Ambient Lighting Generation for Flash Images with Guided Conditional Adversarial Networks

José Chávez[1], Rensso Mora[1] and Edward Cayllahua-Cahuina[2]
[1]*Department of Computer Science, Universidad Católica San Pablo, Arequipa, Peru*
[2]*LIGM, Université Paris-Est, Champs-sur-Marne, France*

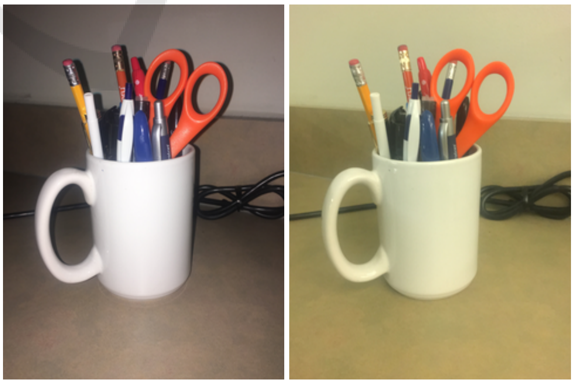
Keywords: Flash Images, Ambient Images, Illumination, Generative Adversarial Networks, Attention Map.

Abstract: To cope with the challenges that low light conditions produce in images, photographers tend to use the light provided by the camera flash to get better illumination. Nevertheless, harsh shadows and non-uniform illumination can arise from using a camera flash, especially in low light conditions. Previous studies have focused on normalizing the lighting on flash images; however, to the best of our knowledge, no prior studies have examined the sideways shadows removal, reconstruction of overexposed areas, and the generation of synthetic ambient shadows or natural tone of scene objects. To provide more natural illumination on flash images and ensure high-frequency details, we propose a generative adversarial network in a guided conditional mode. We show that this approach not only generates natural illumination but also attenuates harsh shadows, simultaneously generating synthetic ambient shadows. Our approach achieves promising results on a custom FAID dataset, outperforming our baseline studies. We also analyze the components of our proposal and how they affect the overall performance and discuss the opportunities for future work.

1 INTRODUCTION

Scenes with low light conditions are challenging in photography, cameras usually produce noisy and/or blurry images. In these situations, people usually use an external device such as a camera flash, thus, creating flash images. However, when the light from the flash is pointing directly at the object, the light can be too harsh for the scene and create a non-uniform illumination. Comparing a flash image with its respective image with ambient illumination, it is clear that the illumination is more natural and uniform because the available light can be more evenly distributed (see Figure 1).

Researchers have studied the enhancement of flash images (Petschnigg et al., 2004; Eisemann and Durand, 2004; Agrawal et al., 2005; Capece et al., 2019), producing enhanced images by the combination of such ambient and flash images, or normalizing the illumination on flash image in a controlled environment (backdrop and studio lighting), but without replicating the natural skin tone of people. However, in a real scenario with low light conditions, there is no information about how the ambient image is. On the other hand, on scenarios without a backdrop, objects away from the camera will have very low illumination, thus, creating dark areas in the image, considering that there is only the illumination of the camera flash. Consequently, in a real scenario with low light conditions, creating ambient images from flash images poses a very challenging problem.



(a) Flash image (b) Ambient image

Figure 1: A comparison of a flash image and an ambient image. (a) Image with camera flash illumination. The image suffers from harsh shadows, dark areas and bright areas. (b) Image with available ambient illumination. In this image, the illumination is more uniform, natural, and the image has not sideways shadows. Images extracted from FAID (Aksoy et al., 2018).

Chávez, J., Mora, R. and Cayllahua-Cahuina, E.
Ambient Lighting Generation for Flash Images with Guided Conditional Adversarial Networks.
DOI: 10.5220/0008983603810388
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 381-388
ISBN: 978-989-758-402-2; ISSN: 2184-4321

Prior works handle the enhancement of low light images, where a scene is underexposed; however, on flash images, objects close to the camera tend to be bright and these techniques overexpose these regions. Our method attenuates the illumination that is close to the camera, and illuminates the underexposed regions at the same time. Since flash and ambient images represent the same scene, researchers (Capece et al., 2019) study the lighting normalization on a flash image by learning the relationship between both images to estimate a relationship between these pair of images, which is added to the respective flash image in a next step, thus, normalizing the illumination on flash images but maintaining high-frequency information. This approach is not effective to restore overexposed areas due to this region still needs to compute the final result.

