16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

CNN-based Image Denoising for Outdoor Active Stereo

Chengchao Qu Maksim Moiseikin* Sascha Voth Jürgen Beyerer
Fraunhofer IOSB
Fraunhoferstr. 1, 76131 Karlsruhe, Germany
chengchao.qu@iosb.fraunhofer.de

Abstract

Stereo vision has been the most widely used passive 3D sensing technology for a variety of vision tasks. 3D coordinates are computed by triangulating correspondences found in the stereo image pair. For homogeneous areas where stereo matching fails, a stereo projector system can be employed by actively projecting auxiliary texture onto the scene. However, the applicability of this approach is restricted to indoor scenarios, since in outdoor environment where the sunlight is strong, the projected pattern is almost invisible. A simple increase in contrast of the projection leads to dramatic rise of the noise level, which again has an adverse impact on the matching algorithm.

We propose a novel framework to tackle this problem, exploiting adaptive contrast improvement with denoising techniques using convolutional neural networks (CNNs) on the difference images to digitally enhance the projection, which is later added back onto the image pair to assist stereo matching. In order to learn an optimal denoising network dedicated to the projected pattern, a straightforward workflow is devised to allow for convenient acquisition of noisy and noiseless pattern images for the input and ground truth respectively. Extensive evaluation on real-world data compared to the state of the art justifies the effectiveness of not only the presented denoising CNN architecture and training routine, but also the entire pipeline for outdoor active stereo reconstruction.

1 Introduction

Despite the rapid development in commercial depth sensors leveraging diverse 3D sensing technologies, such as light field for Raytrix cameras [17], structured light (SL) for Kinect v1 and time of flight (ToF) for Kinect v2 [20], stereo systems are still one of the go-to solutions in many research and industrial applications by virtue of the compactness, computational efficiency, high resolution, as well as low cost and power consumption. However, passive stereo usually has difficulty in matching low-texture surfaces, producing invalid depth maps for those regions. A straightforward solution is to "paint" the scene with projected auxiliary texture [13], which provides unique pattern along the horizontal direction. This helps to block-matching algorithms for finding matches in the stereo image pair for computing range by triangulation. This approach has successfully found its application in the newly released Intel RealSense D400 series [10].

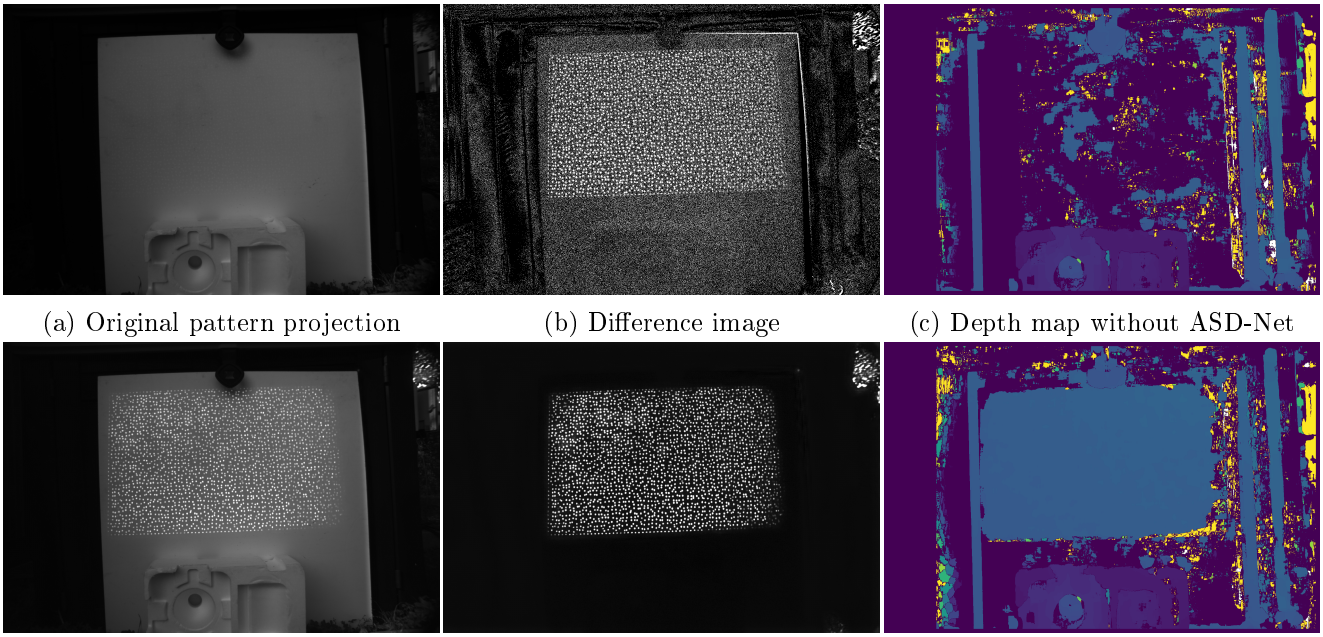
Recent hardware advancement has remarkably improved not only the resolution, frame rate and signal-to-noise ratio (SNR), but also the dynamic range in modern camera sensors. Nonetheless, strong external light sources such as sunlight still pose a huge challenge to active illumination–camera systems. *E.g.*, Kinect, as one of the most successful legacy depth sensors, is only applicable to indoor environment. Even the latest Intel RealSense D400 series need to rely on the original grayscale image texture outdoors, as the IR pattern projection is too weak to be seen when directly exposed in sunlight. Figure 1a illustrates this phenomenon, while the random points projected on the board are barely visible, so that stereo reconstruction in Figure 1c completely fails in this region.

