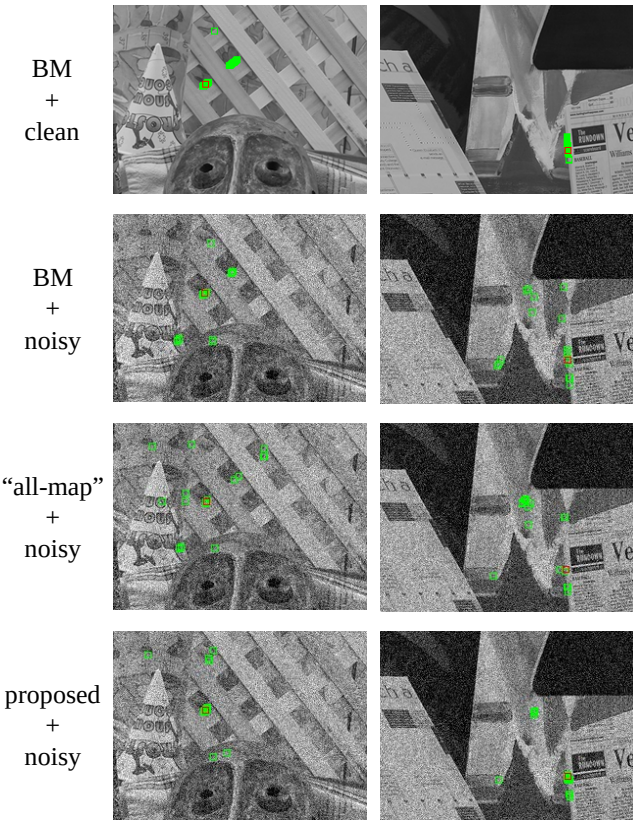


Fig. 6 Comparison of different strategies for selecting similar patches from two stereo pairs (first and second columns). 20 most similar neighbor patches (green boxes) are selected for each reference patch (red box). "all-map" is based on Fig. 5(b), while "proposed" is based on Fig. 5(c). **Better view on a color screen.**

problem, we propose a new patch grouping criterion based on structure correlation between noisy stereo images.

We first estimate an initial disparity map to build a weak relationship between the two stereo images. For simplicity and in order to validate the effectiveness of the proposed scheme, the initial disparity (Fig. 5(a)) is estimated by AD following a BoxFilter aggregation with no refinement. We then take a patch P_i^L centered at pixel i in left view as a reference patch and its neighbor patches as P_j^L. When searching for similar patches in both left and right views, we map the reference patch to the right view according to the initial disparity and get the mapped reference patch P_{i-d}^R. However, unlike the ideal denoiser, we map the "pattern" of the reference patch to the right view. Here, the "pattern" refers to the spatial relationship between the reference patch and the neighbor patches. During the searching process, once a neighbor patch is visited in the left image, its relationship with the reference patch is kept in the right image. As a result, the spatial structure of the neighbor patches with the reference patch in the left view is maintained, and it is duplicated to the right view. Fig. 5(c) shows an example. The similarity between two patches is then defined as:

$$\left\|P_i^L - P_j^L\right\|_2^2 + \left\|P_{i-d}^R - P_{js}^R\right\|_2^2, \tag{3}$$

where P_{js}^R denotes the patch mapped to the right view using the structure correlation derived from the left view. This idea is similar to patch based NLM, which is more robust than the pixel based one [4]. The pattern in Fig. 5(c) can be considered as a "macro patch", leveraging structure information for robustness. As the stereo images are rectified and the cameras are calibrated in advance, the geometry between the two cameras has been taken into account when mapping between the two views. Fig. 6 compares the "all-map" and "pattern-map" strategies. The block matching (BM) result on clean images is included as a benchmark. We can see the patches selected by "all-map" are as random as BM on noisy images, while the proposed structure based mapping has a performance closer to the benchmark.

3.3.2 Patch Similarity

Although the problem caused by low-quality disparity is handled by the stereo structure strategy, the distance between two patches determined by the Euclidean distance (as in [55,30]) is not robust for noisy images, as shown in Fig. 2. Here, we propose a principal component analysis (PCA) based strategy to improve the performance of patch similarity measurement for stereo denoising.

PCA is a traditional decorrelation strategy widely used in dimensionality reduction. It has also been used in image denoising [35,54,11]. When projecting the original data to

a PCA domain, the signal (clean information) and noise can be separated. In addition, by preserving the most significant principal components, noise can be eliminated to some extent. Unlike conventional PCA methods that recover a denoised patch by back-projecting from a lower dimension, in this paper, we simply utilize the projection coefficients in a lower dimension. As the clean information can be well represented in the lower dimension, these coefficients are used to describe patch similarity. Thus, we replace the Euclidean distance on the noisy image patch by the distance between two coefficient vectors of size C ($C \ll r^2$, where C is the number of principal components and r^2 is the patch size). The similarity measurement in Eq. 3 becomes:

$$\left\|\eta(P_i^L)-\eta(P_j^L)\right\|_2^2+\left\|\eta(P_{i-d}^R)-\eta(P_{js}^R)\right\|_2^2, \tag{4}$$

where $\eta(\cdot)$ is a projection function to get the coefficients corresponding to the patch in a lower dimension. In addition to using the new proposed similarity measurement (Eq. 4) for similar patch grouping, we also use it to determine the weights for patch averaging. In the conventional NLM algorithm, the weight is defined as: $e^{-\left\|P_i-P_j\right\|_2^2/h^2}$.

This weight is dominated by the Euclidean distance on noisy images. In the proposed algorithm, the weight between pixels i and j is set to:

$$\begin{aligned} w(i,j) &= e^{-\frac{sim(P_i,P_j)}{h^2}} \\ &= e^{-\frac{\left\|\eta(P_i^L)-\eta(P_j^L)\right\|_2^2+\left\|\eta(P_{i-d}^R)-\eta(P_{js}^R)\right\|_2^2}{h^2}}. \end{aligned} \tag{5}$$

By using the new similarity measurement in Eq. 4 and the new weight in Eq. 5, patch matching accuracy and denoising performance are both improved, as shown in Fig. 7.

In [11], the PCA-based methods are divided into three classes according to the way that the dataset is constructed: global, hierarchical, and local. Global methods are the most efficient ones, while local methods have the highest accuracy. In this paper, we combine the global and local PCAs to leverage both their advantages. Specifically, in the first stage of patch grouping, we use global PCA, while in the following weighted patch averaging stage, we use local PCA. Here, "global" means that the covariance matrix is constructed from all the patches in the image pair, and only needs to be calculated once in advance. "Local" means that for each of the selected patch groups, a covariance matrix is constructed, and each member patch of this group is projected based on this matrix.

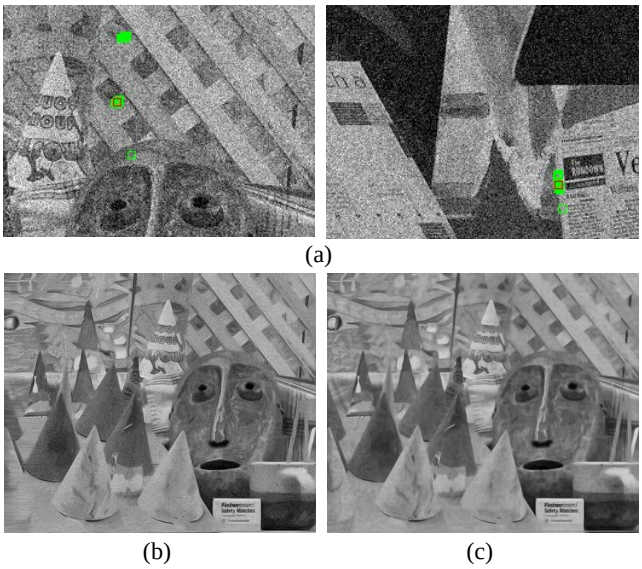


Fig. 7 Improvement by adding the PCA strategy to the similarity measurement, i.e., Eq. 4. Other settings are same as those in Fig. 6. (a) shows the patch matching results. (b) shows the denoised result of using stereo structure as similarity measure, i.e., Eq. 3, with a PSNR=26.69dB. (c) shows the result after adding the PCA, i.e., Eq. 4, with a PSNR=27.15dB.

