Deep learning for robust end-to-end tone mapping

Deep Learning for Robust end-to-end Tone Mapping

Rico Montulet
r.montulet@student.maastrichtuniversity.nl
Alexia Briassouli
alexia.briassouli@maastrichtuniversity.nl

Department of Data Science and Knowledge Engineering
Maastricht University
Maastricht, NL

Abstract

Low-light images require localised processing to enhance details, contrast and lighten dark regions without affecting the appearance of the entire image. A range of tone mapping techniques have been developed to achieve this, with the latest state-of-the-art methods leveraging deep learning. In this work, a new end-to-end tone mapping approach based on Deep Convolutional Adversarial Networks (DCGANs) is introduced along with a data augmentation technique, and shown to improve upon the latest state-of-the-art on benchmarking datasets. We carry out comparisons using the MIT-Adobe FiveK (MIT-5K) and the LOL datasets, as they provide benchmark training and testing data, which is further enriched with data augmentation techniques to increase diversity and robustness. A U-net is used in the generator and a patch-GAN in the discriminator, while a perceptually-relevant loss function based on VGG is used in the generator. The results are visually pleasing, and shown to improve upon the state-of-the-art Deep Retinex, Deep Photo Enhancer and GLADNet on the most widely used benchmark dataset MIT-5K and LOL, without additional computational requirements.

1 Introduction

Low-light conditions can degrade the quality of images, due to the linear mapping carried out by camera sensors, making it difficult to discern details and implement automated computer vision algorithms on them. Since image retouching is a one-to-many problem, a single input image can be retouched in an infinite amount of different ways. This task is best done by hand, for example in Lightroom or Photoshop. Here an expert will change the sliders of white balance, exposure, saturation, etc. until the picture looks optimal (note that optimal is in the eyes of that expert, there is no absolute optimum since this is based on opinion). Tuning all these parameters by hand takes a lot of time, effort and knowledge to be done properly. Various methods have been developed to automatically improve brightness, sharpness and colors in such images, but they are often simple, such as histogram equalization, or semi-automatic, relying on human expertise and interpretation. Histogram equalization and its variants [2] improve contrast by matching an image histogram to a desired distribution, while de-hazing methods [6] can also increase the sharpness of images.

The state-of-the-art in the field is based on deep learning approaches, tailored to the needs of this problem. A series of methods have been developed based on Retinex theory

[5], decomposing images into reflectance and illumination. They can remove noise and increase image visibility, however they are not generally applicable, but need to be tailored for specific images. This has motivated the development of Deep Retinex decomposition [5] for low-light enhancement, which is end-to-end trainable, learning a more generally applicable image decomposition for tone mapping. Here, enhancement is done in three steps: decomposition, adjustment and reconstruction. The first network is trained on decomposing the input image into the reflectance and the illumination. Then a decoder encoder style network enhances the illumination. Finally, a noise removal operation is applied on the reflectance, and the processed reflectance and illumination images are merged to get the enhanced result. GLADNet: Low-Light Enhancement Network with Global Awareness, uses a decoder encoder style network that works on a downsampled version of the image and then applies the estimated "illumination distribution" to the original image to boost it locally. Finally, Deep Photo Enhancer (DPE) [19] proposes an unsupervised deep learning based approach to low-light image enhancement based on two-way Generative Adversarial Networks (GANs) [7] with adaptive weighting [1], for improved convergence. These methods lead to better results than the previous state-of-the-art, however the quality of their outputs has room for improvement, as demonstrated in Sec 4.

This work proposes an approach that is shown to achieve local visually pleasing image enhancement by implementing a variation of DCGANs, which takes into account visually important aspects of imagery. Specifically:

- DCGANs are applied to the lightness channel (the L channel from the LAB colorspace) , instead of the color components of an image, to avoid the color bias (that is observed, for example in DPE). The lightness channel contains the actual structural details of the image and contains the intensity information that needs to be enhanced, while processing all color components does not add information relevant to the enhancement of low-light areas.

- Image-to-image translation is achieved by applying a U-net generator on the lightness channel, with a visually meaningful perceptual loss based on VGG, which leads to improved results.

- The DCGAN discriminator is based on patchGAN, so spatially localised improvements are achieved, overcoming the limitations of state-of-the-art methods that tend to brighten the entire image.

- Data augmentation is applied to the training set to increase diversity and improve robustness.

Experimental results on the benchmark datasets MIT 5K and LOL demonstrate that the approach proposed in this work achieves quantitative and qualitative tone mapping improvements over the existing state-of-the-art. This paper is structured as follows: Sec 2 presents the concept and method proposed in this work. Datasets are described in Sec 3, Sec 4 presents experimental results and Sec 5, Conclusions and plans for future work.