On the other hand, convolutional neural networks (CNNs) have proven to be well-suited for solving a plethora of image restoration tasks [23, 26]. In this paper, we propose a novel framework to address the "disappearing" active projection problem by exploiting proper CNN architectures for enhancing the contrast of the pattern, while, at the same time, suppressing the extreme image noise owing to the short exposure time while acquired outdoors (see Figure 1b). In order to learn an optimal denoising model tailored to the projected pattern, a workflow is developed to allow for convenient acquisition of noisy and noiseless training data. Extensive experiments demonstrate that our network trained with these insights outperforms state-of-the-art algorithms (see Figure 1e). Moreover, we devise a synthetic back-projection scheme to overlay the clean patterns onto the original stereo images (see Figure 1d). In this way, not only the auxiliary projected texture, but also the raw image information can be utilized for stereo reconstruction (see the board and the foam in Figure 1f).

Our main contributions are summarized as follows:

- A straightforward workflow facilitating simultaneous acquisition of realistic noisy and noiseless training data is devised to boost the performance of CNN pattern denoising.

*This work was done when Maksim Moiseikin was an intern at Fraunhofer IOSB.



(a) Original pattern projection (b) Difference image (c) Depth map without ASD-Net

(d) Overlay of (e) on (a) (e) Denoised result using ASD-Net (f) Depth map with (d)

Figure 1: Overview of the proposed framework with example results on real-world image data captured outdoors using the prototype in Figure 4. Best viewed by zooming in the electronic version.

- Building on the U-Net architecture [18] and the residual learning idea [9], we propose a state-of-the-art solution, coined Active Stereo Denoising Network (ASD-Net), for solving the outdoor active stereo problem.
- We present a novel strategy with synthetic overlay of the denoised pattern projection onto the input image to combine the merits of both active and passive stereo vision.
- To the best of our knowledge, the entire pipeline is the first successful attempt on outdoor active stereo using CNN-based denoising.

2 Related work

Active stereo As being one of the classic and most studied methods for 3D sensing, stereo vision has been showing its edge against other popular alternatives such as ToF, SL, LiDAR and light field over the years. Current passive stereo algorithms offer very good precision and efficiency [8], but suffer from dropouts at textureless areas where stereo matching fails, which can be alleviated by supplementing the system with a texture projector. This, in comparison with the passive mechanism, is known as *active stereo* [13]. Unlike SL approaches with a single camera, active stereo does not presume a known geometry of the light pattern, thus possible with unstructured light [16]. A myriad of studies have been presented to investigate the optimal pattern [15], better reconstruction quality [22], *etc.* Recently, Intel released the RealSense D400 camera series with an IR projector module and two cameras for IR as well as visible spectra [10], which can deliver high-quality depth maps of up to 1280×720 pixels in real time thanks to the highly efficient onboard semi-global matching implementation [11], but still suffer from the "disappearing" pattern projection problem in outdoor environment.