3.4 Noise-tolerant Cost for Stereo Matching

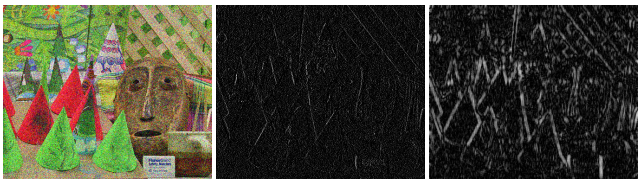
Most existing disparity estimation algorithms are implemented based on the assumption that the intensity of the corresponding pixels is in the left and right views should be consistent. However, this assumption is difficult to guarantee in practice as there are many factors affecting the imaging process of the input images. Most works that handle the intensity-inconsistent problem focus on illumination changes between two views, while the noise-corrupted problem is not well studied. In fact, noise is more commonly observed than illumination changes. As illustrated in Section 3.2, noise has a great impact on the disparity result, even in the ideal case. To achieve a robust disparity result in noisy situation, we propose a new matching cost combining the intensity subspace information and edge similarity.

The matching cost in disparity estimation is also a function measuring the similarity between pixels/patches. Similar to the patch grouping process described in Section 3.3, we apply the idea of principal components here to reduce the impact of noise for matching. The first term of the combined matching cost is defined as:

$$C_{PCAD}(x,d)=\left|\eta(P_x^L)-\eta(P_{x-d}^R)\right|, \tag{6}$$

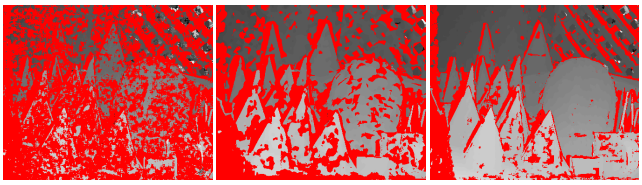
where P_x is a support window (patch) centered at pixel x. P^L and P^R are image patches in the left and right noisy images, respectively. d is the disparity candidate value. $\eta(\cdot)$ is the projection function. PCAD represents the absolute difference of principal components.

Besides PCAD, another important information for similarity measurement in disparity estimation is edge similarity, which has been proved to work well on radiometric changes [20,23,10]. However, when corrupted by noise, the raw edge information (gradient) is not robust. In [48], a cri-



noisy image traditional gradient informative edge

Fig. 8 Effectiveness of the informative edge descriptor. While the traditional gradient detector fails and produces many spikes, the informative edge descriptor can detect the distinctive edges well.



(a) (b) (c)

Fig. 9 Effectiveness of the proposed matching cost, PCIE. (a) disparity from matching cost AD on noisy images. (b) disparity from matching cost PCIE on noisy images. (c) disparity from matching cost AD on clean images. Red regions indicate matching errors larger than 1 pixel.

terion describing informative edges is introduced for blur kernel estimation. It is defined as:

$$iE(x) = \frac{\left| \sum_{y \in P_x} \nabla I(y) \right|}{\sum_{y \in P_x} |\nabla I(y)| + 0.5}, \tag{7}$$

where $\nabla I(x)$ represents the gradient at pixel x, and 0.5 in the denominator is to prevent producing a large iE response in low-texture regions. With this informative edge descriptor, the sum of signed gradients in the numerator exactly eliminates the noise impact, while the sum of unsigned derivatives in the denominator describes how strong the local structure is. Thus, it is robust to noise. Fig. 8 demonstrates the effectiveness of the informative edge descriptor. Hence, we include this descriptor as the second term of our combined matching cost. The informative edge descriptor is defined as:

$$C_{IE}(x,d) = \left| iE(x^L) - iE(x^R - d) \right|. \tag{8}$$

With the above two cost terms, the proposed matching cost PCIE is defined as:

$$C_{PCIE}(x,d) = \alpha C_{PCAD}(x,d) + (1-\alpha) C_{IE}(x,d), \tag{9}$$

where $\alpha \in [0,1]$ is an adjustment parameter used to balance the influence of the principal component term and the informative edge term. It should be noted that as the informative edges lie in the range of $[0,1]$, the principal component term is also normalized to this range. Besides, as the principal component coefficients have already been calculated in the denoising part, the only additional computation is the informative edge. Fig. 9 demonstrates the effectiveness of PCIE. The aggregation of the matching cost is done by BoxFilter with no refinement.

In the previous denoising step, for each patch, we collect a group of similar patches. While the Hausdorff distance (HD) can be used to compare the similarity between each pair of groups, a modified version, the Perceptually Modified Hausdorff Distance (PMHD) [37] for weighted HD is more suitable to our problem as the weights are available. However, considering the computational complexity of PMHD, we propose to use PCIE as our matching cost in this paper.

3.5 Multiscale Joint Stereo Denoising and Disparity Estimation

Humans process visual information in a multiscale manner (i.e., in both coarse and fine scales) [38,32]. This multiscale (or coarse-to-fine) bio-inspiration has been adopted in many computer vision tasks. In image denoising, it has been shown that it is possible to find "clean" versions of almost all the image patches in a noisy image from a coarser scale [56]. Intuitively, an image looks "cleaner" from a distant perspective. For stereo matching, the information from different scales is also leveraged in a cross-scale cost aggregation method [53]. It has also been shown that a cross-scale cost aggregation scheme can reduce the incorrect labellings and produce a more accurate disparity map [43]. As our work focus on noisy stereo images, we therefore apply multiscale in our framework to jointly denoise the stereo images and estimate the corresponding disparity map.

We first downsample the left and right noisy images to construct a multiscale pyramid pair. Assuming the scale ratio is γ, the resolution at scale s is γ^s of the original image. This process can be considered as a downsampling and blurring operation performed in both horizontal and vertical directions. Hence, the noise is reduced as the scale number increases. However, the images are still noisy even in higher scales. As such, we use the proposed SS-PCA denoising algorithm (Section 3.3) to remove the noise in the left and right view pyramids to produce a new denoised pyramid pair.

At the finest scale, besides the cross-view (stereo) correlation, another important information is the cross-scale relationship. Due to the existence of clean versions in a coarser scale [56], we collect a group of patches across all scales with the same patch size as in the finest scale and at the same relative coordinates. Such a group is called a "needle" group [56], as shown in the first row of Fig. 10. As there exists a clean patch in the needle group, similar to the cross-view grouping assumption (Eq. 2), we assume that if two patches are similar to each other, their needle groups should also be similar. The second row of Fig. 10 shows an example to validate this assumption. NLM is used for denoising. Results of with and without "needle grouping" are shown

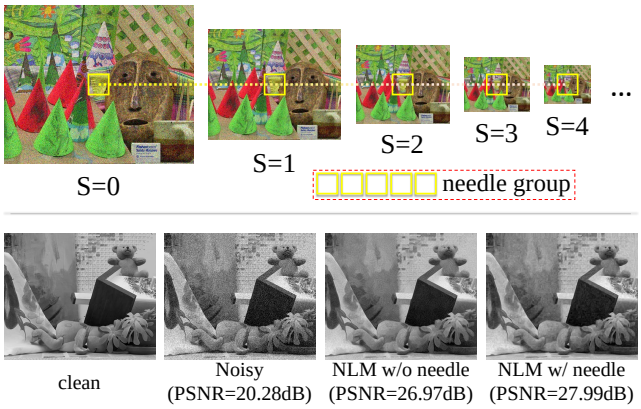


Fig. 10 Needle grouping. First row shows an example of a needle group; second row shows the improvement by using needle grouping.

for comparison. We can see a clear improvement by using needle grouping across scales.

In addition to the cross-view patch group (Eq. 4), for each patch in the left view, we have another needle group. The difference between different needle groups also acts as a component of the similarity measurement. Hence, for patch grouping at the finest scale, the similarity between patches P_i and P_j becomes:

$$\begin{aligned}\left\|\eta(P_i^L)-\eta(P_j^L)\right\|_2^2+\left\|\eta(P_{i-d}^R)-\eta(P_{js}^R)\right\|_2^2\\ +\Psi(P_i^L,P_j^L)+\Psi(P_{i-d}^R,P_{js}^R),\end{aligned} \tag{10}$$

where $\Psi(P_i,P_j)$ is the normalized difference between needle groups P_i and P_j. For simplicity, the difference is depicted by L2-norm. With noise degradation at coarser scales, the proposed SS-PCA is shown to be effective in handling the remaining noise. According to our experiment, adding the needle group to each coarse scale does not contribute to the final result but increases the complexity. Thus, cross-scale denoising is performed only on the finest scale.

From the above denoised pyramids, a cost volume is constructed for each scale, i.e., $(V^0,V^1,\cdots,V^s)$ in Fig. 3. Each cost volume is a cubic structure, in which each point (x,y,d) represents the cost value of pixel (x,y) at disparity d. As the initial denoised pyramid pair still contains residual noise, the cost volume is computed using the proposed PCIE matching cost. For these cost volumes at different scales, a cross-scale aggregation strategy [53] is used to aggregate the inter-scale and intra-scale costs. After the aggregation, a disparity corresponding to the finest scale is estimated, and we refer to it as the initial disparity map. Left-right consistency check [39] is then applied between the left and right disparity maps to reduce the artifacts caused by occlusion.