2 DCGANs for end to end tone mapping

The proposed approach will receive an input image A recorded in low-light conditions, and produce an enhanced image A'. The desired appearance of the output image A' is based on

a set of reference images B, modified by experts (see Sec. 3 for details). Thus, the goal of our method is to find the mapping from A to A', based on the ground truth available in the benchmark data.

In the literature, the task of image enhancement has been attempted in two completely different ways: one way is by using reinforcement learning [8], [16], [20] and the other is by using image-to-image translation [19], [9], [10]. Our work relies on deep learning based style transfer and image-to-image translations, which have made great progress using GANs for the production of retouched images, as described in Sec 1. GANs are well-suited for this problem, as they can learn the mapping from the original to the desired output images, as modified by experts in the training data, and then apply it to the testing input data.

2.1 DCGANs Generator

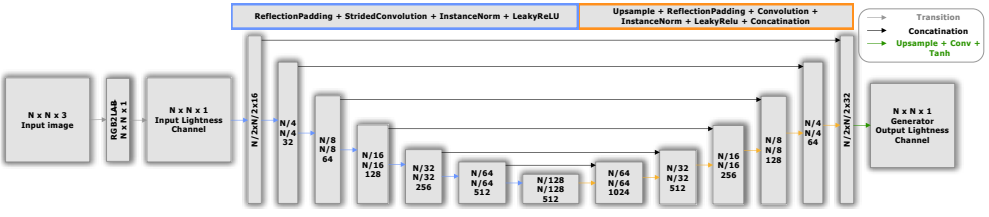

Figure 1: DCGAN U-net Generator network architecture

In this work, we propose an approach that performs end-to-end tone mapping based on a pixel to pixel deep learning architecture. The GAN generator uses U-net [17], as it has been shown to be very successful for image segmentation, and is therefore expected to provide good lightness mapping results. The generator used is shown in Fig 1, where it can be seen that it uses the image lightness channel as input, and provides an enhanced lightness channel as output.

Specifically, given an RGB input image the first step is to extract the lightness channel, resulting in a $N \times N \times 1$ image, which is passed to the first convolutional layer. Each blue arrow in Fig. 1 indicates the following steps: first, reflection padding, followed by a convolutional layer with a stride of two, instance normalisation and finally leakyReLU as the activation function. The block indicates the output size of that layer. These operations are carried out seven times to extract high level features from the image, and will be used to construct the enhanced lightness channel off the image in the next part of the network. This is achieved by applying a different sequence of operations: first we upsample the previous layer, apply reflection padding and a convolutional layer (with stride of one), which is fed into an instance normalisation layer, after which leakyReLU is applied. The intermediate layer is then concatenated with the previous layer (indicated with a black arrow) to create the resulting layer in the gray boxes in Fig. 1. After six blocks, we have an $N/2 \times N/2 \times 32$ layer, which is upsampled one last time, and a convolutional layer with tanh activation gives the final lightness channel.

The DCGAN generator in the proposed architecture employs a loss function that goes beyond simple MSE, as it takes into account visual perception. Specifically, it is based on a combination of perception loss and MSE loss, where the perception loss is computed using VGG. Ledig et al. [12] introduced the idea of using VGG as a loss function in their paper on image super resolution, and it worked very well in practice. The main idea is to take a VGG

network pre-trained on the imageNet dataset for image classification. The network is then truncated, and only the first five layers are kept. The expert image is passed through these five layers, and the outputs of the fifth layer are kept as the higher level features for this image. Next, the generated image is passed through the network and its features are extracted. Now, these feature vectors are compared using MSE to get the difference in features space instead of in pixel space. This preserves high frequencies (details) much better compared to using pixel-wise MSE [12]. It should be noted that in our work, the VGG loss function was adapted to work with grayscale (lightness) images.

