Learning to See in the Dark

Vidhey Oza
Khoury College of Computer Sciences
Northeastern University
oza.vi@northeastern.edu

Abstract

Capturing low-light images is difficult particularly in daily-use devices like smartphones, where long-exposure or multi-exposure HDR captures are prone to shaking and bad post-processing. Using AI-based techniques requires a light model footprint, and hence developing small-scale generative models show strong potential in this problem domain. We propose a simple GAN-based training architecture that makes use of the prowess of supervised learning using the See-in-the-Dark Dataset, which contains dark-bright image pairs across different image contexts, to generate light dynamically on a dark image input to the generator. Results show PSNR of 29.07 and SSIM of 0.791, which fall in line with other benchmark techniques in the problem domain.

1. Introduction

Dark image illumination is an open research problem that has the attention of machine learning and mathematical researchers alike. Initial dark-image enhancement techniques targeted the problem mathematically, and it was quickly understood that capturing images in low light introduced 2 issues in them: first is the introduction of noise due to the range of minimum and maximum pixel values in the image, and second is the distortion in true-to-life color which would be amplified if the image were simply upscaled. Commonly used techniques include multi-exposure image captures, which are useful in theory but require significant stabilization for it to be comparable to an ideal bright image.

Current techniques in this problem domain commonly used by professional software like Photoshop, Affinity, GIMP, etc. include a two-step procedure of first upsampling the dark image, then applying mathematical denoising to it to remove the amplified noise. The main problem with such an approach is that upsampling messes with the color balance of the image, and must be manually fixed to resemble real-life. Also, the denoising reduces the sharpness of the image, reducing the general quality of the image.

Computer vision based techniques have been proposed by [2][3][10][11] that make use of convolutional networks and its variants to perform image illumination on varying image samples. Since the problem involves just adding features into the input image without altering its size or shape, fully convolutional networks can be used with the goal of encoding the image structure and dynamically adding light on a global as well as local level.

An important problem with training CNNs, especially in a supervised setting, is that datasets are not abundantly available that provide variations across different aspects of low light images like indoors-vs-outdoors setting, objects vs people in the foreground, scene variations in the background, etc. Because of this, researchers are forced to fuse multiple datasets together or create their own database, which affects the standardization of research in the problem domain.

With this in mind, I designed a simple GAN-based training architecture that uses a fully convolutional U-Net based network structure for the generator, making use of the See-in-the-Dark (SITD) dataset proposed in CVPR 2018 by [1] that includes a robust set of dark-bright image pairs that can be used for fully supervised training.

The main contributions of this project are: (a) building on the work by [1] on their SITD dataset by proposing a GAN-based supervised learning procedure, with the goal of designing a pipeline that can be easily modified in the future with more robust convolutional modules. (b) designing a data pipeline that can be used to train model in low-hardware settings by proposing an overlapping patching system that is robust against common pitfalls of convolutional networks.

The next section of the report presents some of the relevant work and a brief history of some of the unique approaches used for dark image illumination. This is followed by detailed explanation of the dataset and processing, which in turn leads into a discussion into my entire training pipeline. I then discuss the results qualitatively as well as quantitatively, then conclude the paper with my findings and what I believe can be explored in the future.

2. Related Work

Research work in image illumination is divided generally

into 2 unique ways. When represented in RGB color space, the image is enhanced separately on all three channels and fused together to get the final enhanced image [7]. On the other hand, the HSI (Hue, Saturation, Intensity) space contains a separate channel for hue, so it is left unchanged while the other two channels are altered to enhance the image [4][5][6]. This is increasingly being used as the common technique, separating the image into a color and grey space such that only the grey space, that would encode light balance and image structure, needs to be changed.

With the advent of deep learning and convolutional networks, a new method of image processing was made accessible for researchers to explore this problem domain. The idea of Fully Convolutional Networks (FCNs) was proposed as a solution of the problem of semantic image segmentation [8], where only convolutional filters are used in the entirety of the neural network, opening the idea of image-to-image training for one of the first times. This technique was used later in various problem domains where input-output image pairs could be extracted.

Papers like [9][10] approached the problem from an unsupervised learning perspective, where the model learns features about the image and uses it directly to enhance it without any ground truth data. On the other hand, image datasets like [11][12] propose that using a supervised learning approach is much simpler to execute and can be easily generalized to multiple image categories. But the problem with supervised learning is that the collected data needs to be balanced across categories like portraits, object captures and landscape photos, but also needs to be voluminous enough to reduce inevitable bias.

Authors of [1] attempt to solve this problem by standardizing their image gathering process. They release full technical details of their capturing procedure, and use it to train a simple FCN and show promising results. This helps in establishing a common ground to test unsupervised as well as supervised learning methods. The next section will describe the dataset in more detail.