At this point, we have a pair of initial denoised images and an initially estimated disparity map for the finest scale (i.e., original resolution). For the initial denoised images, noise has been reduced to a large extent, while for the initial disparity map, the accuracy has also been improved. By using these initial images, image patch grouping and averaging can be further improved. As a result, the SS-PCA denoising algorithm can be performed in a more accurate way. On the other hand, the disparity can also be updated. With the stereo image pair becoming "cleaner" and disparity map more accurate, this iterative process approaches to the ideal algorithm mentioned in Section 3.2. Algorithm 1 summarizes all the steps.

Algorithm 1 Multiscale Joint Denoising and Disparity Estimation.

Input: Noisy image pair, I_n^L and I_n^R;
Output: Denoised image pair, $\hat{I}^L and \hat{I}^R$, estimated disparity map, D;

```
Construct multiscale pyramid with scaling factor γ^s;
Denoise each scale pair by SS-PCA (Section 3.3);
if not converge then
    Calculate a cost volume for each scale using PCIE;
    Aggregate cross-scale costs, and generate disparity map D^s;
    Using stereo structure and needle to form patch group P;
    Denoise each patch by P̂(i) = Σ_{j∈P} w(i,j)·P(i);
end if
Denoise the whole image Î by P̂;
```

As the proposed method normally converges in 2-3 iterations, we fix it to two iterations, which is similar to the strategy of some "two-stage" methods [9,50]. Note that in our implementation, each iteration utilizes the same pipeline, instead of using different schemes, e.g., Wiener filtering. In the first iteration, the disparity map for denoising at each scale is estimated by AD + BoxFilter without any refinement. We have also tried other methods (e.g., HMI [18]) for estimating the initial disparity. Although HMI can achieve a slightly better initial disparity (3% error), the final result is similar to that of using AD + BoxFitler (less than 0.01dB for denoising and 0.1% for disparity estimation). Thus, for simplicity, we use AD + BoxFilter in the initialization. Fig. 11 shows the effectiveness of each step for denoising. We can see that the performance of the proposed framework improves continuously with each step.

4 Experimental Results

In this section, we evaluate the proposed method qualitatively and quantitatively through a number of experiments. As our work focuses on noisy stereo images, our experiments are conducted based on two popular stereo datasets, Middlebury [40] and KITTI [33]. The evaluation is divided into two parts, the denoising performance and disparity estimation performance. The images are corrupted by white Gaussian noise at noise levels: 25, 35, 45 and 55. In addition, we also report the performance on images corrupted by real noise.

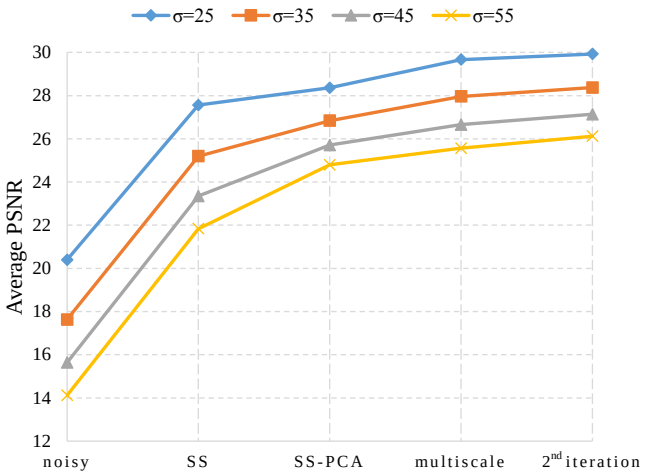


Fig. 11 Denoising performance of each step on the Middlebury dataset.

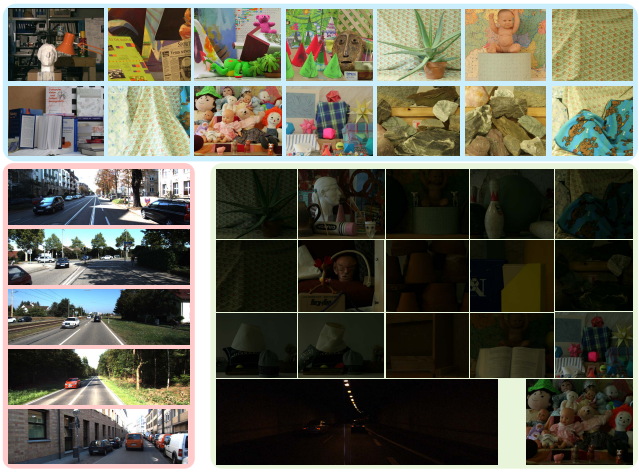


Fig. 12 Evaluation test sets: the Middlebury indoor test set (blue), the KITTI outdoor test set (red), and the low-light test set (green). Only left view images are shown here.

4.1 Dataset

The Middlebury dataset is captured in indoor scenes with various difficulties, e.g., low-texture and occlusion, that may occur in real scenes, while the KITTI dataset is captured using a car equipped with a LiDAR sensor and color cameras on challenging outdoor scenes, e.g., thin objects, shadows, slant surfaces and specular reflections. Hence, the stereo images from Middlebury form our indoor test set, and data from KITTI form our outdoor test set. Besides, as noise often appears in low-light conditions, we have also constructed a test set consisting of low-light images, which are obtained from the Middlebury and KITTI datasets. Fig. 12 shows example images from these three test sets. For the indoor test set, randomly selected images are denoted as $i1, i2, \ldots, i14$ from top-left to bottom-right; for the outdoor test set, the first frame of each scene in the training subset is selected and denoted as $o1, o2, \ldots, o200$; for the low-light test set, the images are denoted as $l1, l2, \ldots, l17$ from top-left to bottom-right. Section 4.5 discusses details of the images captured in real-noise scenes.

4.2 Parameter Setting

In our experiments, the patch size r and the searching window size S are set to 7 and 19, respectively. The number of similar patches for patch grouping between the left and right views is set to 18. Similar to NLM [3], the averaging parameter h for the weights is set according to the noise level as $h = 6\sigma + 14$. The number of principal components in patch grouping of SS-PCA is set to 3. The pyramid size s and the scaling factor γ are set to 3 and 0.5, respectively. All the parameters are tuned on the Middlebury dataset and applied to all three test sets.

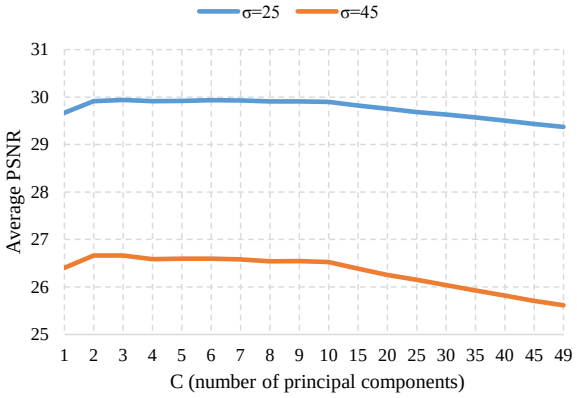


Fig. 13 Denoising performances w.r.t. different numbers of principal components at two different noise levels.

The number of principal components C for SS-PCA is set experimentally. The relationship between denoising performance and the number of principal components C is illustrated in Fig. 13. We show two noise levels ($\sigma = 25, 45$) based on the images from the Middlebury dataset. The denoising performance is from the proposed SS-PCA algorithm and all other parameters are fixed. At the right end of the figure, the dimensionality of the subspace of PCA is equal to the total number of pixels in a patch, which means that no PCA is performed. We can also see from the figure that the best performance occurs at a relatively low dimension, around $C = 3$, and the performance decreases as C is away from the point $C = 3$. Hence, we set the dimensionality parameter C to 3. From the two curves in Fig. 13, we can observe that the proposed method performs similarly under different noise levels, demonstrating its robustness to noise levels. Other parameters are also set experimentally.

4.3 Evaluation on Denoising Performance

As the proposed method focuses on stereo scenarios, in evaluating the denoising performance, we mainly compare it to the state-of-the-art non-single image denoising methods, including the ideal stereo denoiser (Section 3.2), BM4D [31] used for multi-image denoising, TID [29] for targeted external database, and PMHD [16] for stereo images. Note that the volumetric data of BM4D consists of the noisy image pair, while one of the noisy images for TID acts as the supporting database. In addition to the above non-single denoising methods, as our algorithm is inspired by NLM [3], we also compare it with NLM. Another single-image denoising method BM3D [9] is also included, as it is considered as the state-of-the-art single-image denoising method. PMHD is based on our own implementation, and the other methods are based on the authors' codes.