To rise to the challenge of noise removal from external light, some authors adopt spacetime stereo (STS) by using many frames with different projection patterns [5, 28]. Sagawa and Satoh [19] extract the light signals of the illuminant by removing ambient light with spread spectrum modulation for better SNR, which is also suitable for moving objects, however, with the minimum requirement of 40 mW laser power.

Image denoising As a long-time research topic in low-level vision, image denoising has received substantial attention in the image processing community, which results in diverse methods for image prior modeling, including Markov Random Fields (MRFs) [14], self-similarity-based non-local means (NLM) [1] and block-matching and 3D filtering (BM3D) [4], sparse representation [6], weighted nuclear norm minimization (WNNM) [7], *etc.* Lately, the breakthrough in deep learning and especially CNNs has revolutionized image denoising. In [2], Burger *et al.* leverage multi-layer

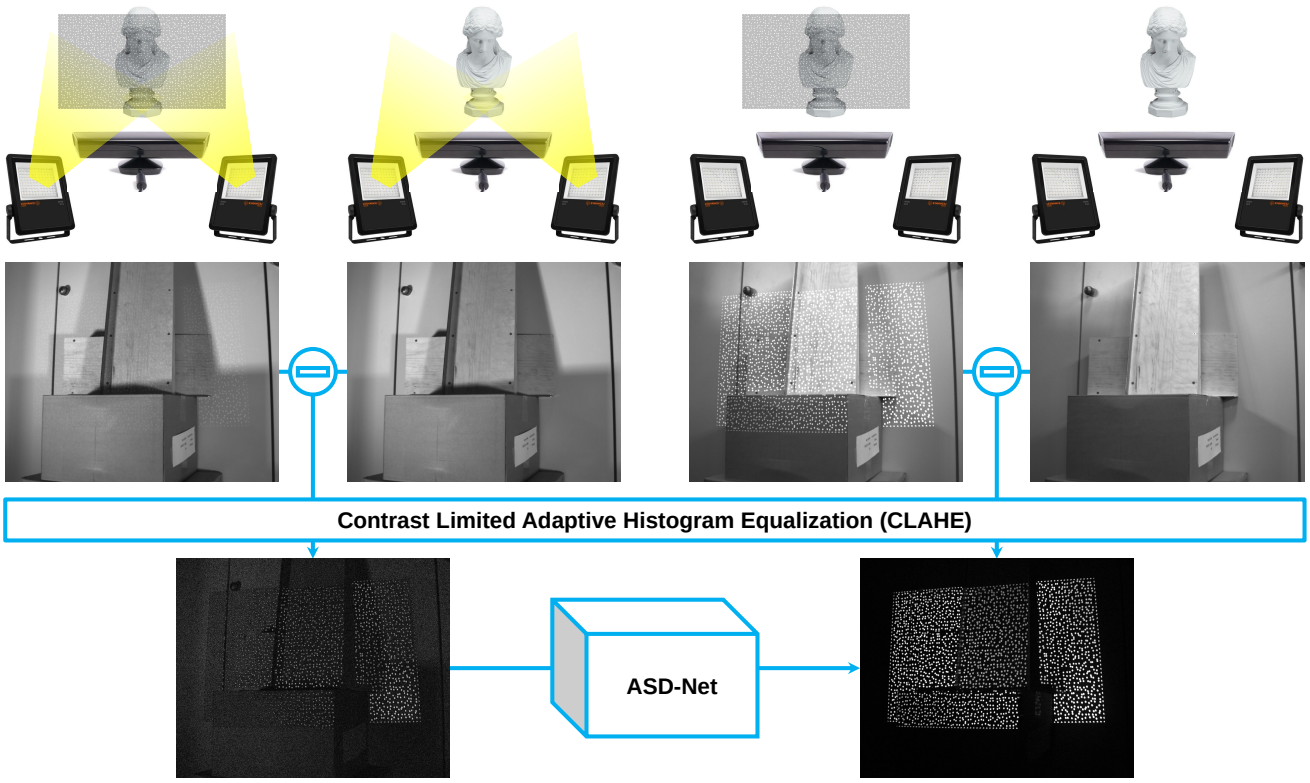


Figure 2: Pipeline for training our ASD-Net using realistic images collected with a special hardware setup.