2.2 DCGANs Discriminator

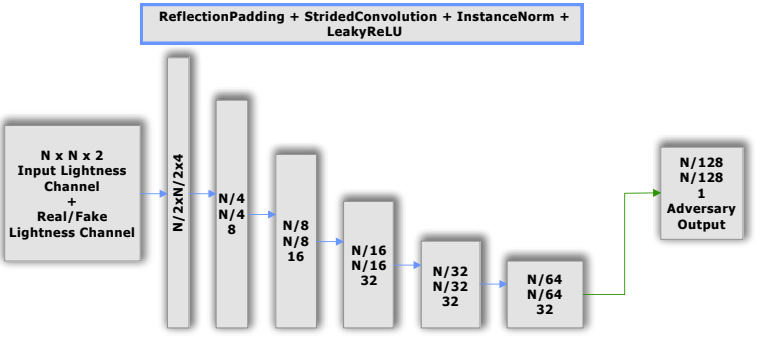

Figure 2: DCGAN Discriminator network architecture

The patchGAN discriminator is used, as it does not only output a single scalar indicating if the image is real or fake, but instead, it maps the input to a $N \times N$ array of outputs X , where each $X_{i,j}$ signifies whether the patch at location i, j in the image is real or fake. Thus, the resulting $N/128 \times N/128$ adversary output in Fig 2 is a map of 1 or -1, indicating if a patch in the input image is real or fake [13].

Unlike the generator, which uses a psycho-visually relevant error metric, the discriminator uses the MSE loss, comparing the values in the $N/128 \times N/128$ output to an array of the same shape, either filled with ones or negative ones, depending on if the image being input into the discriminator is real or fake. Since the discriminator is based on convolution, we can trace these outputs back to the input patch they belonged to, which will in turn tell the generator where it should make the biggest changes to the weights. Thus, using the patchGAN discriminator is equivalent to breaking up the input image into patches, running each of these individually through a regular discriminator and, finally, arranging the outputs. An overview of the discriminator architecture used in this work can be seen in Fig. 2.

3 Benchmark Datasets

In order to leverage the advantages of deep learning and be able to train a neural network, a benchmarking dataset is needed, which provides training and testing data. As a supervised approach is used in this work, the data needs to provide before/after style images, i.e. ground truth information. For this reason, and due to its widespread use in all the state-of-the-art tone-mapping research papers, the MIT-5K dataset created by Adobe and MIT is used [3].

This dataset contains 5000 RAW images taken with high-end DSLR cameras, which were then given to five experts to color correct in Lightroom. A user study was conducted and a group of people ranked the five expert images from most preferred to least preferred color grading. The conclusion was that expert C gave the most desirable results [3] and is therefore used in the rest of this paper, as well as in the state-of-the-art. We pre-processed the MIT-5K images with various data augmentation techniques, in order to create training samples. From every image, 5 random pairs of partially overlapping patches were selected and augmented to create the final dataset. The patches contain 512×512 pixels and chosen based on one condition, which is that they should be non-flat. This means they should contain some edges, which is verified by applying Canny edge detection on the patches [4].

The MIT-5K dataset is split into a train and test set. For testing, about 10% of the dataset was selected by sampling the uniform pseudo random number generator from numpy. This resulted in 442 images that were randomly added to the test set (from MIT-5K only), the other 4558 images are used for generating the training set.

In addition to MIT-5K, another benchmark dataset that has been used in a few of the state-of-the-art papers, is the LOw Light paired dataset (LOL). This dataset was developed in [5] for training and testing their models. In this work, LOL was only used for enriching the testing data, specifically in the qualitative comparison in Sec 4.2.

4 Experiments

Our model is trained on the MIT-5K dataset like stated before. However, since the problem of tone-mapping often refers to enhancing images obtained in low-light conditions by security cameras with small sensors, the input images are darkened to simulate this effect. This is achieved by taking off the gamma correction from the image, dividing it by a random integer in $[1, 128]$, and adding gamma again. The resulting darkened images are saved and used as input for all further tests. This introduces an additional challenge to our model, as it needs to perform local boosting, where in some areas it has to boost more than in others.

4.1 Experimental setup

The model used (Sec. 2) itself consists of the generator and the discriminator. The generator consists of 8 layers of down-sampling with stride 2 and kernel size 3, and 8 layers of up-sampling with size 2 and kernel size 3. There are skip connections between equally sized layers in the down and up-sampling. The network is optimised using the Adam optimiser [11]. The learning rate is 0.01 and has a decay factor 0.9. The learning rate is cyclic, with a decreasing reset point to avoid getting stuck in a saddle point or local minimum. The discriminator network is also based on convolution and also uses stride 2 and kernel size of 3. Adam was also used for optimization as well as the cyclic learning rate at 0.01. All experiments were performed on a NVidia Titan X GPU, and the model is implemented in Keras with a Tensorflow backend.