In this article, we propose a conditional adversarial network in a guided mode, which follows two objective functions. First, the reconstruction loss generates uniform illumination and synthetic ambient shadows. Second, the adversarial loss, which represents the objective function of GANs (Goodfellow et al., 2014), forces to model high-frequency details on the output image, and perform a more natural illumination. Both loss functions are guided through the attention mechanism, which is performed by attention maps based on the input image and ground truth. The attention mechanism allows to the model to be more robust to overexposed areas and sideways shadows presented on flash images. It also improves the robustness of the model on inconsistent scene match between pairs of flash and ambient images since they are both usually not perfectly aligned at the moment of capture. We compare against state-of-the-art enhancement techniques for low light images (Fu et al., 2016; Guo et al., 2017), and flash images (Capece et al., 2019). Ablation studies are also performed on the architecture.

Then, the major contributions of this article are:

- An attention mechanism to guide a conditional adversarial network on the task of translating from flash images to ambient images. Giving robustness against overexposed areas and shadows presented on flash and ambient images, and the misaligned scene between both images. This mechanism guides the adversarial loss to avoid blurry results on regions by discriminating these cases.
- Our proposed attention mechanism also guides the reconstruction loss to be robust against high-frequency details thought the texture information that the attention map gives.

2 RELATED WORK

2.1 Low Light Image Enhancement

Prior works (Petschnigg et al., 2004; Eisemann and Durand, 2004; Agrawal et al., 2005) combine the advantages of both ambient and flash images. These image processing techniques use the information of the image with the available illumination (ambient image) and the image with light from the camera flash (flash image) and create an enhanced image based on both images. In contrast with these techniques, our model enhances the flash image but without any kind of information of the ambient image.

In SRIE (Fu et al., 2016), the reflectance and illumination are estimated by a weighted variational model, then, the images are enhanced with the reflectance and illumination components. LIME (Guo et al., 2017), on the other hand, enhance the images by the estimation of their illumination maps. More specific, the illumination map of each pixel is first estimated individually by finding the maximum value in the R, G and B channels, then the illumination map is refined by imposing a structure prior. This refined illumination map has smoothness texture details. Both methods SRIE and LIME do not contemplate sideways shadows removal, reconstruction of overexposed areas or generation of synthetic ambient shadows.

2.2 Image-to-Image Translation

Prior works use symmetric encoder-decoder networks (Ronneberger et al., 2015; Isola et al., 2017; Chen et al., 2018) for image-to-image translation such as: image segmentation, synthesizing photos, enhancing low light images, etc. These networks are composed of various convolutional layers, where the input is encoded to a latent space representation and then decoded to estimate the desired output. Inspired on the U-Net architecture (Ronneberger et al., 2015), our model employs skip connections to share information between encoder and decoder, to recover spatial information lost by downsampling operations.

In (Capece et al., 2019), a deep learning model turns a smartphone flash selfie into a studio portrait. The model generates a uniform illumination, but not reproduce the same skin tone of the person under studio lighting. The encoder part of the network represents the first 13 convolutional blocks of the VGG-16 (Simonyan and Zisserman, 2015), and the weights of the encoder are initialized with a pre-trained model for face-recognition (Parkhi et al., 2015). The inputs and target of this network are given filtered, to estimate an image with low-frequency details, which represent the relationship on illumination between the ambient and flash image. This pre-processing step is the drawback of this model because it can not learn a high-quality relationship of illumination between the flash and the ambient image. This step also has a computation time due to the model uses a bilateral filter.

We exploit the transfer learning approach of this model, but we proposed an end-to-end architecture where the encoder path is initialized with the VGG-16 pre-trained on the ImageNet dataset (Deng et al., 2009), thus, making our model for general scenes, not only for faces. And the decoder part is symmetric respect to the encoder. The end-to-end architecture also avoids an additional pre-processing step.