Indoor: The comparison results are presented in Tables 1 and 2. From the tables, we can see that the ideal denoiser is the best performer in all cases, as expected. The proposed method clearly outperforms NLM by over **3dB** at both low and high noise levels. It even outperforms the state-of-the-art single image denoising method BM3D. The specific designed multiple-image denoiser BM4D, TID and PMHD perform not as good as the proposed method. Specially, TID [29] is reported to significantly outperform state-of-the-art single image denoising methods. However, it was based on the assumption that only one image was corrupted by noise. When both views are corrupted by noise as in our experiment, its performance is not as good. Compared with the multi-image denoisers, the proposed method outperforms them by **1.3dB** to **3.8dB** on average. This shows that the proposed method is effective and robust to high noise levels.

Fig. 14 shows a qualitative comparison of the above denoising methods. The denoising results of images $i1$ and $i5$ when noise level σ is 35 are presented. Selected regions (red and green boxes) are magnified for comparison. As we can see, when compared with other methods, the proposed method reduces the noise significantly and preserves more fine details. Specifically, in image $i1$, the proposed method reconstructs the camera handle well, while other methods either smooth it out or fail. The proposed method also performs better on the edges of the statue and the characters on the book spine. While for image $i5$, the proposed method faithfully reconstructs the pattern on the background cloth, as well as the crotch of the aloe. In summary, the proposed method performs well in indoor scenes.

Outdoor: The KITTI outdoor dataset is more realistic and challenging than the Middlebury indoor dataset, since various factors in an outdoor environment could influence the performance of denoising as mentioned in Section 4.1. In Fig. 15(top), we compare the performances of both single-image and multi-image denoising methods over a range of noise levels defined in Section 4.1. Note that for the outdoor KITTI dataset, the ground truth disparity is not a dense map. To show the result of the ideal denoiser, we use the disparity computed from a learning based method [52] (on clean images), which is considered as the state-of-the-art stereo matching method that produces a low error on KITTI.

From Fig. 15(top), we observe that the proposed method consistently outperforms all the other denoising methods, except the ideal denoiser. Specifically, it outperforms BM3D and BM4D on average by **1.09 dB** and **1.83dB**, respectively, over the noise levels. As the noise level increases, the proposed method performs even better compares with other denoising methods. For example, when $\sigma = 55$, the proposed method is **4.52dB** better than NLM, and **1.67dB** better than BM3D. Fig. 14 shows some visual results. We can see that the proposed method reconstructs the details and edges well. In summary, this experiment shows that the proposed method performs well in outdoor scenes.

Low-light: As images captured under low-light conditions tend to be noisy, in this experiment, we compare the proposed method with other state-of-the-art methods, by adding noise of low and high levels ($\sigma = 25, 55$) to the original test set, as shown in Fig. 15(bottom). We can see that under low-light conditions, the proposed method outperforms other methods even more than under normal-light conditions. For example, at $\sigma = 25$, it outperforms NLM and PMHD by **4.04dB** and **2.85dB**, respectively. At $\sigma = 55$, it even outperforms them by **6.1dB** and **5.38dB**, respectively. Fig. 14 shows some examples for visual evaluation. Image $l17$ is corrupted by noise at $\sigma = 55$, while $l16$ is at $\sigma = 25$. From the visual results, even at a large noise, the proposed method can still reconstruct the structure and details under low-light conditions, e.g., the face of the doll and the traffic lines on the road.

4.4 Evaluation on Disparity Estimation

To evaluate the quality of the estimated disparity map, error percentage is the most commonly used criterion defined as:

$$Err(D) = \frac{\sum_{i=1}^{N} (|D(i) - GT(i)| > \delta)}{N}, \tag{11}$$

where D is the estimated disparity map and GT is the ground truth. δ is an error threshold set to 1.0 and 3.0 for the Middlebury and KITTI datasets, respectively, by default. N is the total number of pixels in the image. The summation in the numerator is to sum up the number of pixels which disparity errors exceed the error threshold.

In this paper, we use the error percentage to evaluate the performance of the algorithms and all the regions in the disparity map are taken into consideration. As our algorithm focuses on the noisy scenario, we compare the proposed method with some state-of-the-art stereo matching

Table 1 Comparison of denoising performance (PSNR) with six other algorithms on the indoor test set, at $\sigma = 25, 35$.

σ	25							35						
Image	NLM	BM3D	BM4D	TID	PMHD	Proposed	Ideal	NLM	BM3D	BM4D	TID	PMHD	Proposed	Ideal
*i*1	27.35	29.89	29.03	27.46	28.31	**30.66**	31.50	25.11	27.74	26.91	25.06	26.13	**28.92**	29.77
*i*2	27.89	29.53	28.49	27.63	28.56	**30.13**	31.33	26.18	27.92	26.90	25.31	26.99	**28.49**	29.94
*i*3	27.97	29.52	28.69	27.06	28.53	**30.19**	31.24	26.34	27.99	27.22	24.90	27.06	**28.70**	29.91
*i*4	26.52	28.07	27.10	26.09	27.22	**28.73**	29.86	24.89	26.58	25.71	24.24	25.80	**27.37**	28.71
*i*5	25.18	27.05	26.57	26.26	26.45	**27.89**	29.35	24.00	25.95	25.55	24.42	25.30	**26.78**	28.28
*i*6	28.25	30.43	29.67	27.76	29.32	**30.99**	32.45	27.51	29.24	28.66	25.68	28.12	**29.61**	31.12
*i*7	24.19	26.37	25.79	26.01	25.78	**27.10**	29.11	23.25	25.17	24.73	23.99	24.61	**26.05**	28.10
*i*8	28.71	**31.42**	30.05	25.45	29.75	31.38	32.75	26.45	29.71	28.45	23.94	28.04	**29.75**	31.13
*i*9	25.20	27.29	26.81	25.97	26.81	**28.03**	29.83	24.04	26.01	25.66	24.34	25.53	**27.06**	28.73
*i*10	27.15	29.32	28.60	25.99	28.42	**29.85**	31.20	24.89	27.89	27.10	24.35	26.90	**28.62**	29.77
*i*11	29.00	31.36	30.24	27.57	29.78	**31.80**	32.95	26.89	29.77	28.75	25.55	28.07	**30.15**	31.30
*i*12	27.52	29.68	29.11	27.13	28.73	**30.33**	31.63	25.98	28.43	27.89	25.05	27.38	**29.11**	30.39
*i*13	27.37	29.31	28.80	26.99	28.49	**29.98**	31.33	25.92	28.12	27.68	25.11	27.20	**28.81**	30.17
*i*14	27.10	29.17	28.40	26.45	28.23	**29.80**	31.30	25.10	27.81	27.01	24.60	26.87	**28.54**	30.17
Ave.	27.10	29.17	28.38	26.70	28.17	**29.78**	31.13	25.47	27.74	27.02	24.75	26.71	**28.43**	29.82

Table 2 Comparison of denoising performance (PSNR) with six other algorithms on the indoor test set, at $\sigma = 45, 55$.

σ	45							55						
Image	NLM	BM3D	BM4D	TID	PMHD	Proposed	Ideal	NLM	BM3D	BM4D	TID	PMHD	Proposed	Ideal
*i*1	23.53	25.82	25.26	23.27	24.44	**27.58**	28.35	20.73	24.34	23.88	21.85	23.11	**26.48**	27.17
*i*2	24.86	26.41	25.71	23.54	25.71	**27.17**	28.80	22.49	25.35	24.69	22.06	24.53	**26.10**	27.81
*i*3	25.11	26.70	26.13	23.24	25.85	**27.47**	28.80	22.86	25.78	25.22	21.90	24.81	**26.46**	27.81
*i*4	23.78	25.34	24.77	22.64	24.66	**26.32**	27.73	21.89	24.56	24.03	21.46	23.74	**25.44**	26.89
*i*5	23.16	25.08	24.83	23.04	24.46	**26.08**	27.49	22.40	24.39	24.19	21.84	23.76	**25.43**	26.76
*i*6	27.02	28.44	27.96	24.17	27.30	**28.48**	30.02	26.57	27.89	27.36	22.68	26.67	**27.45**	29.00
*i*7	22.87	24.36	24.12	22.66	23.89	**25.36**	27.39	22.60	23.83	23.69	21.48	23.39	**24.70**	26.76
*i*8	24.65	28.28	27.19	22.81	26.64	**28.44**	29.78	23.10	27.13	26.12	21.57	25.46	**27.32**	28.62
*i*9	23.27	24.95	24.80	22.97	24.60	**26.32**	27.89	22.59	24.08	23.95	21.76	23.78	**25.62**	27.12
*i*10	23.21	26.56	25.99	22.90	25.56	**27.50**	28.61	21.86	25.54	25.03	21.58	24.44	**26.52**	27.61
*i*11	25.34	28.59	27.68	23.90	26.78	**28.81**	29.92	24.15	27.66	26.82	22.38	25.81	**27.63**	28.72
*i*12	24.80	27.47	26.99	23.53	26.31	**28.01**	29.31	23.78	26.66	26.22	22.15	25.43	**26.97**	28.31
*i*13	24.89	27.18	26.85	23.47	26.18	**27.76**	29.15	23.99	26.43	26.14	22.17	25.37	**26.79**	28.20
*i*14	23.85	26.61	25.96	23.09	25.69	**27.47**	29.20	22.86	25.65	25.08	21.86	24.70	**26.53**	28.31
Ave.	24.31	26.56	26.02	23.23	25.58	**27.34**	28.74	22.99	25.66	25.17	21.91	24.64	**26.39**	27.79

