Lighting Enhancement Aids Reconstruction of Colonoscopic Surfaces

Yubo Zhang[1], Shuxian Wang[1], Ruibin Ma[1], Sarah K. McGill[2], Julian Rosenman[3], and Stephen Pizer[1]

[1] Department of Computer Science, University of North Carolina at Chapel Hill
`{zhangyb,shuxian,ruibinma,pizer}@cs.unc.edu`
[2] Department of Medicine, University of North Carolina at Chapel Hill
`mcgills@email.unc.edu`
[3] Department of Radiation Oncology, University of North Carolina at Chapel Hill
`rosenmju@med.unc.edu`

Abstract. High screening coverage during colonoscopy is crucial to effectively prevent colon cancer. Previous work has allowed the alerting the doctor to unsurveyed regions by reconstructing the 3D colonoscopic surface from colonoscopy videos in real-time. However, the lighting inconsistency of colonoscopy videos can cause a key component of the colonoscopic reconstruction system, the SLAM optimization, to fail. In this work we focus on the lighting problem in colonoscopy videos. To successfully improve the lighting consistency of colonoscopy videos, we have found necessary a lighting correction that adapts to the intensity distribution of recent video frames. To achieve this in real-time, we have designed and trained an RNN network. This network adapts the gamma value in a gamma-correction process. Applied in the colonoscopic surface reconstruction system, our light-weight model significantly boosts the reconstruction success rate, making a larger proportion of colonoscopy video segments reconstructable and improving the reconstruction quality of the already reconstructed segments.

Keywords: Colonoscopy, image enhancement, 3D reconstruction

1 Introduction

Colonoscopy is an effective examination to prevent colon (large intestine) cancer by screening for lesions. During a colonoscopy a flexible tube called a colonoscope is inserted up to the distal end of the patient's colon, and then it is withdrawn through the colon while producing a video by a camera that is attached at the tip of the colonoscope. A point light attached to the colonoscope tip moves with the camera to provide the lighting source. During a colonoscopy the camera sends back live images of the colon wall to the doctor, who watches and detects and removes cancerous or pre-cancerous lesions with a surgical knife inside the colonoscope.

Due to time limitation, colonoscopy is a one-pass exam, but high coverage of screening and high polyp detection rate are crucial. Previous work tried to
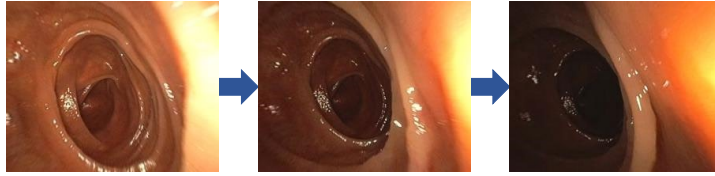
Fig. 1. The lighting in the colonoscopy video can change rapidly from frame to frame, and a large part of the colon interior can be poorly lit due to the complicated colon terrain. This lighting issue can bring troubles to the diagnosis and the colon surface reconstruction system.

aid the colonoscopy process by alerting the doctor to unsurveyed colon regions ("blind spots") revealed by reconstructing the already screened surfaces in real-time from colonoscopy videos [11]. However, as shown in Fig. 1, the complex topography makes a notable proportion of the colon surface be poorly lit; also, it causes brightness and contrast to change rapidly from frame to frame. The poor lighting situation challenges 1) lesion detection by the doctor and 2) successful reconstruction of the surfaces, thereby allowing alerts as to blind spots.

In this work we focus on this lighting issue in colonoscopy. We aim to brighten the dark regions in colonoscopy video frames and to make the lighting of colonoscopy videos consistent between consecutive frames. Such image enhancement is commonly carried out by curve adjustment approaches [7, 6] such as gamma correction, and recent deep learning methods can apply more sophisticated adjustment to each image pixel with CNN- [5, 19] or GAN-based models [1, 8]. Considering that our inputs are video frames, we adopt the learning-based approach and develop an RNN-based network for adaptive curve adjustment. Our network is trained in an unsupervised fashion using a loss measuring the lighting consistency among nearby frames and will produce well-lit frame sequences at test time. Applying our model as a pre-processing step to the frames for reconstruction leads to significant improvement of both the number of reconstructable colonoscopy video segments and the quality of reconstructions, with time overhead small enough to be ignored.