perceptron (MLP) to achieve fast inference and comparable performance as traditional BM3D. Zhang *et al.* propose DnCNN [25] with a deep sequential architecture and residual learning [9], which recovers the noise map for additive white Gaussian noise (AWGN). Later, they present FFDNet [27] to first estimate pixel-wise noise levels in the image before the denoising network. A summary of recent work can be found in [21].

Most discriminative denoising algorithms are learned on synthetic data due to the simplicity of generating large amount of ground truth training data based on clean images with AWGN or salt-and-pepper noise. However, it is proved that realistic images provide unique characteristics, which is beneficial to the actual task. *E.g.*, the See-in-the-Dark (SID) dataset [3] captured with raw sensor data under extreme low-light circumstances can help to train a deep CNN suitable for noise removal with high color fidelity. This inspires us to acquire our own dataset for the specific random point pattern in active stereo.

3 Pipeline for outdoor active stereo

The first goal of this work is to recover a clean pattern image $\mathbf{x}$ from a noisy input $\mathbf{y}$, which is obtained by subtracting a captured image with active projection from that without it, under the assumption of an additive noise $\mathbf{y} = \mathbf{x} + \mathbf{n}$. Unlike most related work with AWGN, the noise term $\mathbf{n}$ follows an unknown prior, which should be learned in a data-driven manner.

3.1 Acquisition of training data

A large number of realistic images is essential for training CNNs. In the case of active stereo, we need difference image pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$ of the same scene, while the following challenge must be addressed:

- For capturing $\mathbf{x}_i$ where the clean projection pattern is well visible, it should be acquired *indoors with long exposure time.*
- For capturing $\mathbf{y}_i$ where the projection pattern is contaminated with strong image noise, it should be acquired *outdoors with short exposure time.*

So question arises as to where to collect the training data to meet the requirement for variable length of exposure. The strategy employed in [3] is not applicable, as their objective is to obtain the ground truth clean images in low-light environment.

We present a special hardware setup to solve this issue. Two LED floodlights with 20 000 lm are deployed to serve as artificial sunlight sources, which facilitates convenient indoor acquisition. The concrete workflow is depicted in the upper part of Figure 2. Obviously,

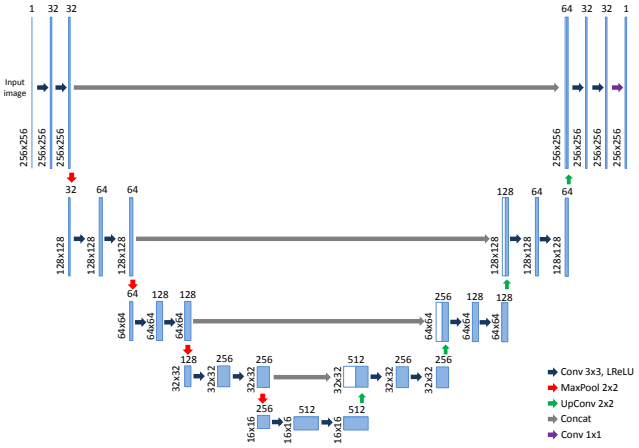


Figure 3: Network architecture of our ASD-Net, which is a modified version of U-Net [18].