2.3 Conditional GANs

Conditional GANs (Mirza and Osindero, 2014) have been proposed as a general purpose for image-to-image translation (Isola et al., 2017). A cGAN is composed of two architectures, the generator, and the discriminator. Both architectures are fully convolutional networks (Long et al., 2015). On the generator, which represents an encoder-decoder network, each step of the encoder and decoder is mainly composed by convolutional layers. The generator G and discriminator D are conditioned on some type of information such as images, labels, texts, etc. In our case, this information represents the flash images I_f, and our cGAN learns to map from flash images I_f to ambient images I_a. Thus, the generator synthesizes ambient images $\hat{I}_a$, which can not be distinguished from the real ambient images I_a, while the discriminator is trained in adversarial form respect to the generator to distinguish between I_a and $\hat{I}_a$. As it shows in pix2pix model (Isola et al., 2017), this min-max game ensure the learning of high-frequency details unlike using only a reconstruction loss like a MAE (Mean Absolute Error), which output smoothed results.

3 PROPOSED METHOD

Our model is composed of two architectures, generator G, and discriminator D; and translate from flash images I_f to ambient images I_a. Then, the training procedure follows two objectives: the reconstruction loss **R**, which aims to minimize the distance between the input image (I_f) and the target image (I_a); and the adversarial loss **A**; which represent the objective of the cGAN (Isola et al., 2017). Figure 2 illustrates an overall of our architecture model.

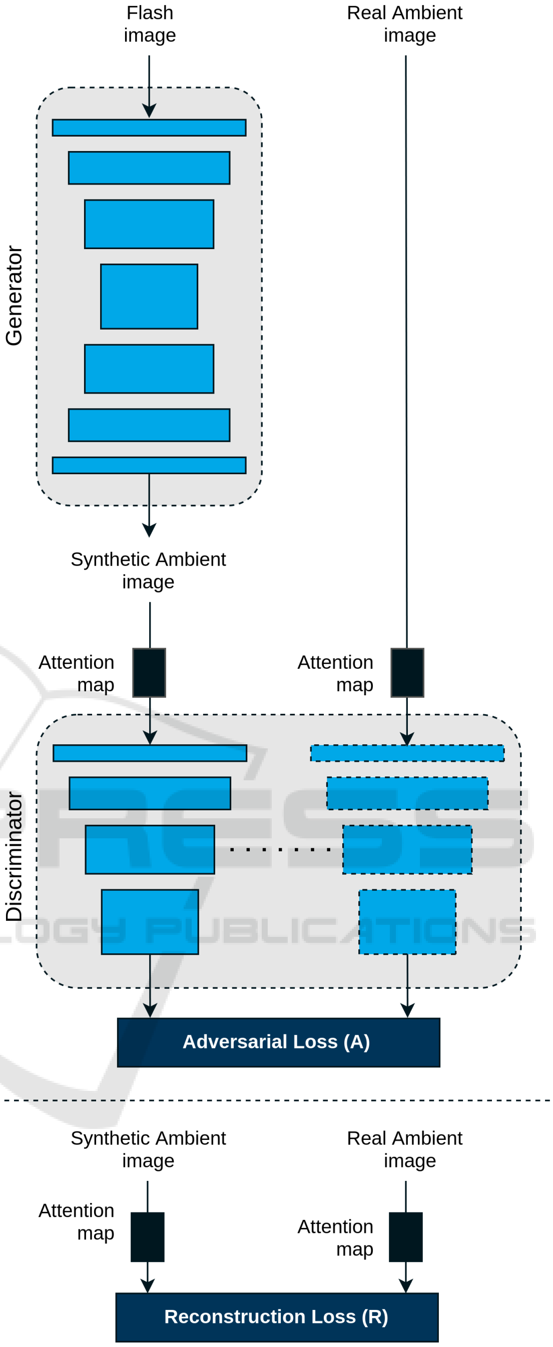


Figure 2: Network architecture. The generator has as its input the flash image I_f and as its output the synthetic ambient image $\hat{I}_a$. The discriminator network learns through the adversarial loss **A** to classify between the real ambient image I_a, this is the ambient image that belongs to the training set, and the synthetic ambient image $\hat{I}_a$. We also set the reconstruction loss **R** between I_a and $\hat{I}_a$. All attention maps are compute thought I_f and I_a.