Generative Adversarial Networks (GANs) [13] are recently being used to train robust sets of neural networks, where the end goal is to either use the generator model(s) to generate the required output, or the discriminator model(s) to classify an artificially generated or altered image from a real one. While initial generators used seed noise as input to generate image or other data, they are now being trained on images as seed that need to be altered.

Authors of [9][14] modified the training procedure of GANs to perform self-supervised learning for low-light image enhancement. These are among the state-of-the-art, and involve complex model architectures to achieve the model efficacy. On the other hand, the method proposed in this project outlines a simpler GAN based technique that makes use of the supervised learning dataset proposed in [1].

Figure 1: Dark-bright image pair samples from SITD dataset. Top - high difference in illumination, bottom - low difference in illumination.

3. Dataset Extraction & Processing

The data used in the project comes from the proposed See-In-The-Dark (SITD) dataset that includes dark-bright image pairs in different settings. The images were captured using two different cameras, Fujifilm and Sony, which is important to note since both use their own kind of sensors to capture the image. The Sony α7S uses a Bayer-style sensor, while Fujifilm X-T2 uses a more modern APS-C X-Trans sensor. Both have differing ways to arrange the sensed light using filter arrays for the green light wavelength. The cameras were mounted on tripods to avoid any shaky imaging in high-exposure captures, and both were chosen to be mirrorless to avoid any vibration during the capture event. Exposure is the main variant between the dark and the bright images, with dark images exposed at 1/10s to 1/30s and bright images exposure set anywhere from 10s to 30s. Other differences like focal length and aperture are used to counter any unwanted changes between the pairs.

The main drawback of using this dataset is the size of the dataset, not just in terms of volume but of each dark-bright pair. The resolution is 4240×2832 for Sony and 6000×4000 for the Fuji images, which loaded together with the fully convolutional networks make the model size upwards of 4M parameters. This not only hampers the model processing speed but is also difficult to train without incorporating bias.

To solve this problem, the idea of patching is used to increase the data volume while reducing the size of each pair. The patches are generated by extracting a section with a given dimension and moving across the image with a given stride or step size. For maintaining image structure while balancing for reducing memory usage, I selected a patch size of 100x100, but chose a stride of 50. This is because patching and merging the output image back would lead to unwanted noise at the borders since convolution operation is weaker at the edges than it is with the pixels in

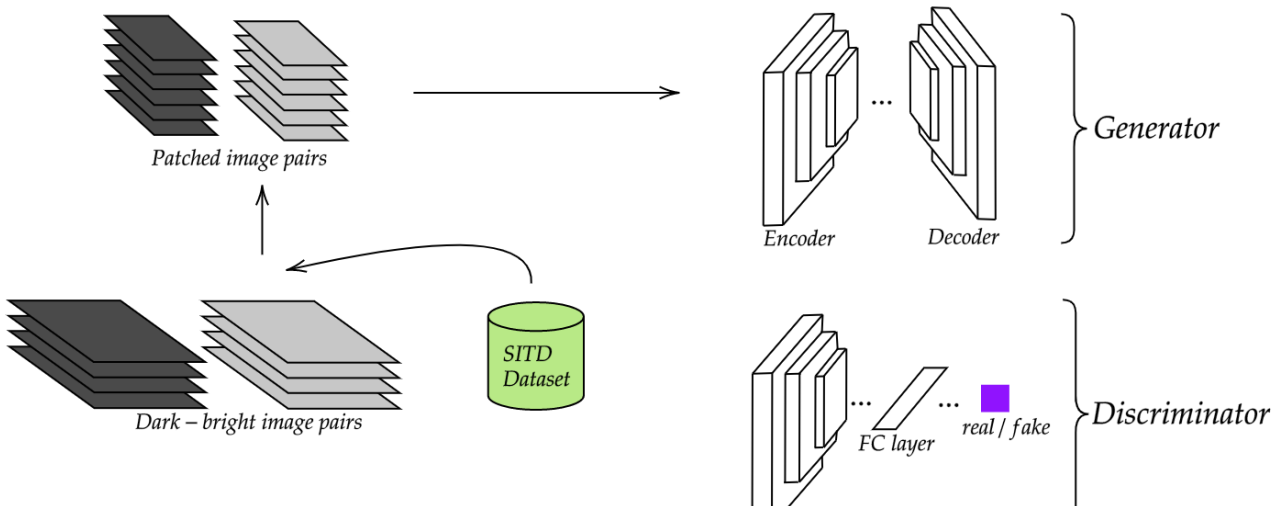

Figure 2: Proposed technique architecture. First, the images are extracted from the dataset and dark-bright pairs are generated. The pairs are then patched with overlapping, then used to train the generator and discriminator in the dual GAN training architecture.

the central area.