Overall, the contributions of this work are

- To our knowledge, our work is the first one to focus on the lighting issue in the colonoscopy video.
- We propose a light-weight RNN model for lighting enhancement of an image sequence, which can be trained without ground-truth. Tested on real colonoscopy images, our model can effectively brighten the poorly lit regions in each frame, and make the lighting consistent from frame to frame.
- Most importantly, applied in the colonoscopic surface reconstruction system, our model effectively boosts the reconstruction success rate and improves the reconstruction quality without sacrificing the time efficiency of the system.

2 Background: SLAM and colonoscopic reconstruction

Here, we review the mechanism behind the 3D reconstruction technique SLAM, which is a key component in colonoscopic reconstruction systems. We then analyze why the lighting issue in colonoscopy can cause the SLAM system to fail.

2.1 SLAM mechanism

3D reconstruction is a challenging task in computer vision. One of the most successful methods for this task is Simultaneous Localization And Mapping (SLAM). SLAM systems have achieved tremendous success in indoor and outdoor-scene reconstruction [4, 17]. Recently, with the development of deep learning, neural networks have been applied to aid SLAM for better reconstruction [18, 20, 24], especially in the more challenging scenarios such as reconstruction from colonoscopy video [11].

SLAM is an algorithm that can achieve real-time dense reconstruction from a sequence of monocular images [13, 3, 12, 2]. As the name suggests, SLAM has a localization component and a mapping component; the two components operate cooperatively. The localization (tracking) component predicts the camera poses from each incoming image frame. Based on the visual clues extracted from the images, the mapping component optimizes especially the pose predictions but also the keypoints' depth estimates. The objective function used in the optimization of the SLAM of [2], which is applied in colonoscopic reconstruction [11], is

$$E_{ij} = \sum_{\boldsymbol{p} \in \mathcal{P}_j} \omega_{\boldsymbol{p}} \left\| (I_j[\boldsymbol{p}'] - b_j) - \frac{e^{a_j}}{e^{a_i}} (I_i[\boldsymbol{p}] - b_i) \right\|_{\gamma} \tag{1}$$

$$\boldsymbol{p}' = \Pi(T\Pi^{-1}(\boldsymbol{p}, d_{\boldsymbol{p}})) \tag{2}$$

Each keypoint $\boldsymbol{p}$ in a resource image I_i can be projected to a location $\boldsymbol{p}'$ in a target image I_j, based on its predicted depth $d_{\boldsymbol{p}}$ and the predicted camera transformation T. Π denotes the projection function. Supposing the transformation and depth predictions are correct, $\boldsymbol{p}$ of image I_i and $\boldsymbol{p}'$ of image I_j come from the same surface point, so $I_i[\boldsymbol{p}]$ and $I_i[\boldsymbol{p}]$ should have the same intensity. This assumption is the mechanism behind the energy function E_{ij} which is the photometric error between two frames I_i and I_j.

2.2 The lighting problem in colonoscopic surface reconstruction

The photometric-base optimization in SLAM depends on the light consistency between frames, that assuming the same physical location in the environment is shown in similar intensities in different frames. This is a prerequisite that can be mostly fulfilled in an indoor or outdoor scenario with the steady lighting source, e.g., sunlight. When it comes to colonoscopy, this lighting consistency assumption can often be violated since the point light is moving with the camera and can change rapidly due to motion and occlusion. Although the SLAM

algorithm also optimizes an additional brightness transformation $e^{-a_i}(I_i - b_i)$ as a compensation, it is only a global adjustment for each frame to fix exposure and cannot handle the more complicated lighting changes in colonoscopy videos such as contrast difference (bright regions become brighter and dark regions become darker). So without specific brightness adjustment, this kind of rapid lighting change in colonoscopy makes the SLAM system unstable, often leading to tracking failure.