Both the reconstruction loss **R** and the adversarial loss **A** are guided by our attention mechanism to ensure a better learning procedure. The attention mechanism is performed on the entries of **R** and **A**, that is, the ambient image I_a and synthetic ambient image $\hat{I}_a$ first pass through the attention map before the computation of **R** and **A**.

3.1 Attention Mechanism

The attention mechanism that we propose aims to guide the reconstruction and adversarial loss. The mechanism is simple but efficient, we guide both **R** and **A** with an attention map base on the flash image I_f and the ambient image I_a. We define the attention map $\mathcal{M}$ as:

$$\mathcal{M}(i,j) = 1 - \frac{1}{C}\sum_{k=1}^{C} \mid I_a(i,j,k) - I_f(i,j,k) \mid . \qquad (1)$$

In Equation 1, C represents the number of channels and $\mathcal{M}(i,j)$ the value of the attention map at the position (i,j). $I(i,j,k)$ represent the pixel value at (i,j) and channel k. Then, I_a and $\hat{I}_a$ pass though the attention map before compute the reconstruction loss **R** and the adversarial loss **A**,

$$I_a := I_a \otimes \mathcal{M}, \hat{I}_a := \hat{I}_a \otimes \mathcal{M}. \qquad (2)$$

The operation $\otimes$ represents the element-wise multiplication. Equation 2 guides **A** and **R** to a better learning procedure through the discrimination of overexposed areas, shadows, and scene misalignment, between I_f and I_a. Then **R**, which represent the L1 distance, and **A** are defined as:

$$\begin{aligned} \mathbf{R}(G) &= \mathbb{E}_{I_f \sim p_{\mathbf{data}}, I_a \sim p_{\mathbf{data}}}\left[\left\| I_a - G(I_f) \right\|_1 \right] \\ \mathbf{A}(D,G) &= \mathbb{E}_{I_a \sim p_{\mathbf{data}}}\left[\log D(I_a) \right] \\ &\quad + \mathbb{E}_{I_f \sim p_{\mathbf{data}}}\left[\log(1 - D(G(I_f))) \right]. \end{aligned} \qquad (3)$$

By this operation, the reconstruction loss **R** is conducted to learn the normalization of the lighting, discriminating the high-frequency details because the attention map $\mathcal{M}$ gives this information by the element-wise multiplication. $\mathcal{M}$ also guides **R** to be robust for the misaligned scene between flash and ambient images. On the other hand, the adversarial loss **A** is focused on generating realism and high-frequency details on the regions indicated by $\mathcal{M}$. **A** not allows blurry outputs where the attention map $\mathcal{M}$ indicates, because all blurry regions are classified as fake and the adversarial loss tries to fix it by generating high-frequency details on these regions.

Finally, our full objective $\mathcal{L}$ is a mix of the reconstruction and the adversarial loss, maintaining the relevance of the reconstruction loss and scaling the adversarial loss by the hyperparameter λ. Equation 4 allows determining to what extent the adversarial loss **A** should influence to $\mathcal{L}$, thus, controlling the generation of artifacts in the output images.

$$\mathcal{L}(G,D) = \mathbf{R}(G) + \lambda \cdot \mathbf{A}(G,D). \qquad (4)$$

We perform ablation studies on the architecture, and verify the improvements of using our proposed attention mechanism. Our ablation studies also consider the use and not of a pre-trained model in the generator.

4 EXPERIMENTS

In this section, we describe the Flash and Ambient Illumination Dataset (FAID) and the custom set of these images that we use. We present the training protocol that we followed and show the quantitative and qualitative results that validate our proposal. Finally, we present the controlled experiments that we perform to determine how the components of our architecture affect the overall performance.

4.1 Dataset

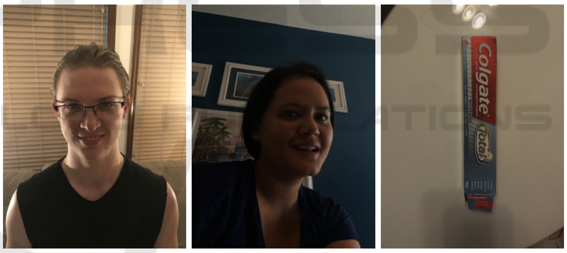


Figure 3: Ambient images from FAID (Aksoy et al., 2018) with low illumination, reflections, and shadows from external objects.