4. Proposed Technique

The proposed technique is an extension to the generative adversarial network architecture that includes a fully convolutional network. The entire technique with the data processing is given in Fig. 2, which shows the data flow from the database all the way to the training procedure.

To reiterate, the GAN design is a dual model architecture, which are trained simultaneously with the end goal of making the generator robust to artifacts in the output image. The generator is trained to accept seed input that is passed through the network to generate desired target output image. The discriminator is trained to classify between the real bright image and the generated (or fake) image, with the end goal that the discriminator is fooled by the generator.

The generator in the GAN model pair is a fully convolutional network that takes input the dark image, and is trained to produce a corresponding bright image. With the SITD dataset, such training is possible in a supervised setting, with each bright image having multiple associated bright images with different exposure settings. This is done with the goal of balancing the training of the generator and including not just the darkest images but also lower-mid level illumination. The model is designed in a U-Net architecture with 3-layer encoder and decoder modules. The encoder module uses max pooling layers after convolutional filters to bottleneck the model into encoding important features from the input image. After the final encoding layer, the decoder uses an unpooling layer after every convolutional layer that corresponds to each max pooling layer directly. Since each encoding step is responsible for learning hidden features at different scales with respect to the image, each pre-max-pooling layer has a skip connection to the post-un-pooling layer. This is done to propagate knowledge during backpropagation without having to encounter problems like vanishing gradient or altered encoding.

The discriminator is designed as a simple convolutional model, with 2 convolutional blocks containing max pooling layers and 2 fully connected blocks with dropout layers. The convolutional layer is designed so as to isolate any image size reduction to max pooling layers only. Such a simple model is designed so that the discriminator stays robust against various artifacts that may be created by the generator on the output image, and the latter is forced to generate more realistic looking images. These models are trained simultaneously while alternating between backpropagation in the generator and the discriminator.

5. Results & Discussion

The proposed technique is trained on the Sony and Fujifilm datasets separately, and results from them are reported separately. First, I discuss a higher-level qualitative analysis of the results, then take a deeper dive with quantitative evaluation using different metrics.

The pipeline is compared to two categories of models: traditional denoising techniques, and other state-of-the-art computer vision models. The traditional pipeline involves an upscaling module followed by denoising. The choice of denoising model is done in line with current state-of-the-art techniques used widely across domains, and BM3D is used as the baseline benchmark. Other computer vision techniques used for comparison are the model proposed by SITD paper [1], EnlightenGAN [2] and LEUGAN [3]. This selection is done to compare the proposed technique with the baseline model proposed by the authors of the dataset

and compare against other state-of-the-art generative techniques using unsupervised learning procedures. Since these models show the best industry-level performance, this comparison is done to gauge the efficacy of the proposed technique that is not expected to be better than these models. Instead, this is a relative benchmarking for different learning procedures. Another logistic issue with comparing the models was the incompatibility of the data, since the SITD images are 4-channel while all other techniques use 3-channel RGB images as input. To solve this, a simple channel conversion function is designed to convert both Sony and Fujifilm images to RGB arrays before testing. Image samples of all techniques are given in the code repository for reference, provided in the Appendix.

Qualitatively, we see that the traditional pipeline has a clear pitfall in image enhancement. The upscaling+BM3D pipeline processes the image by increasing pixel values by a constant value, which not only introduces noise in the image but also significantly disturbs its color balance. BM3D solves one of the problems by attempting to denoise the image but in relation to the botched color balance BM3D is either not effective at all or ends up considerably reducing the sharpness of the image. Hence such a pipeline is only useful in a manual setting where the color balance is manually fixed before applying denoising.

The ML techniques show a more realistic color rebalancing since they have looked at bright image counterparts. But since the unsupervised techniques have not seen the ground truth images, they too struggle with color balance, and end up botching at least one portion of the image. However, LEUGAN generally shows the best performance, with a better understanding of reflections and light sources, and retaining color close to real life even with pitch-black images.

Quantitatively, the model is compared on 2 different metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). PSNR is a commonly used metric for comparing similarity between images and detecting noise difference between them. SSIM is a more robust technique that considers 3 aspects of image similarity: luminance, contrast and structure. Together these metrics are useful in comparing the models against one another. Table 1 reports the metrics for all the techniques discussed.