Fig. 14 Qualitative results on the three test sets (indoor (blue), outdoor (red), and low-light (green) corresponding to the first two, following two, and the last two images in row one). Numbers under each method represent the PSNR values for the outdoor/low-light images.

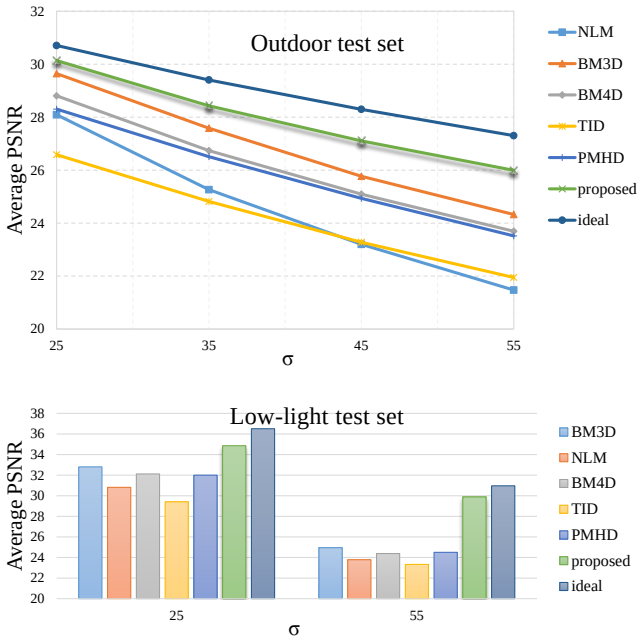


Fig. 15 Quantitative evaluation on the outdoor test set (top) and the low-light test set (bottom), over a range of noise levels.

methods that are robust to radiometric changes: HMI [18] (which is shown to perform better on image noise [20]), ANCC [17] (which is also reported to be robust against noise variations), ADSM [24] (which is the state-of-the-art radiometric-change-resistant method), and PMHD [16] for noisy stereo images. In addition to these state-of-the-art radiometric-robust methods, we also compare with some sequential implementations of denoising followed by stereo matching to validate the effectiveness of the proposed joint framework. The first one (NLM + cross-scale stereo matching), denoted as 'NLM+CS', is similar to the unroll of our approach. The second one (MLP+CS) is a state-of-the-art denoising method MLP (multi layer perceptron) [6] followed by cross-scale stereo matching. As the MRF-based global stereo matching method is supposed to be robust to radiometric changes like noise, the third sequential baseline (NLM+MRF) applies an MRF-based method [13] to image pair denoised by NLM. Again, PMHD is based on our implementation, while the other methods are based on the authors' own codes.

Fig. 16 shows quantitative and qualitative results on the indoor test set. We can see that the error percentage of the proposed method outperforms other methods significantly, and performs consistently across all noise levels. Although ANCC [17] is supposed to be robust to various illumination and exposure changes, only the right images were contaminated by noise (while the left images were kept clean) in their experiments. In practice, if one image is contaminated by noise, the other image is very likely to be similar. Hence, in our experiment, both images are corrupted by noise. Under this situation, ANCC fails to estimate a disparity map with reasonable accuracy. HMI and PMHD show similar performances, while PMHD performs better on higher noise levels. HMI was shown to perform best in noisy scenarios among the selected matching costs in [20]. However, at high noise levels, PMHD performs slightly better than HMI as it utilizes the non-local information for matching, which is more robust to noise. ADSM has a similar performance to HMI and PMHD. We also observe that NLM+MRF performs better than other non-sequential methods. The unrolled sequential setup NLM+CS outperforms the above methods, but increases in error percentage as the noise level increases. MLP+CS performs better than the NLM+CS method, but also increases in error percentage as the noise level increases. On the other hand, the proposed method performs significantly better than all the above methods, and its performance is nearly independent of the noise level.

The outdoor test set from KITTI was originally proposed for stereo matching, including several practical factors that influence the matching between two views. The first row of Fig. 17 compares the performances of different methods on KITTI over a range of noise levels. The second row compares their visual qualities. We can see that the result is similar to that of the indoor test set. The sequential methods perform better than the other non-sequential methods, with NLM+CS performing slightly better than MLP+CS here. The proposed method continues to outperform other methods, and has an approximately **50%** reduction in error percentage over NLM+CS (the best performing sequential method). This demonstrates the effectiveness of the proposed method on outdoor scenes.

Disparity estimation is difficult under the low-light condition, as the image contrast becomes low. When corrupted by noise, it becomes even more difficult to estimate the correspondences. As in the denoising experiment, we evaluate the proposed method using two representative noise levels ($\sigma = 25, 55$), as shown in Fig. 18. We can see that the proposed method outperforms all the other methods at both low and high noise levels, and has a lower error percentage than the sequential methods by more than **20%**.

To evaluate the effectiveness of the proposed PCIE matching cost, in this experiment, we keep other settings the same but replace PCIE with other matching costs: HMI [18], PCAD-only and IE-only. As shown in Fig. 19, both HMI and IE-only perform similarly and their performances decrease linearly with the increase in the noise level. Although HMI is computed based on the mutual information, it still depends on the intensity. Likewise, IE-only is based on edge information. As the noise level increases, the intensity and the edge information are affected and thus, the performances of HMI and IE-only degrade. In general, PCAD-only performs better than HMI and IE-only. Its performance is less affected