We want to alleviate the failures caused by lighting changes. In this work we accomplished this by explicitly adjusting the brightness of the image sequence to make them more consistent. This requires solving an adaptive image enhancement problem that each image in the sequence be enhanced in relation to adjacent frames. We develop a deep learning method for this task, which will be discussed in detail in the next section.

3   Method

In this work we apply an adaptive intensity mapping to enhance the colonoscopy frame sequence with the help from an RNN network, whose implementation and unsupervised training strategy will be introduced in this section.

3.1   Adaptive gamma correction

Image enhancement is a classical topic in computer vision. Multiple directions have been proposed to resolve the issue, such as histogram equalization and its variants [14, 22], unsharp masking [15, 10] and more recently deep learning pixel-wise prediction approaches [5, 8, 9].


(a) The curve of gamma correction

(b) An image corrected by different gamma values
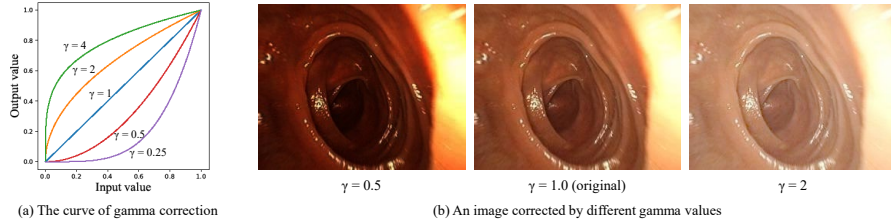
Fig. 2. The gamma correction: (a) Gamma correction's effect on image intensity; (b) The figure in the middle is been adjusted by different gamma values.

Our driving problem of the reconstruction from colonoscopy video requires real-time execution, so our lighting enhancement must be accomplished at the video frame rate. This requires that only a handful of parameters be adjusted from frame to frame in an enhancement method. An example approach that

meets this requirement is gamma correction [7, 6]. In gamma correction, for every pixel in an image, its input value I_{in} is adjusted by the power γ, then multiplied by a constant A to get the enhanced value I_{out}:

$$I_{out} = AI_{in}^{\gamma} \tag{3}$$

In practice, the input intensities will be normalized to $[0, 1]$, so the constant A will be set to 1 and γ is the only parameter controlling the adjustment. By applying a different γ value, the same image can be brightened or darkened to a different extent, as the example shown in (b) of Fig. 2.
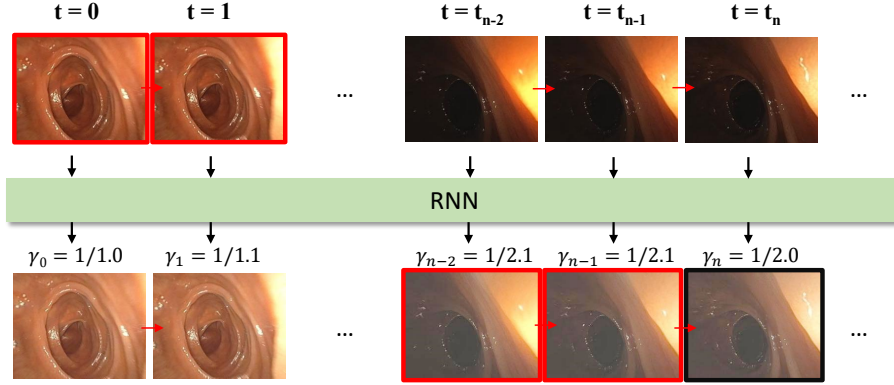
Fig. 3. At each time step, the RNN takes an input image and predicts its γ value to enhance the image. The network is trained by contrasting the current corrected image with two previous corrected images and the first two input images of the sequence.