Introduced by (Aksoy et al., 2018), the FAID(Flash and Ambient Illumination Dataset) is a collection of pairs of flash and ambient images, which present 6 categories: *People*, *Shelves*, *Plants*, *Toys*, *Rooms*, and *Objects*. As a result, we have 2775 pairs of flash and ambient images. We inspected each image in the dataset and found that there exist ambient images that have problems such as low illumination, shadows from external objects or even reflections. Therefore, we used a reduced set of the entire FAID dataset for our experiments. Finally, our custom dataset has 969 pairs of images for training and 116 for testing and all images were resized to 320 × 240 or 240 × 320 depending on their orientation.

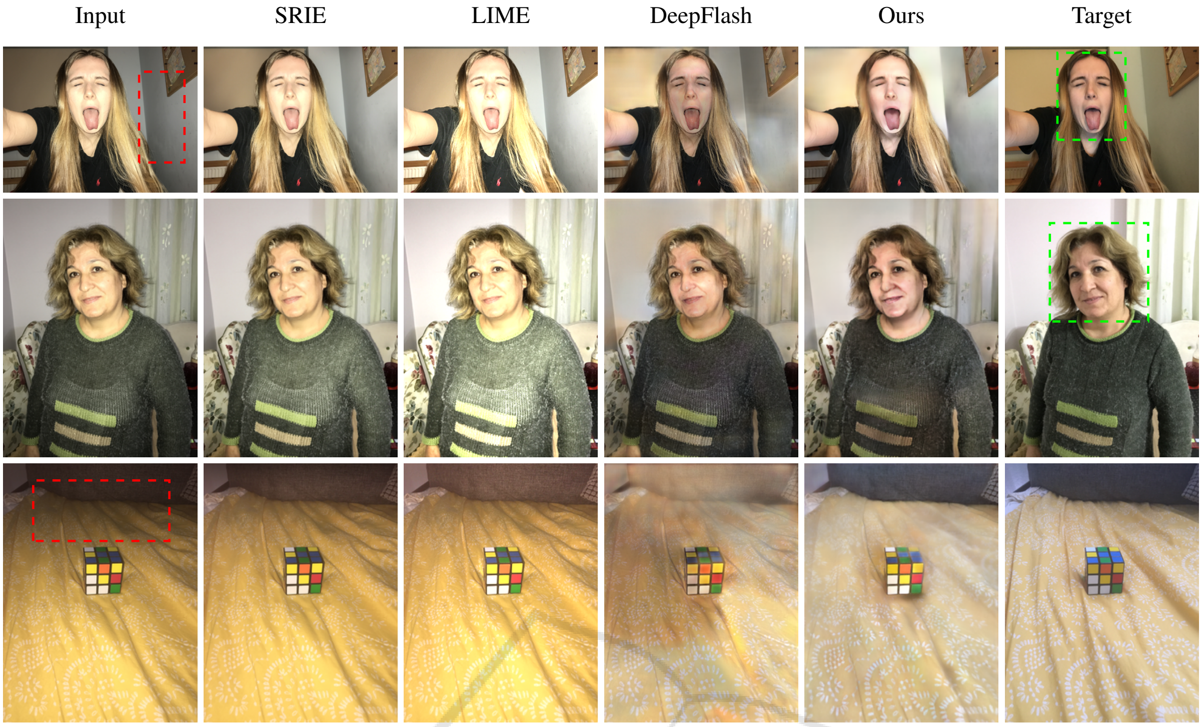


Figure 4: Qualitative comparison. Enhancement of low-illuminated areas (red), and estimation of natural skin and air tone of people (green). We compare with SRIE (Fu et al., 2016), LIME (Guo et al., 2017), and DeepFlash (Capece et al., 2019).