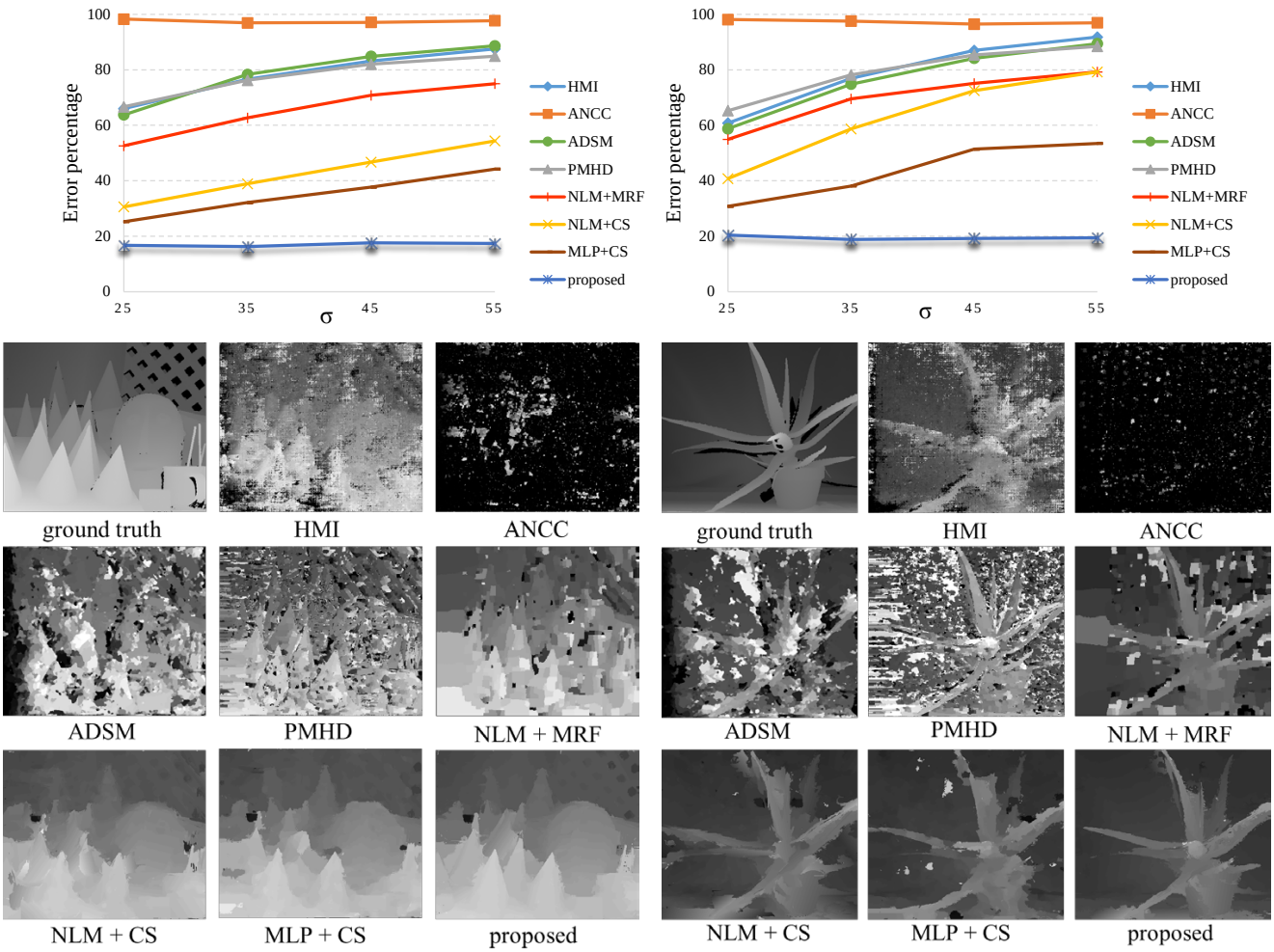


Fig. 16 Performance comparison of disparity estimation on the indoor test set. First row shows the quantitative comparison, while second row shows the visual comparison at $\sigma = 25$. The left/right diagrams on the first row correspond to the left/right images in the second row.

by the noise level, as it is computed in a lower-dimension projected by PCA. On the other hand, the proposed PCIE, which combines PCAD and IE, performs consistently better than all the other matching costs. Its performance is also nearly independent of the noise level.

4.5 Performance on Images with Real Noise

In addition to the evaluations on public datasets with synthetic noise, we have also tested the performance of the proposed method on images corrupted by real noise. The test set on real noise is captured by a stereo camera. Here, we experiment with two image pairs, referred to as *console* and *shelf*, as shown in Fig. 20. As their noise levels are unknown, a method similar to [27] is applied to do the noise estimation. Fig. 20 compares the denoising performance of the proposed method with BM3D and PMHD. We can see that the proposed method performs well on images corrupted by real noise. For example, the noise on the floor of *console* and on the pictures of *shelf* is significantly reduced. Fig. 21 compares the disparity estimation performance of the proposed method with HMI and NLM+CS. The proposed method can estimate the disparity well in low-texture regions and on object edges, as compared with the other two methods. This demonstrates that the proposed method also performs well on the captured noisy data.

5 Applications

There are many depth-aware applications for stereo scenarios. They typically leverage both the intensity/color images and the depth/disparity image as input to generate some visual pleasing or synthetic results. To demonstrate the effectiveness of the proposed method, we discuss two applications: digital refocusing and virtual view synthesis. The only input to both applications is a stereo pair of noisy images.

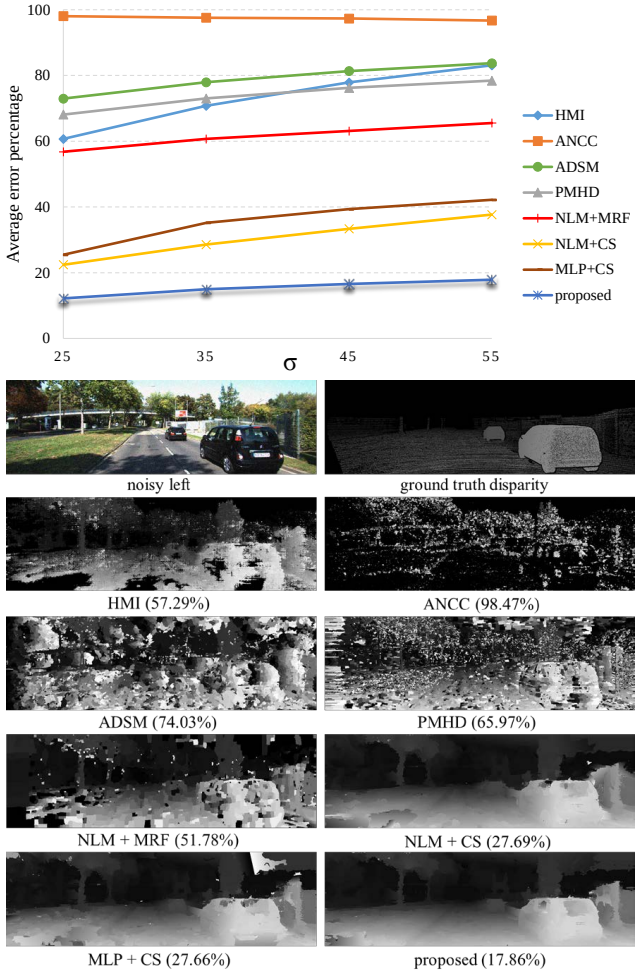


Fig. 17 Performance comparison of disparity estimation on the outdoor test set. Top row shows the average error percentages while bottom row shows a visual example (noise level is set to 25). Numbers in brackets are the error percentages.

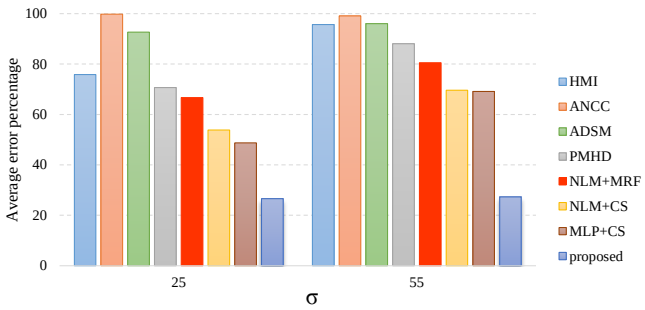


Fig. 18 Disparity performance on the low-light test set.

5.1 Digital Refocusing

Digital refocusing or time-shift photography is an image editing strategy in computational photography. It is a process to simulate the focusing of a camera as an image is being captured. However, unlike focusing with a hardware camera, digital refocusing can shift the in-focus region to anywhere on the image, at least theoretically. It is one of

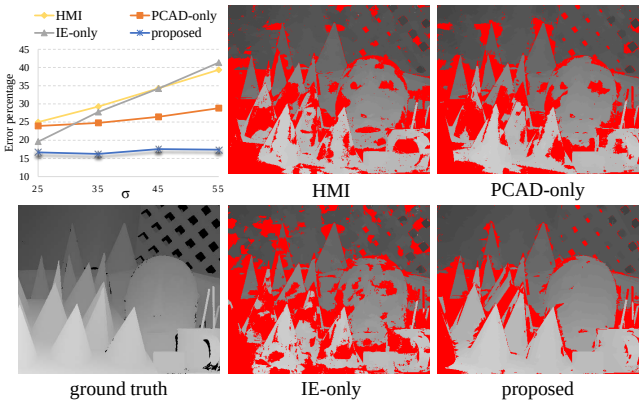


Fig. 19 Quantitative and qualitative performance comparisons on different matching costs.

the most popular depth-aware image processing techniques. Once an accurate depth is acquired, we may set to focus on a specific depth range, producing the depth-of-field effect. However, in a noisy situation, the captured images will be noisy and the estimated depth from these noisy images will also have many outliers. As such, traditional refocusing methods will not perform well. Since the proposed method aims at addressing the problems due to the input noisy images, it can be directly applied here. The digital refocusing result is shown in Fig. 22. To better imitate the real depth-of-field effect produced by a camera, we apply the physical model in [45] to blur the out-focus regions. From the result, we can see that the proposed method produces much cleaner and better depth-of-field effect than the traditional method.