Our goal is to improve the lighting consistency of an image sequence. Adjusting the way in which each frame is enhanced requires an adaptive version. To model this adaptation, we specifically design a recurrent neural (RNN) network to capture the temporal information needed in the enhancement and to predict the γ value for each image. The overall pipeline of our RNN-based adaptive gamma correction is shown in Fig. 3. The details of the RNN network will be discussed next.

3.2 RNN network

Although the gamma correction is cheap in computation, our enhancement RNN network also needs to be light-weight in order to serve as an additional preprocessing step for the current real-time colonoscopic surface reconstruction system. Thus, we build our RNN network on the ConvLSTM unit [23], with only

two additional convolution layers:

$$x_1 = \mathrm{ReLU}(\mathrm{Conv}(I_i)) \qquad (4)$$

$$[x_2, h_i] = \mathrm{ConvLSTM}([x_1, h_{i-1}]) \qquad (5)$$

$$x_3 = \mathrm{ReLU}(\mathrm{Conv}(x_2)) \qquad (6)$$

$$f = \mathrm{AvgPool}(x_3) \qquad (7)$$

$$\gamma_i = \mathrm{ReLU}(Af + b) \qquad (8)$$

At test time, at each time step the network takes the current image frame I_i as the input, includes latent information of the previous frames h_{i-1} into the computation, and predicts the gamma correction value γ_i for the current frame. The prediction happens recurrently, enhancing the entire sequence as shown in Fig. 3.

3.3 Training strategy

With no groundtruth to supervise the network training, our RNN is trained in an unsupervised fashion to achieve lighting consistency. We achieve this consistency by comparing the current adjusted frame to four frames that we call reference frames. These reference frames consist of the previous two adjusted frames and the first two images of the entire sequence, which serve as the "seed" images. For example in Fig. 3, the target frame is surrounding in black and all the reference frames are surrounding in red.

The loss we are optimizing utilizes the structural similarity measurement [21]:

$$L_{ssim} = \mathrm{mean}(1 - \mathrm{SSIM}(I_r, I_t)) \qquad (9)$$

$$= \mathrm{mean}(1 - \frac{(2\mu_{I_r}\mu_{I_t} + c_1)(2\sigma_{I_r I_t} + c_2)}{(\mu_{I_r}^2 + \mu_{I_t}^2 + c_1)(\sigma_{I_r}^2 + \sigma_{I_t}^2 + c_2)}) \qquad (10)$$

where at each pixel location, μ_I is the local average, σ_I^2 is the local variance and $\sigma_{I_r I_t}$ is the covariance of two images. The SSIM function is composed of three comparisons: luminance, contrast and structure. When applying to training, the luminance and contrast comparisons will force the network to produce adjusted images with similar lighting to the reference frames.

Specular regions have very high intensities that inordinately influence the lighting estimation in a frame. Therefore, in training when computing the L_{ssim}, we mask out the pixels with input intensity larger than 0.7. Each training sequence contains 10 sequential frames, with the first two as seeds, and the loss of each target frame is computed as the average L_{ssim} of its four reference frames.

4 Experiments

4.1 Implementation details

The training data for the RNN network comes from 75 colonoscopy video snippets, each containing 50 to 150 frames. We divided the snippets into 10-frame

overlapping sub-sequences. In this way created about 1800 training sequences. We used the Adam optimizer with a fixed learning rate of 5×10^{-5} to train the RNN network for 10 epochs. A batch size of 4 was used. We implemented our method using the PyTorch framework on a single Nvidia Quadro RTX5000 GPU.

At test time the average run time of our network to enhance an image is less than a millisecond. Considering the frame rate we use to extract images from video is 18 frames per second, the overhead brought by our method is small enough and does not violate the real-time execution when added to the colonoscopic reconstruction system.