4.2 Training

We freeze all convolutional layers of in the encoder part of the generator, and train our model using the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.5$, based on (Isola et al., 2017). Using learning rates $2 \cdot 10^{-5}$ and $2 \cdot 10^{-6}$ for the generator and the discriminator respectively, equal or higher learning rate of the discriminator respect to the generator results on a divergence. To regularize the adversarial loss A, we set $\lambda = 1$, fewer values for λ results on blurry outputs and higher values of λ results on many artifacts. The training procedure is performed using random crops of 224 × 224 and horizontal random flipping for data augmentation. The implementation of our architecture is in Pytorch, and the training process takes approximately one day using an NVIDIA graphics card GeForce GTX 1070.

4.3 Quantitative and Qualitative Validation

We use the PSNR (Peak Signal-to-Noise Ratio) and the SSIM (Structural Similarity) to measure the performance of our quantitative results. Table 1 reports the mean PSNR and the mean SSIM on the test set, for 1000 epochs. All hyperparameters are setting on the same way for (Capece et al., 2019), and the encoder-decoder network was pre-trained on the ImageNet dataset (Deng et al., 2009) instead on a model used for face recognition (Parkhi et al., 2015). Our quantitative results do not significantly outperform the state-of-the-art image enhancement methods, but at least shows improvements on the flash image enhancement task.

Table 1: Reporting the mean PSNR and the mean SSIM with SRIE (Fu et al., 2016), LIME (Guo et al., 2017), and DeepFlash (Capece et al., 2019).

Method	PSNR	SSIM
LIME	12.38	0.611
SRIE	14.09	0.659
DeepFlash	15.39	0.671
Ours	**15.67**	**0.684**

Estimation of the skin tone of people is shown in Figure 4, where the illumination map created by LIME (Guo et al., 2017) conducts to brightening and overexposing the flash images. LIME (Guo et al., 2017), can not distinguish the natural color of dark objects and tend to illuminate them. Results in SRIE (Fu et al., 2016) do not present considerable changes concerning the flash images on these kind of scenes. DeepFlash (Capece et al., 2019) present non-

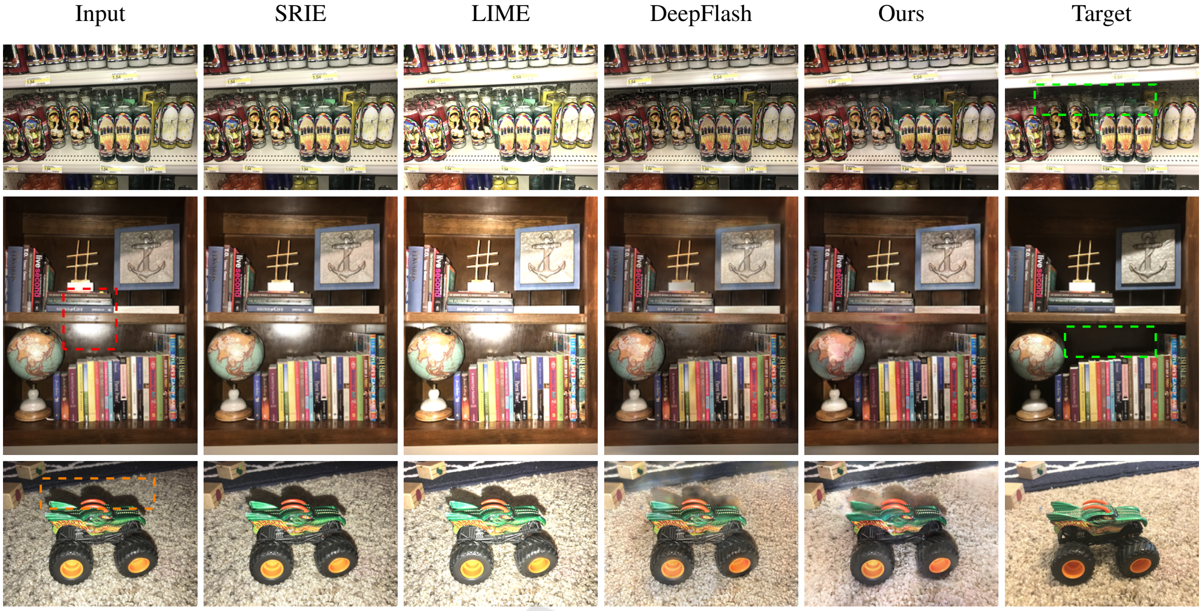


Figure 5: Qualitative comparison. Generation of ambient shadows (green), attenuation of overexposed areas (red), and sideways shadow removal (orange). We compare with SRIE (Fu et al., 2016), LIME (Guo et al., 2017), and DeepFlash (Capece et al., 2019).

uniform illumination on flash images of people, apparently this is due to trying to simulate shadows. In the case of flash images that have low illuminated areas and also high illuminated areas like the Rubik's Cube on Figure 4, (Capece et al., 2019) present meaningless illumination on their results, and our method shows considerable better results, that is, our result looks much more similar to the ground truth.