4.2 Experimental results: comparison with state-of-the-art

The model will be compared qualitatively and quantitatively to the state-of-the-art models that are currently achieving the best performance, as described in Sec. 1, i.e. Deep Photo Enhancer (DPE) [19], GLADnet (GLobal illumination Aware and Detail preserving Network)

Network	PSNR	MSE	FSITM
Deep Photo Enhancer	13.04	3912.6	0.8490
GLADNet	13.79	3185.17	0.8455
Deep Retinex	14.24	1692.41	0.8431
proposed network	**19.72**	**1054.49**	**0.9243**

Table 1: Quantitative comparison between the proposed method and the state-of-the-art DPE, GLADNet and Deep Retinex. The proposed approach achieves better results for all scores.

[18] and Deep Retinex [5]. The first one is, like the name suggests, a general image enhancement network that aims to improve the overall quality of the image. The second network is designed to improve low-light images using global awareness, while preserving details. The last one is similar to GLADNet in functionality but has a different implementation. For both GLAD- Net and Deep Retinex, the weights provided on their respective Github pages were used for all experiments. For Deep Photo Enhancer the demo website was used, using their best (HDR) model.

Quantitative Comparison:
The first comparison will be quantitative: 442 images from the MIT-5K dataset are fed through all networks described above and through the proposed network, giving the scores in Table 1. These scores are computed by taking the outputs of each of the networks and comparing them to the expert reference image, and comprise of: (1) Peak Signal to Noise Ration (PSNR), where higher is better, (2) Mean Squared Error (MSE), where lower is better, and (3) Feature Similarity Index for ToneMapped images [15] (FSITM), which is higher when the resulting image quality is better. As Table 1 shows, our method clearly outperforms the state-of-the-art on all measures. These were chosen since they are used in other papers in the image enhancement community to compare models.[21] No cross-validation was used to generate these results since retraining the model on every fold would take too much time and the dataset after augmentation was quite large ($\pm$25000 samples).

Qualitative Comparison:
The improvements brought on by the proposed method are more clearly visible in qualitative comparisons, and since the method is focussed on extremely dark images, several examples are shown below where the input is very dark. Next to the input image are the enhanced versions by GLADNet, DPE, Deep Retinex and our method. We have chosen images with a different kind of content, and with very low-light, to demonstrate the effectiveness of our approach and compare it with the state-of-the-art. The images are taken from the test set and are thus unseen data to all model. Important to note is that this includes images from both MIT-5k as well as from the LOL dataset. In all figures below, one can observe that DPE produces images that are quite dark and with details still lacking. GLADNet tends to brighten the entire image, so it works like a boosting filter, instead of locally highlighting dark areas and revealing details, while reducing noice. Deep Retinex produces unnatural looking images, which maybe originate from the separately processing of reflectance and illumination. Supplementary material has been provided with this submission, with additional images for qualitative comparison of the different methods, where it can be seen that these differences in image quality persist.

Figure 3: (a) Original, (b) Expert C, (c) DPE, (d) GLADNet, (e) Deep Retinex, (f) Proposed approach. The proposed method is qualitatively closest to the outcomes from Expert C.

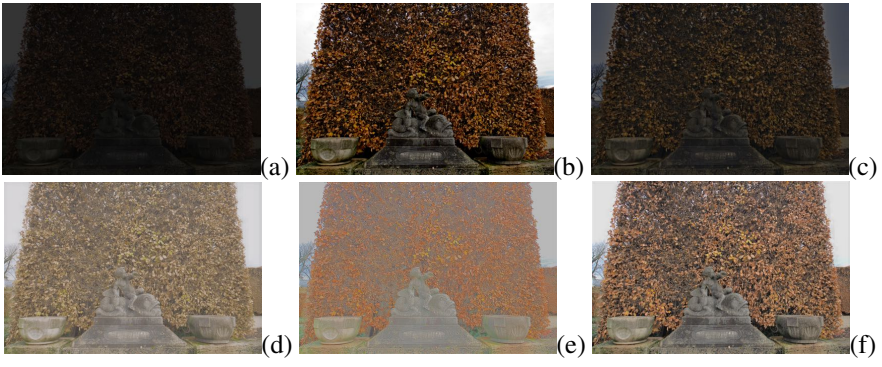

Figure 4: (a) Original, (b) Expert C, (c) DPE, (d) GLADNet, (e) Deep Retinex, (f) Proposed approach. The proposed method is qualitatively closest to the outcomes from Expert C.

Figure 5: (a) Original, (b) Expert C, (c) DPE, (d) GLADNet, (e) Deep Retinex, (f) Proposed approach. The proposed method is qualitatively closest to the outcomes from Expert C.