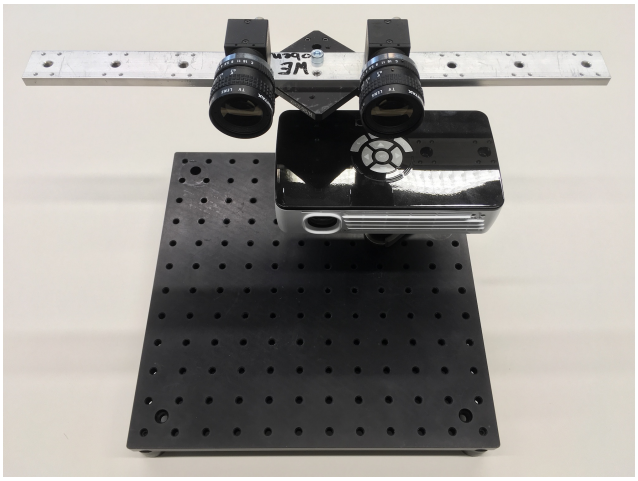


Figure 4: Prototype of our active stereo system.

by turning on the lights, the exposure time can be suppressed to a very short duration, so that the projected pattern is barely visible on the right part of the first image in the middle row, similar to the outdoor scenario in Figure 1a. Afterwards, contrast-limited adaptive histogram equalization (CLAHE) [30] is utilized to enhance the contrast of the point pattern in both $\mathbf{x}_i$ and $\mathbf{y}_i$. Note that although this step is optional by virtue of the capabilities of CNNs, we find it offers small performance boost as well as better visualization. Totally 100 image pairs with different objects in the scene are acquired, which are partitioned into 75, 15 and 10 for training, validation and evaluation respectively.

3.2 Active Stereo Denoising Network (ASD-Net)

After collecting the appropriate training data, our ASD-Net is now described in detail. In the state-of-the-art CNN denoiser DnCNN [25], residual learning [9] is proved not only to allow for greater network depth, but also to ease the learning of the denoising task. In this sense, we adopt this principle to estimate the noise map $\mathbf{n}$ from the input $\mathbf{y}$. As such, this is equivalent to adding an identity shortcut into the network.

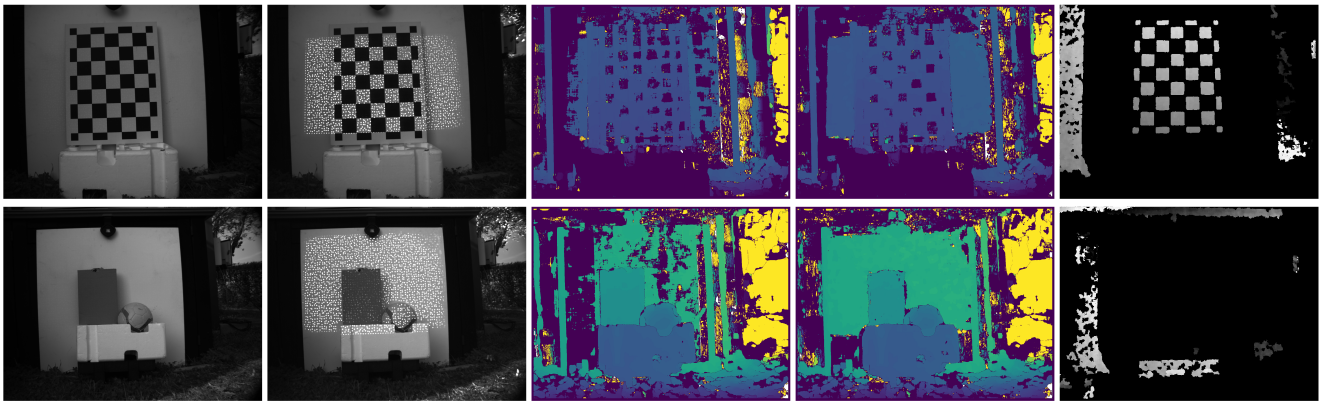
Unlike the sequential single-scale approach in DnCNN, a multi-scale network, *i.e.*, U-Net [18], is investigated in our work, since it is proved to be effective for many pixel-level tasks. Furthermore, its architecture similar to an auto-encoder has the potential to effectively encode the spatial distribution of the specific random point pattern in our active stereo setup.