5.2 Virtual View Synthesis

Another popular depth-aware application is virtual view synthesis or depth-based image rendering (DBIR). It uses one color image and the corresponding depth/disparity map as inputs and generates another or several virtual views. The intensity/color information is warped to other pixel positions based on the corresponding depth values. This is usually applied to the 1-to-N or 2-to-N video system for synthesizing novel views and saves the hardware cost. Although there are many methods proposed for virtual view synthesis based on depth information, they all assume clean input images. Fig. 23 shows an example result in synthesizing novel views from noisy input images. The proposed method produces far fewer artifacts in the synthesized images than those from the traditional method.

6 Conclusions

In this paper, we have proposed a joint framework to iteratively optimize image denoising and disparity estimation on noisy stereo images. We achieve this by mutually

Fig. 20 Denoising performance on images captured with real noise. The first row is the results of image *console*, while the third row is the results of image *shelf*. Enlarged regions are shown in the second and fourth rows.

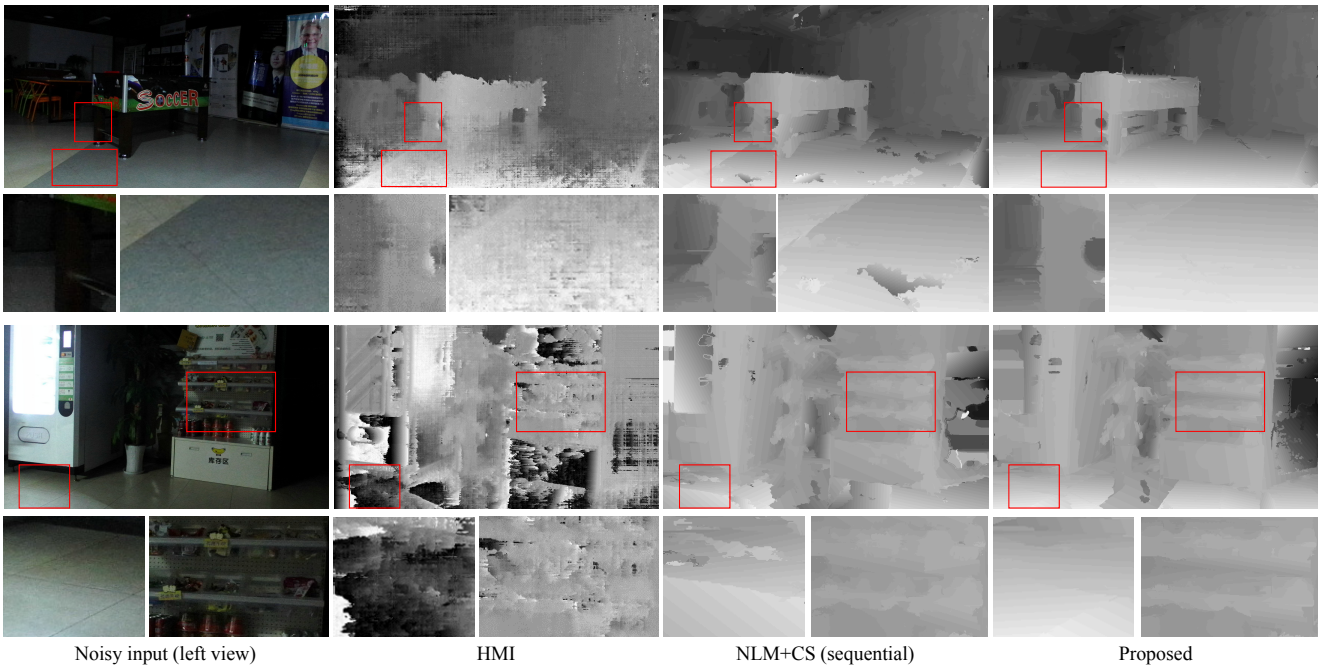


Fig. 21 Disparity estimation performance on images *console* (first row) and *shelf* (third row) captured with real noise. Enlarged regions are shown in the second and fourth rows.

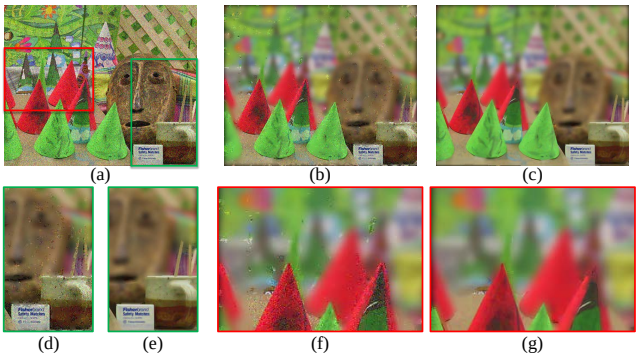


Fig. 22 Digital refocusing results. (a) left input noisy view. (b) using a traditional method. (c) refocusing with the proposed method. (d) and (f) are two cropped regions from (b), while (e) and (g) are the corresponding regions from (c).

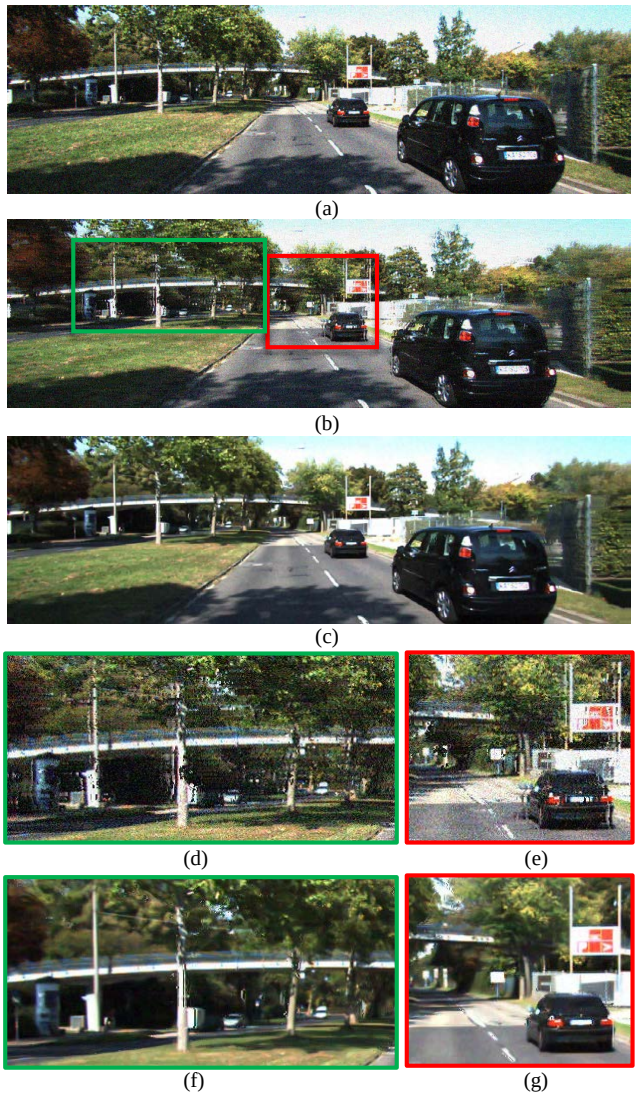


Fig. 23 Results of virtual view synthesis. (a) left noisy view. (b) traditional method. (c) the proposed algorithm. (d)-(e) are two magnified regions from (b), while (f)-(g) are the corresponding regions from (c).

leveraging the information from the stereo images and the estimated disparity. The proposed method is mainly based on the idea from non-local means, which groups similar patches non-locally. Unlike single-image denoising, we propose a new stereo structure algorithm to group similar patches from both intra and inter stereo images. For disparity estimation, we propose a combined matching cost to better describe the similarity in noisy situation. In addition, a multiscale framework is proposed to integrate denoising and disparity estimation, in which both performances are iteratively improved. We have evaluated the effectiveness of the proposed method both quantitatively and qualitatively through different test sets (indoor, outdoor and low-light). Experimental results show that the proposed method outperforms both single-image and multi-image denoising methods, and achieves better performance than the state-of-the-art radiometric-change-robust stereo matching methods. The proposed method also performs well on images with real noise. As the proposed framework leverages stereo information for denoising while simultaneously improving the disparity quality, it can be considered as a generic framework, in which each part can be adapted for specific operations (e.g., BM3D with Wiener filter or multiple images with optical flow).

As a future work, we plan to migrate the proposed framework to the mobile environment to support mobile applications.

Acknowledgements We would thank the anonymous reviewers for their constructive suggestions and insightful comments. This work is partially supported by the Hong Kong PhD Fellowship Scheme (HKPFS) from the RGC of Hong Kong, a SRG grant from City University of Hong Kong (No. 7004416), and a GRF grant from the RGC of Hong Kong (PolyU 152124/15E).