Model	PSNR	SSIM
Upscale+BM3D	18.23	0.674
SITD [1]	27.74	0.733
EnlightenGAN	31.40	0.809
LEUGAN	32.98	0.823
Proposed	29.07	0.791

The proposed technique, at PSNR 29.07 and SSIM 0.791, performs significantly better than the traditional baseline technique (PSNR 18.23, SSIM 0.674). This is in line with the discussions done in the previous section about the shortcomings of upscaling+BM3D. My technique is also slightly better than the model proposed by authors of SITD [1] (PSNR 27.74, SSIM 0.733). This can be attributed to the robust adversarial learning procedure of the GAN architecture, combined with the simplicity of supervised learning. The model falls slightly short of the goal of PSNR 30 and SSIM 0.8, which is widely considered as a threshold for highly comparable image pairs, but reaches close to those metric values.

EnlightenGAN and LEUGAN both perform better than the proposed technique, showcasing the prowess of unsupervised learning in computer vision domains. While supervised learning is simple and effective on its own, unsupervised training makes the model more generalized and robust to unseen images.

6. Conclusion & Future Work

The project proposes a variation on the supervised GAN training procedure as a viable solution to the problem of low light image enhancement. This is made possible with the SITD dataset which is generalized enough to include images from various scenes, backgrounds and foregrounds, along with standardized image capturing techniques to assure similarity between characteristics of the image which being diverse enough to stay robust against any model bias. The generator is designed in a U-Net like architecture with an encoder-decoder architecture made entirely of convolutional layers, while the discriminator remains fairly unchanged.

I believe that there can be improvements made in many aspects of the proposed pipeline. the biggest drawback of using RAW images for training is their 4-channel image structure, which prevents it from being used directly for commonly found images in JPG or PNG formats. So, gathering images in common image formats or constructing an efficient module to converting the RAW images to these formats can be a step in the right direction. The generator processing time is ~50ms, which is too high for processing a 30fps video in real time, so reducing model size can be another direction forward. This can also help in future deployment on mobile devices with the TFLite backend so as to locally use these illumination models.

Appendix

The code is available on GitHub at: https://github.com/vidheyoza/CS7180-AdvancedCV/tree/master/SeeInTheDark/see-in-the-dark. The code is a part of the main course repository, and following this link gives access to the main project code with a detailed README.md for instructions.

References

[1] Chen, Chen, et al. "Learning to see in the dark." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018*.

[2] Jiang, Yifan et al. “EnlightenGAN: Deep Light Enhancement Without Paired Supervision.” *IEEE Transactions on Image Processing 30 (2021): 2340-2349.*

[3] Qu, Yangyang, and Yongsheng Ou. "LEUGAN: Low-Light Image Enhancement by Unsupervised Generative Attentional Networks." *arXiv preprint arXiv:2012.13322 (2020).*

[4] X. Jie, H. LiNa, G. GuoHua and Z. MingQuan, "Based on HSV Space Real-color Image Enhanced by Multi-scale Homomorphic," *2009 WRI Global Congress on Intelligent Systems, 2009*, pp. 160-165, doi: 10.1109/GCIS.2009.295.

[5] B. Gupta and T. K. Agarwal, "New contrast enhancement approach for dark images with non-uniform illumination", *Comput. Electr. Eng.*, vol. 70, pp. 616-630, Aug. 2018.

[6] M. Iqbal, S. S. Ali, M. M. Riaz, A. Ghafoor and A. Ahmad, "Color and white balancing in low-light image enhancement", *Optik*, vol. 209, May 2020.

[7] X. Yang, K. Xu, Y. Song, Q. Zhang, X. Wei, and R. Lau, ‘‘Image correction via deep reciprocating HDR transformation,’’, *arXiv preprint arXiv:1804.04371 (2018).*

[8] Long, J. et al. “Fully convolutional networks for semantic segmentation.” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015): 3431-3440.

[9] H. Lee, K. Sohn and D. Min, "Unsupervised Low-Light Image Enhancement Using Bright Channel Prior,", *IEEE Signal Processing Letters*, vol. 27, pp. 251-255, 2020.

[10] W. Wang, C. Wei, W. Yang and J. Liu, "GLADNet: Low-Light Enhancement Network with Global Awareness," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 751-755, doi: 10.1109/FG.2018.00118.

[11] W. Yang, W. Wang, H. Huang, S. Wang and J. Liu, "Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 2072-2086, 2021, doi: 10.1109/TIP.2021.3050850.

[12] Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014). “Generative Adversarial Networks”. *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*. pp. 2672–2680.

[13] Qu, Yangyang et al. “LEUGAN: Low-Light Image Enhancement by Unsupervised Generative Attentional Networks.” *ArXiv preprint ArXiv 2012.13322* (2020).