Figure 3 illustrates our ASD-Net. In this symmetric architecture, each convolution block is composed of two 3×3 convolution layers, which is followed by a max pooling or deconvolution layer to shrink or enlarge the resolution respectively. The deeper it goes, the more feature maps are used to extract higher-level information without adding too much computational burden thanks to the smaller resolutions. The outputs of the left blocks are concatenated to the right part to promote and combine lower-level information with higher-level abstraction. An important difference of ASD-Net to the original U-Net lies in the convolution layers. Since no padding is implemented in U-Net, the right part of the network has a lower resolution than the counterparts on the left, which is not desired in our task. Instead, we make zero-padding to ensure same resolution of input and output images. Moreover, the number of feature maps in each convolution layer is halved to accelerate the inference while still able to outperform DnCNN.

Implementation details Our ASD-Net is implemented with the TensorFlow framework. During the training, each image from Section 3.1 is divided into 820 patches of 256×256 pixels, leading to 61 500 training samples in total. To further enrich the training corpus, common data augmentation techniques are employed, including random rotation, mirroring and gray-level jittering. Besides, we impose extra AWGN to degrade the input images, as the two floodlights are still not as strong as the sun. The Adam optimizer [12] is utilized with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and the initial learning rate of 10^{-4} with a decay rate of 0.1 applied halfway through the totally 1000 epochs. We adopt the ℓ_2 loss for training, although other prevailing losses are also easily applicable [29].

3.3 Denoised pattern overlay for stereo matching

After the difference images are processed by ASD-Net, an intuitive way for the final step is to directly run stereo algorithms on the outputs for the scene reconstruction. Nevertheless, these pure pattern images (see Figure 1e) do not convey any more information than those with SL solutions. In order to maximize the ben-



(a) (b) (c) (d) (e)

Figure 5: Qualitative results on real-world image data captured outdoors: (a) original pattern projection; (b) overlay of denoised pattern using ASD-Net; (c) depth map with (a); (d) depth map with (b); (e) depth map using Kinect v1. Best viewed by zooming in the electronic version.

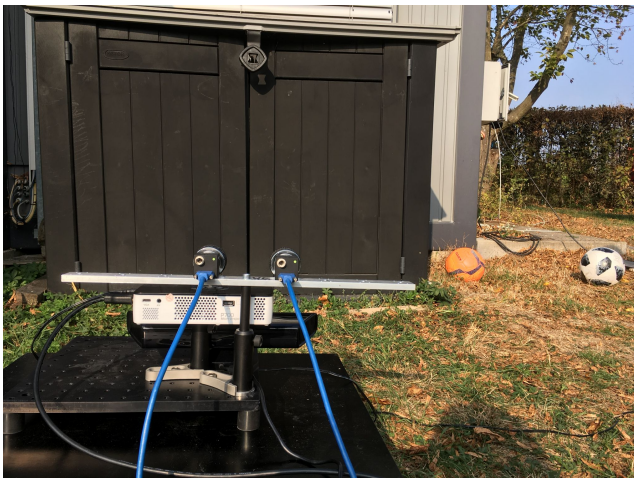


Figure 6: Outdoor data acquisition for evaluation with the prototype system in Figure 4 and a Kinect v1.

efit of the stereo system, we devise a posterior back-projection scheme to overlay the clean patterns onto the original stereo images (see Figure 1d). As such, the supplemental image texture can add extra robustness and precision to the active stereo framework. As an example, the foam in Figure 1f outside the projection area is now reconstructible. The merits of both active and passive stereo vision are thereby fully exploited.

4 Experiments

In this section, we validate the effectiveness of the proposed solution for outdoor active stereo. Standard metrics peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [24] are used to quantitatively benchmark the denoising performance of ASD-Net compared to state-of-the-art approaches NLM [1], BM3D [4] and DnCNN [25] on the test images of the collected indoor data with ground truth (see Section 3.1), which are reported in Table 1. It can first be seen that the original DnCNN model for natural images from the authors [25] slightly outperforms the classic non-deep NLM and BM3D. On the other side, the baseline DnCNN model trained on our active projection data gives a huge boost, justifying the necessity of realistic data for these domain-specific images. The investigated ASD-Net, as expected, provides the highest quantitative scores, especially in the perception-based SSIM, which is attributed to the multi-scale architecture for better extraction of semantic information. Moreover, for input images of 1920×1200 pixels, the runtime measured on an NVIDIA GeForce GTX 1080 Ti GPU is significantly reduced from 0.55 s with DnCNN down to 0.19 s with our ASD-Net.