4.2 Visual effect
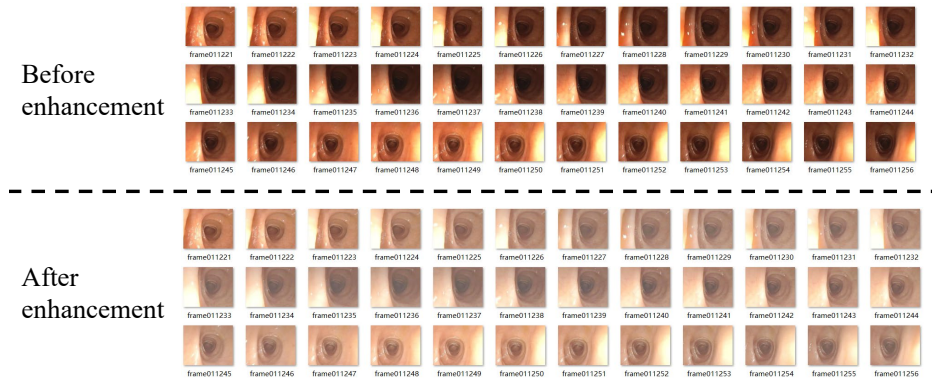
Fig. 4. The visual comparison of the same sequence before (top) and after (bottom) the RNN enhancement. After enhancement, the extreme dark regions have been brightened up and the lighting consistency improves.

In Fig. 4 we show a snippet that is been enhanced using our trained RNN network, where the original frames are shown on the top and the enhanced ones on the bottom. The overall brightness of the sequence has been largely improved and the "down the barrel" regions are brightened up and clearly revealed after enhancement. Moreover, the lighting consistency is significantly improved. The image contrast changes rapidly in the original sequence, due to a haustral ridge blocking much of the light. After enhancement, these changes across frames from bright on average to dark on average and then back to bright on average are far less obvious: the lighting of the sequence becomes much more consistent.

4.3 Application in colonoscopic surface reconstruction

In the colonoscopic surface reconstruction system RNNSLAM [11], a deep learning RNN network runs in parallel with a standard SLAM system. The RNN

component predicts the depths and pose of each camera frame to initialize their SLAM optimization, and the SLAM output updates the hidden states of RNN. This combination leads to low tracking error. Our RNN lighting enhancement network can also run in parallel with the RNNSLAM. The enhanced images produced by our method are given to its SLAM component as the image inputs.

Success on the cases that fail with original images With our method inserted, we tested the RNNSLAM system on the cases which failed with the original images. Most of those failure cases include significant lighting changes or large camera motion that leads to lighting occlusion, and the SLAM tracking fails when it cannot predict reasonable camera poses. We collected 32 cases that have the above failure with original images. When testing them with our enhancement module inserted, 21 of 32 succeeded without tracking failure: the reconstruction success rate significantly increased. We illustrate a case of improved tracking in Fig. 5.
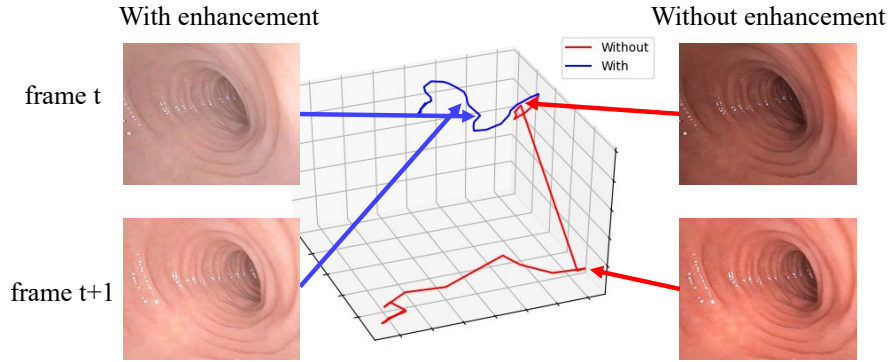
Fig. 5. The comparison of pose trajectories for the same sequence without (red) and with (blue) the RNN enhancement.