References

1. Alter, F., Matsushita, Y., Tang, X.: An intensity similarity measure in low-light conditions. In: ECCV (2006)
2. Birchfield, S., Tomasi, C.: A pixel dissimilarity measure that is insensitive to image sampling. IEEE TPAMI **20**(4), 401–406 (1998)
3. Buades, A., Coll, B., Morel, J.: A non-local algorithm for image denoising. In: CVPR, pp. 60–65 (2005)
4. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Modeling & Simulation **4**(2), 490–530 (2005)
5. Buades, A., Coll, B., Morel, J.M.: Nonlocal image and movie denoising. IJCV **76**(2), 123–139 (2008)
6. Burger, H., Schuler, C., Harmeling, S.: Image denoising: Can plain neural networks compete with BM3D? In: CVPR, pp. 2392–2399 (2012)
7. Chan, S., Zickler, T., Lu, Y.: Monte Carlo non-local means: Random sampling for large-scale image filtering. IEEE TIP **23**(8), 3711–3725 (2014)
8. Chen, F., Zhang, L., Yu, H.: External patch prior guided internal clustering for image denoising. In: ICCV, pp. 603–611 (2015)

9. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE TIP **16**(8), 2080–2095 (2007)
10. De-Maeztu, L., Villanueva, A., Cabeza, R.: Stereo matching using gradient similarity and locally adaptive support-weight. Pattern Recognition Letters **32**(13), 1643–1651 (2011)
11. Deledalle, C., Salmon, J., Dalalyan, A.: Image denoising with patch based PCA: local versus global. In: BMVC, pp. 425–455 (2011)
12. Dong, W., Li, G., Shi, G., Li, X., Ma, Y.: Low-rank tensor approximation with laplacian scale mixture modeling for multiframe image denoising. In: ICCV, pp. 442–449 (2015)
13. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. IJCV **70**(1), 41–54 (2006)
14. Fu, Y., Lam, A., Sato, I., Sato, Y.: Adaptive spatial-spectral dictionary learning for hyperspectral image restoration. IJCV pp. 1–18 (2016)
15. Healey, G., Kondepudy, R.: Radiometric ccd camera calibration and noise estimation. IEEE TPAMI **16**(3), 267–276 (1994)
16. Heo, Y., Lee, K., Lee, S.: Simultaneous depth reconstruction and restoration of noisy stereo images using non-local pixel distribution. In: CVPR, pp. 1–8 (2007)
17. Heo, Y., Lee, K., Lee, S.: Robust stereo matching using adaptive normalized cross-correlation. IEEE TPAMI **33**(4), 807–822 (2011)
18. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE TPAMI **30**(2), 328–341 (2008)
19. Hirschmüller, H., Innocent, P., Garibaldi, J.: Real-time correlation-based stereo vision with reduced border errors. IJCV **47**(1), 229–246 (2002)
20. Hirschmüller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. IEEE TPAMI **31**(9), 1582–1599 (2009)
21. Honda, H., Timofte, R., Gool, L.V.: Make my day - high-fidelity color denoising with near-infrared. In: CVPR Workshops, pp. 82–90 (2015)
22. Joshi, N., Cohen, M.: Seeing Mt. Rainier: Lucky imaging for multi-image denoising, sharpening, and haze removal. In: ICCP, pp. 1–8 (2010)
23. Jung, I., Sim, J., Kim, C., Lee, S.: Robust stereo matching under radiometric variations based on cumulative distributions of gradients. In: ICIP, pp. 2082–2085 (2013)
24. Kim, Y., Koo, J., Lee, S.: Adaptive descriptor-based robust stereo matching under radiometric changes. Pattern Recognition Letters **78**, 41–47 (2016)
25. Levin, A., Nadler, B.: Natural image denoising: Optimality and inherent bounds. In: CVPR, pp. 2833–2840 (2011)
26. Levin, A., Nadler, B., Durand, F., Freeman, W.: Patch complexity, finite pixel correlations and optimal denoising. In: ECCV, pp. 73–86 (2012)
27. Liu, C., Freeman, W.: A high-quality video denoising algorithm based on reliable motion estimation. In: ECCV, pp. 706–719 (2010)
28. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.: Image-specific prior adaptation for denoising. IEEE TIP **24**(12), 5469–5478 (2015)
29. Luo, E., Chan, S., Nguyen, T.: Adaptive image denoising by targeted databases. IEEE TIP **24**(7), 2167–2181 (2015)
30. Luo, E., Chan, S., Pan, S., Nguyen, T.: Adaptive non-local means for multiview image denoising: Searching for the right patches via a statistical approach. In: ICIP, pp. 543–547 (2013)
31. Maggioni, M., Katkovnik, V., Egiazarian, K., Foi, A.: Nonlocal transform-domain filter for volumetric data denoising and reconstruction. IEEE TIP **22**(1), 119–133 (2013)
32. Menz, M., Freeman, R.: Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism. Nature Neuroscience **6**(1), 59–65 (2003)
33. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR, pp. 3061–3070 (2015)
34. Mosseri, I., Zontak, M., Irani, M.: Combining the power of internal and external denoising. In: ICIP, pp. 1–9 (2013)
35. Muresan, D., Parks, T.: Adaptive principal components and image denoising. In: ICIP, pp. 101–104 (2003)
36. Nir, T., Kimmel, R., Bruckstein, A.: Variational approach for joint optic-flow computation and video restoration. Computer Science, Technion (2005)
37. Park, B., Lee, K., Lee, S.: A new similarity measure for random signatures: perceptually modified hausdorff distance. In: ACIVS, pp. 990–1001 (2006)
38. Romeny, B.M.T.H., Florack, L.: A multiscale geometric model of human vision. In: The Perception of Visual Information, pp. 73–114 (1993)
39. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV **47**(1), 7–42 (2002)
40. Scharstein, D., Szeliski, R.: Middlebury stereo dataset. http://vision.middlebury.edu/stereo/data/ (2007)
41. Shao, L., Yan, R., Li, X., Liu, Y.: From heuristic optimization to dictionary learning: A review and comprehensive comparison of image denoising algorithms. IEEE Trans. on Cybernetics **44**(7), 1001–1013 (2014)
42. Shen, X., Yan, Q., Xu, L., Ma, L., Jia, J.: Multispectral joint image restoration via optimizing a scale map. IEEE TPAMI **37**(12), 2518–2530 (2015)
43. Tan, X., Sun, C., Wang, D., Guo, Y., Pham, T.: Soft cost aggregation with multi-resolution fusion. In: ECCV, pp. 17–32 (2014)
44. Vemulapalli, R., Tuzel, O., Liu, M.: Deep gaussian conditional random field network: A model-based deep network for discriminative denoising. arXiv **1511.04067** (2015)
45. Vu, D., Chidester, B., Yang, H., Do, M., Lu, J.: Efficient hybrid tree-based stereo matching with applications to postcapture image refocusing. IEEE TIP **23**(8), 3428–3442 (2014)
46. Xu, J., Yang, Q., Tang, J., Feng, Z.: Linear time illumination invariant stereo matching. IJCV pp. 1–15 (2016)
47. Xu, J., Zhang, L., Zuo, W., Zhang, D., Feng, X.: Patch group based nonlocal self-similarity prior learning for image denoising. In: ICCV, pp. 244–252 (2015)
48. Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. In: ECCV, pp. 157–170 (2010)
49. Yue, H., Sun, X., Yang, J., Wu, F.: Cid: Combined image denoising in spatial and frequency domains using web images. In: CVPR, pp. 2933–2940 (2014)
50. Yue, H., Sun, X., Yang, J., Wu, F.: Image denoising by exploring external and internal correlations. IEEE TIP **24**(6), 1967–1982 (2015)
51. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: ECCV, pp. 151–158 (1994)
52. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research **17**(1), 2287–2318 (2016)
53. Zhang, K., Fang, Y., Min, D., Sun, L., Yang, S., Yan, S., Tian, Q.: Cross-scale cost aggregation for stereo matching. In: CVPR, pp. 1590–1597 (2014)
54. Zhang, L., Dong, W., Zhang, D., Shi, G.: Two-stage image denoising by principal component analysis with local pixel grouping. Pattern Recognition **43**(4), 1531–1549 (2010)
55. Zhang, L., Vaddadi, S., Jin, H., Nayar, S.: Multiple view image denoising. In: CVPR, pp. 1542–1549 (2009)
56. Zontak, M., Mosseri, I., Irani, M.: Separating signal from noise using patch recurrence across scales. In: CVPR, pp. 1195–1202 (2013)
57. Zoran, D., Weiss, Y.: From learning models of natural image patches to whole image restoration. In: ICCV, pp. 479–486 (2011)