Table 1: Quantitative denoising results in PSNR and SSIM on the indoor test data.

Approach	PSNR	SSIM
NLM [1]	15.50	0.37
BM3D [4]	15.54	0.39
DnCNN [25] (original model)	15.68	0.46
DnCNN [25] (own model)	19.96	0.68
ASD-Net	19.98	0.83

A second dataset is acquired on a sunny day to evaluate our outdoor active stereo pipeline (see Figure 6). Sample results in Figure 5 show that the original projection in Figure 5a offers very limited help for stereo matching. In comparison, the presented synthetic overlay of the denoised pattern in Figure 5b demonstrates promising improvement for the depth maps. On a final

note for completeness, Kinect v1, which is designed for indoor usage, works only at dark areas where its pattern is still visible.

5 Conclusions

This paper proposes the first effort on outdoor active stereo using CNN-based denoising. A straightforward workflow facilitating convenient acquisition of realistic ground truth data is devised, which, coupled with our dedicated ASD-Net, leads to state-of-the-art denoising performance on active stereo images. Finally yet importantly, a novel strategy with synthetic overlay of the denoised pattern onto the input image successfully combine the merits of both active and passive stereo vision. Future work on our ASD-Net will focus on simultaneous exploitation of stereo images for better consistency and direct depth inference.

Acknowledgment

This work was supported by the FhG Internal Programs under Grant No. WISA 831395.

References

[1] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *CVPR*, 2005.

[2] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" In *CVPR*, 2012.

[3] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *CVPR*, 2018.

[4] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE TIP*, 2007.

[5] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," *IEEE TPAMI*, 2005.

[6] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE TIP*, 2013.

[7] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *CVPR*, 2014.

[8] R. A. Hamzah and H. Ibrahim, "Literature survey on stereo vision disparity map algorithms," *Journal of Sensors*, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[10] Intel Corporation. (2018). Intel RealSense depth camera D435, [Online]. Available: https://click.intel.com/intelr-realsensetm-depth-camera-d435.html.

[11] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel(R) RealSense(TM) stereoscopic depth cameras," in *CVPRW*, 2017.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[13] K. Konolige, "Projected texture stereo," in *ICRA*, 2010.

[14] X. Lan, S. Roth, D. Huttenlocher, and M. J. Black, "Efficient belief propagation with learned higher-order Markov random fields," in *ECCV*, 2006.

[15] J. Lim, "Optimized projection pattern supplementing stereo systems," in *ICRA*, 2009.

[16] H. K. Nishihara, "PRISM: A practical real-time imaging stereo matcher," MIT, Tech. Rep., 1984.

[17] Raytrix GmbH. (2018). 3D light field camera solutions, [Online]. Available: https://raytrix.de/products/.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[19] R. Sagawa and Y. Satoh, "Illuminant-camera communication to observe moving objects under strong external light by spread spectrum modulation," in *CVPR*, 2017.

[20] H. Sarbolandi, D. Lefloch, and A. Kolb, "Kinect range sensing: Structured-light versus time-of-flight Kinect," *CVIU*, 2015.

[21] C. Tian, Y. Xu, L. Fei, and K. Yan. (2018). Deep learning for image denoising: A survey. arXiv: 1810.05052 [cs.CV].

[22] F. Tombari and K. Konolige, "A practical stereo system based on regularization and texture projection," in *ICINCO*, 2009.

[23] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *CVPR*, 2018.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE TIP*, 2004.

[25] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE TIP*, 2017.

[26] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep CNN denoiser prior for image restoration," in *CVPR*, 2017.

[27] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE TIP*, 2018.

[28] L. Zhang, B. Curless, and S. M. Seitz, "Spacetime stereo: Shape recovery for dynamic scenes," in *CVPR*, 2003.

[29] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE TIP*, 2017.

[30] K. Zuiderveld, "Contrast limited adaptive histogram equalization," *Graphics Gems IV*, 1994.