5 Conclusions

This work introduced a network that achieves tone mapping of low-light images using GANs, but only leveraging the lightness from a given input image to generate a new lightness map

Figure 6: (a) Original, (b) Expert C, (c) DPE, (d) GLADNet, (e) Deep Retinex, (f) Proposed approach. The proposed method is qualitatively closest to the outcomes from Expert C.

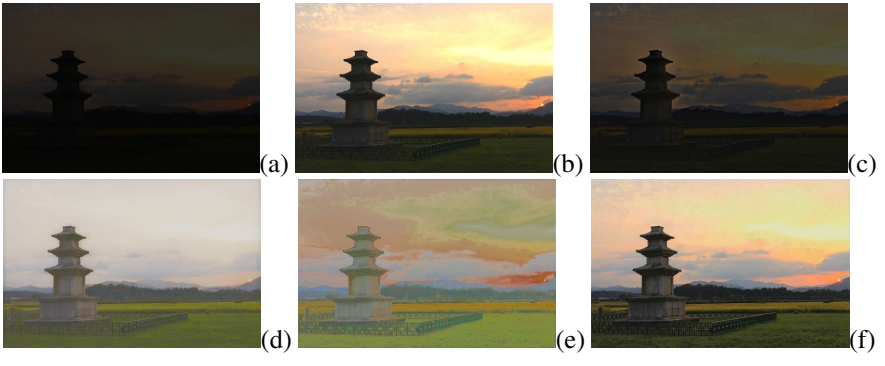

Figure 7: (a) Original, (b) Expert C, (c) DPE, (d) GLADNet, (e) Deep Retinex, (f) Proposed approach. The proposed method is qualitatively closest to the outcomes from Expert C.

on a per pixel basis. The consideration of the lightness alone, which contains the essential appearance information of the image, in combination with a psycho-visually relevant error metric in the generator and patchGAN discriminator is expected to lead to visually pleasing results. Indeed, extensive experimental testing on the benchmark dataset MIT 5K demonstrated the superiority of our approach, both quantitatively and qualitatively, over the state-of-the-art. The proposed method leads to promising results on image tone mapping, especially on very dark input images where the tested state of the art image enhance networks lag behind. This is because it can boost the input image on a local level, which means different boost strengths can be used for different parts of the image. When using already good images, the network will leave them basically unchanged, avoiding over-brightening them.

Further research directions include extending this work so as to address even more challenging scenarios, involving back-light conditions, problems related to High Dynamic Range (HDR) and complex shadows. Also, the metrics in the quantitative comparison could be extended with more metrics that capture better the essence of what tone mapping is trying to achieve. For example, the Natural Image Quality Evaluator (NIQE) could be used which is a well known image quality index that is used in evaluating image restoration[14]. In addition to the NIQE, a user study would provide the most relevant qualitative results, as long as it is

over a large and diverse group.

Acknowledgement: *This work was carried out while the first author was an intern at Bosch, Eindhoven.*

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] A. Boschetti, N. Adami, R. Leonardi, and M. Okuda. High dynamic range image tone mapping based on local histogram equalization. In *2010 IEEE International Conference on Multimedia and Expo*, 2010.

[3] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 97–104. IEEE, 2011.

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[5] Wenhan Yang Jiaying Liu Chen Wei, Wenjing Wang. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*. British Machine Vision Association, 2018.

[6] John Y. Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Transactions on Image Processing*, 2012.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)*, 37(2):26, 2018.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[10] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[12] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2017.

[13] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[14] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[15] Hossein Ziaei Nafchi, Atena Shahkolaei, Reza Farrahi Moghaddam, and Mohamed Cheriet. Fsitm: A feature similarity index for tone-mapped images. *IEEE Signal Process. Lett.*, 22(8):1026–1029, 2015.

[16] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5928–5936, 2018.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[18] Wenjing Wang, Chen Wei, wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. *2018 13th IEEE International Conference on Automatic Face ands Gesture Recognition (FC 2018)*, 2018. doi: 10.1109/fg.2018.00118.

[19] Yu-Sheng Chen Yu-Ching Wang and Man-Hsin Kao Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans.

[20] Huan Yang, Baoyuan Wang, Noranart Vesdapunt, Minyi Guo, and Sing Bing Kang. Personalized attention-aware exposure control using reinforcement learning. *arXiv preprint arXiv:1803.02269*, 2018.

[21] Runsheng Yu, Wenyu Liu, Yasen Zhang, Zhi Qu, Deli Zhao, and Bo Zhang. Deepexposure: Learning to expose photos with asynchronously reinforced adversarial learning. In *Advances in Neural Information Processing Systems*, pages 2149–2159, 2018.