In Fig. 5, the predicted camera trajectory without enhancement is shown in red, and the trajectory after enhancement is shown in blue. In each case the camera is predicted to start from the top right position and to move along the respective trajectory to the bottom left position. Particularly in the figure, we show the positions predicted for two successive keyframes t and $t + 1$. Due to a lighting change, the camera pose prediction changes by a dramatic distance when using the original frames, although it is clear to the human eye that the motion between these two images is far more subtle. This issue of discontinuity in predicted poses does not occur when we enhance the input images, where the predicted camera poses move smoothly.

Pose improvement of previously reconstructable sequences Not only does our lighting enhancement allow previously un-reconstructable sequences to be reconstructed, but as demonstrated in the following, it improves the pose trajectories produced from sequences that were previously reconstructable. To quantitatively measure the pose improvement, we adopt the evaluation method in [11], using COLMAP [16] on the un-enhanced images to generate a high-quality baseline trajectory as the virtual "ground-truth" for each video sequence. As the evaluation metric we compute the absolute pose error (APE) of each timestamp in the generated trajectory compared to its "ground-truth". For each trajectory, the RMSE of the APE statistics of all timestamps is computed. We tested 12 colonoscopic sequences that were already reconstructable without our enhancement strategy. We computed the RMSE and other APE statistics (mean, std, etc.) of each sequence respectively, and in Table 1 we list the summary statistics across these 12 sequences. For better reference, we also show the errors of the trajectories produced by COLMAP when using enhanced images. Since we are measuring the error, the lower the result is, the better.

Method	mean RMSE	mean APE	std APE	min APE	median APE	max APE
COLMAP w/ enhanced img	0.329	0.295	0.143	0.077	0.278	0.736
RNNSLAM w/ original img	0.840	0.752	0.371	0.203	0.694	1.666
RNNSLAM w/ enhanced img	**0.680**	**0.606**	**0.302**	**0.105**	**0.554**	**1.388**

**Table 1.** The average APE statistics of 12 colonoscopic sequences

For every sample in these 12 sequences, even though the original lighting is consistent enough for the SLAM to produce decent pose predictions, our enhancement method brought further improvement, decreasing their pose errors. This result shows that besides significantly increasing the reconstruction success rate, our method can also bring minor improvement to the reconstruction quality in cases that can be reconstructed successfully without lighting enhancement.

5   Discussion and conclusion

Due to the complexity of colon geometry, the lighting in colonoscopy videos tends to have rapid changes that lead to failures in the colonoscopic surface reconstruction system. In this work we focused on improving the lighting consistency in colonoscopy videos. We proposed a light-weight RNN network for an adaptive lighting enhancement method to enhance the colonoscopy image sequence, which brightens the dark regions and makes lighting consistent from frame to frame. With the help of our enhancement module, a larger portion of colonoscopy videos can now be successfully reconstructed and the reconstruction quality is improved.

Future work

1. Our method is initially designed to improve the lighting consistency of colonoscopy sequences, but it is also applicable for other video modalities. For example, for other types of endoscopy videos and outdoor driving sequences with large lighting changes, our network can be trained on these data to improve their lighting consistency for better reconstruction.