Figure 4 reveals some aspects about the generation of ambient lighting on people. Note the synthetic shadows in mouth and under the chin. Almost all ambient images from train data was taken with light source that came from above through a typical light source that exists in homes. Therefore, the model learns to generate synthetic ambient lighting simulating a light source that comes from above.

Figure 5 shows that our model synthesizes ambient shadows on flash images such as shelves, but suffer for restoring overexposed areas produced by the camera flash. LIME (Guo et al., 2017), and SRIE (Fu et al., 2016) do not attenuate overexposed areas or synthesize ambient shadows on these type of scenes, these methods do not handle this kind of issues of flash images. DeepFlash architecture (Capece et al., 2019) performs weak ambient shadows, attenuate overexposed areas without restoring them, and outputs many artifacts on their results. In the case of sideways shadow removal, all models fail (including ours).

4.4 Ablation Study

We perform different experiments to validate the final configuration of our architecture. Table 2 reports the quantitative comparison between our controlled experiments. Furthermore, we also show in Figure 6 qualitative compositions between conditions in Table 2.

Table 2: Controlled experiments. This table reports the mean PSNR and the mean SSIM for distinct architecture configurations.

Condition	PSNR	SSIM
1. **Default** ($\mathbf{R}_{\mathcal{M}} + \mathbf{A}_{\mathcal{M}}$)	**15.67**	**0.684**
2. $\mathbf{R} + \mathbf{A}$	15.55	0.676
3. $\mathbf{R}$	15.64	0.681
4. U-Net	14.81	0.643

Our quantitative assessments show that using a pre-trained model improves significantly the model trained from scratch (condition 4). The other methods seem to have similar results. This is because these models, which use the MAE for the objective function (condition 3), generate blurry results to minimize the error between estimated images and the targets. Condition 2, which is similar to the default model without the attention mechanism, has less quantitative values than condition 3 because the adversarial loss gives some sharpness on their output images.

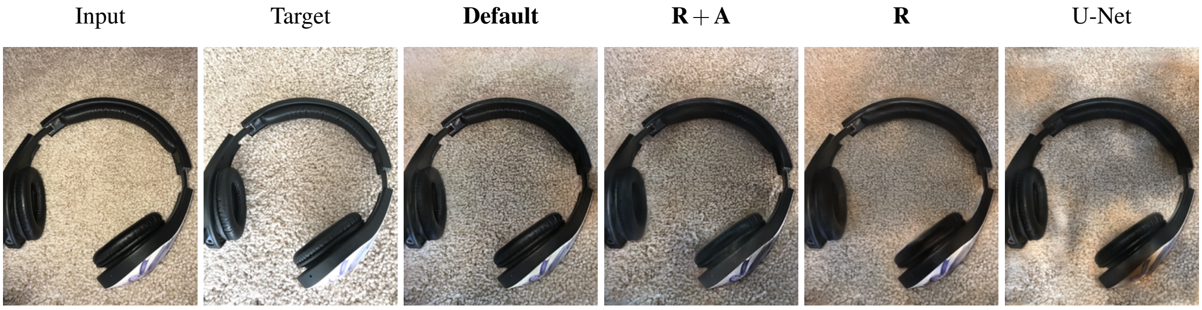


Figure 6: Qualitative comparison for each condition in our controlled experiments on the loss function, the attention map, and the network architecture.

We explore our qualitative results (Figure 6) for different loss functions, the attention map, and network architectures.

Loss Function. Table 2 reports the influence by using the adversarial loss. Condition 3 represents the same structure of the generator without considering the adversarial loss, i.e., just an encoder-decoder network, without a discriminator. This architecture presents blurred results comparing with our default model. In this case the reconstruction loss **R** is not enough to generate high-frequency details on their results, note the blurry image of the headphones (Figure 6). The adversarial loss **A** ensure a better quality due to the deep discriminator network, which classifies blurry results as fake. Condition 2 presents also blurry results; however, the output images present more uniform illumination due to the adversarial loss.

Attention Map. Condition 1, which represent our default model, present uniform illumination, and high-frequency details (note the sharpness on the headphone respect to the other conditions). Our attention mechanism guides the reconstruction and adversarial loss to obtain uniform illumination and also sharpness results with less artifacts. However, due to the robustness for overexposed areas and shadows, our model can not re-lighting dark areas with high-frequency details. We believe that a better formulation of the attention mechanism could address this problem.

Network Architecture. As we report in Table 2, we perform the well known U-Net (Ronneberger et al., 2015) architecture in condition 4. We adopt the model proposed by (Chen et al., 2018) for enhancing extreme low light images, and train it from scratch. U-Net present blurry output images and also non-uniform illumination. Our default model, which uses transfer learning, performs better quantitative and qualitative results. We believe this is due to the few samples in the training set.

5 CONCLUSIONS

Ambient lighting generation is a challenging problem, even more on flash images under low light conditions. Shadows on the flash image have to be removed, overexposed areas should be reconstructed, and ambient shadows must be synthesized as a part of the simulation of an ambient light source. In this paper, we propose a model with a guided reconstruction loss for normalizing the illumination and a guided adversarial loss to model high-frequency illumination details on flash images. Our results show that our guided mechanism estimated high-frequency details without introducing visual artifacts in our synthetic ambient images. The guided adversarial loss also produces more realistic ambient illumination on flash images than the state-of-the-art methods. Our current results are promising, nonetheless, there are cases where our model fails such as: restoring overexposed areas, normalizing the lighting for flash images on extreme low light conditions, and sideways shadow removal on flash images (see Figure 4). We believe that a more dedicated approach on the adversarial loss would be useful to address these issues.

Other methods based on intrinsic image decomposition (Shen et al., 2013) would be also useful by recovering the albedo (reflectance) and shading of the flash image, then, modifying directly the shading component to obtain the ambient image. As we show on this article, some cases need a more dedicated treatment. We aim to further study these cases and evaluate new techniques to improve the ambient lighting generation for flash images in such situations.

ACKNOWLEDGEMENTS

This work was supported by grant 234-2015-FONDECYT (Master Program) from Cienciactiva of the National Council for Science, Technology and

Technological Innovation (CONCYTEC-PERU). I thank all the people who directly or indirectly helped me with this work.

REFERENCES

Agrawal, A., Raskar, R., Nayar, S. K., and Li, Y. (2005). Removing photography artifacts using gradient projection and flash-exposure sampling. *ACM Trans. Graph.*, 24(3):828–835.

Aksoy, Y., Kim, C., Kellnhofer, P., Paris, S., Elgharib, M., Pollefeys, M., and Matusik, W. (2018). A dataset of flash and ambient illumination pairs from the crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 634–649.

Capece, N., Banterle, F., Cignoni, P., Ganovelli, F., Scopigno, R., and Erra, U. (2019). Deepflash: Turning a flash selfie into a studio portrait. *Signal Processing: Image Communication*, 77:28 – 39.

Chen, C., Chen, Q., Xu, J., and Koltun, V. (2018). Learning to see in the dark. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Eisemann, E. and Durand, F. (2004). Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph.*, 23(3):673–678.

Fu, X., Zeng, D., Huang, Y., Zhang, X.-P., and Ding, X. (2016). A weighted variational model for simultaneous reflectance and illumination estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Guo, X., Li, Y., and Ling, H. (2017). Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In Xianghua Xie, M. W. J. and Tam, G. K. L., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press.

Petschnigg, G., Szeliski, R., Agrawala, M., Cohen, M., Hoppe, H., and Toyama, K. (2004). Digital photography with flash and no-flash image pairs. *ACM Trans. Graph.*, 23(3):664–672.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Shen, L., Yeo, C., and Hua, B. (2013). Intrinsic image decomposition using a sparse representation of reflectance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2904–2915.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.