Before enhancement

After our enhancement
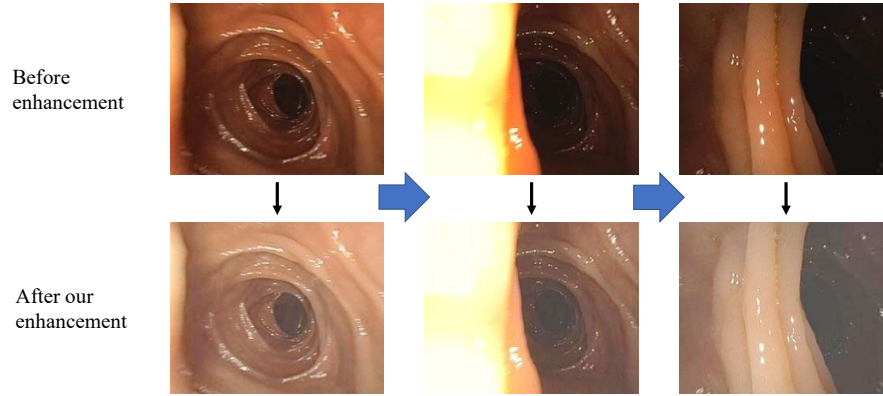
Fig. 6. An extreme tracking condition, the "close occluder". Its complicated lighting situation is beyond our enhancement method's capacity.

2. Although our method successfully increases the reconstruction success rate, there is still a notable fraction of colonoscopy sequences that cannot be reconstructed after the lighting fix. They usually contain some extreme tracking conditions, and one of them is what we call the "close occluder". In this case, as the colonoscopy moving behind a haustral ridge, or moving side-way and being really close to the colon surface, the image foreground occupies a large portion of an image frame, as shown in the last two images of Fig. 6. The lighting in these images is usually extremely bright in the foreground and extremely dark in the background. Moreover, when using these frames, the SLAM system will choose keypoints for optimization not only from the "down the barrel" portion of the image, as it usually does with the normal frames, but also from the bright foreground. Therefore, in order for these cases to succeed, the bright and the dark region both need to be more elaborately adjusted to achieve global lighting consistency. Currently our method cannot handle this complicated scenario and a more sophisticated image enhancement technique is needed; we leave it to future work.

References

1. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6306–6314 (2018)
2. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence **40**(3), 611–625 (2017)
3. Engel, J., Sturm, J., Cremers, D.: Semi-dense visual odometry for a monocular camera. In: Proceedings of the IEEE international conference on computer vision. pp. 1449–1456 (2013)
4. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
5. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
6. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1780–1789 (2020)
7. Huang, S.C., Cheng, F.C., Chiu, Y.S.: Efficient contrast enhancement using adaptive gamma correction with weighting distribution. IEEE transactions on image processing **22**(3), 1032–1041 (2012)
8. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
9. Ke, Z., Qiu, D., Li, K., Yan, Q., Lau, R.W.: Guided collaborative training for pixel-wise semi-supervised learning. In: European Conference on Computer Vision (ECCV) (August 2020)
10. Luft, T., Colditz, C., Deussen, O.: Image enhancement by unsharp masking the depth buffer. ACM Transactions on Graphics (TOG) **25**(3), 1206–1213 (2006)
11. Ma, R., Wang, R., Pizer, S., Rosenman, J., McGill, S.K., Frahm, J.M.: Real-time 3d reconstruction of colonoscopic surfaces for determining missing regions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–582. Springer (2019)
12. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE transactions on robotics **31**(5), 1147–1163 (2015)
13. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: Dtam: Dense tracking and mapping in real-time. In: 2011 international conference on computer vision. pp. 2320–2327. IEEE (2011)
14. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
15. Polesel, A., Ramponi, G., Mathews, V.J.: Image enhancement via adaptive unsharp masking. IEEE transactions on image processing **9**(3), 505–510 (2000)
16. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

17. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: Proc. of the International Conference on Intelligent Robot Systems (IROS) (Oct 2012)
18. Tateno, K., Tombari, F., Laina, I., Navab, N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6243–6252 (2017)
19. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6849–6857 (2019)
20. Wang, S., Clark, R., Wen, H., Trigoni, N.: Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 2043–2050. IEEE (2017)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
22. Xiao, B., Xu, Y., Tang, H., Bi, X., Li, W.: Histogram learning in image contrast enhancement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
23. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802–810 (2015)
24. Yang, N., Wang, R., Stuckler, J., Cremers, D.: Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 817–833 (2018)