FlashGAN: Generating Ambient Images from Flash Photographs

Abdul Wasi[1], Iktaj Singh Bhinder[1], O. Jeba Shiney[1],
Mahesh Krishnananda Prabhu[2], and L. Ramesh Kumar[2](✉)

[1] Chandigarh University, Mohali, India
[2] Samsung Research Institute, Bangalore, India
{mahesh.kp,ram.kumar}@samsung.com

Abstract. Mobile Cameras capture images deftly in scenarios with ample light and can meticulously highlight even the finest detail from the visible spectrum. However, they perform poorly in low-light setups owing to their sensor size, and so, a flash gets triggered to capture the image better. Photographs taken using a flashlight have artefacts like atypical skin tone, sharp shadow, non-uniform illumination, and specular highlights. This work proposes a conditional generative adversarial network (cGAN) to generate ambient images with uniform illumination from the flash photographs and mitigate other artefacts introduced by the triggered flash. The proposed architecture's generator has a VGG-16 inspired encoder at its core, pipelined with a decoder. A discriminator is employed to classify patches from each image as real or generated and penalize the network accordingly. Experimental results demonstrate that the proposed architecture significantly outperforms the current state-of-the-art, performing even better on facial images with homogenous backgrounds.

Keywords: Conditional Generative Adversarial Network (cGAN) · Generator · VGG-16 · Encoder · Decoder · Discriminator

1 Introduction

Since the advent of cameras on mobile phones, there has been an increasing shift in capturing images using them. Attributable to the limited size of their sensors, the resolution and quality of the images captured using mobile cameras are relatively low when compared to professional DSLR cameras. Despite this, these cameras are able to capture the finest details of the objects in evenly-distributed lighting conditions. However, due to hardware constraints, the camera lens and aperture are relatively small and using them to capture images in low-light conditions results in poor quality. As a result, these cameras employ a flashlight to compensate for the low-light conditions. Such photographs, captured using a flash have a number of anomalies induced in them. These include an even and

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
D. Gupta et al. (Eds.): CVIP 2022, CCIS 1776, pp. 129–140, 2023.
https://doi.org/10.1007/978-3-031-31407-0_10

unnatural skin tone, shadows that are more sharp than normal, areas overexposed to a flashlight, and specular highlights resulting in a degenerated image [1]. Consequently, reconstructing such images into ambient ones poses a challenge. Prior work done to address this problem involved training a convolutional neural network (CNN) to generate uniformly illuminated portraits from flash images [2]. However, these models have been trained in studio environment with a homogeneous background. As a result, they fail to perform on random day-to-day images captured with flash in uncontrolled environments. Also, even though they perform well under control setups, they fail to successfully reconstruct certain features like the hue of the face and also, have a low peak signal-to-noise ratio.

The dataset used in this research consists of pairs of flash and no-flash images of human faces and an array of other objects captured in an indoor setup with diverse backgrounds, generally 0.5–1.0 s apart, with the source of illumination being indoor lights [3]. The absence of natural light source helps in proper flash exposure of the input images. This work seeks to investigate the prospect of taking the images captured with a flash and using a conditional generative adversarial network (cGAN) to convert them into ambient images that seem to be captured using a uniformly distributed lighting condition. The generator takes a flash image as an input and reconstructs the output image without the flash artifacts. The discriminator takes pairs of the input and target image as inputs and tries to infer if it is the ground truth or a generated image, minimizing the loss function accordingly. Finally, it learns an optimal function to map an input flash image to the desired output image. The output image is then subtracted from the unfiltered flash image to get the uniformly illuminated ambient image.

2 Related Work

This section provides a concise overview of the literature relevant to our problem. Ronneberger *et al.* [4] demonstrate how an encoder-decoder based architecture, trained on a few input images helps generate a corresponding segmented image in the output. The encoder here converts the input images to a lower-dimensional latent space representation from which the decoder generates an output. Isola *et al.* [5] propose a cGAN inspired generic solution to problems where the output image is generated from a paired input image. The authors employ a U-Net and a Markovian discriminator to generate the output images while trying to minimize the GAN loss and the mean absolute error between the target and generated image. Liu *et al.* [6] use coupled GANs based framework for image-to-image translation. Their model is essentially an unsupervised one and performs well on animal and facial image translation, among others. Pertinent to mention is that most of these methods use pairs of aligned images for training purposes. Zhu *et al.* [7] attempted to solve the problem by presenting an architecture that tries to learn a mapping from a source image to the target without depending on paired examples. This method gives good results in photo enhancement, style transfer, etc. Capece *et al.* [2] propose an encoder-decoder based CNN, to be

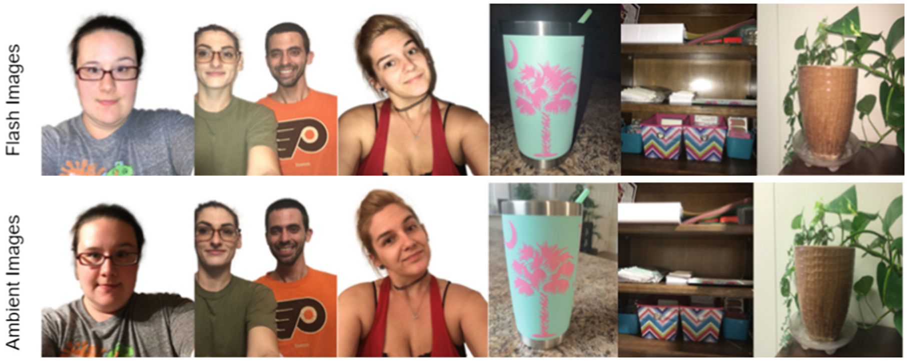

Fig. 1. Samples of flash and ambient images from the dataset. For pairs of images with faces, the foreground is extracted using MODNet [8].

used for translating pictures captured using flash into portraits with uniform illumination. It uses a dataset of human faces captured in a studio, with and without flash. The encoder part of this model uses the weights of a pre-trained VGG-16 which helps in extracting low-dimensional features from the image. The decoder is symmetric to the encoder and it up-samples the features to generate the output image. In order to minimize the Spatial Information loss, the model further uses skip connections between encoder and decoder layers. Although this architecture mitigates the anomalies introduced by a flash, it fails to regenerate the actual skin tone of the subject. Chávez *et al.* [1] use a conditional GAN to generate uniformly illuminated images from flash ones. Although they are able to generate somewhat uniform skin texture in the output and reduce the effect of flash, their model gives a less score on the structural similarity index measure. This indicates that even though the model removes the abnormalities that come from flash and generate a realistic skin tone, the output image has features that are divergent from normal. Inspired by the above-mentioned work, we propose a conditional GAN which fixes the shortcomings and generates the output with higher accuracy.

3 Methodology

3.1 Dataset

The dataset used in this research is acquired from the Flash and Ambient Illuminations Dataset (FAID) [3]. It consists of 2775 pairs of properly aligned images of People, Shelves, Plants, Toys, Rooms, and Objects captured with and without using a flash, usually 0.5 to 1 s apart (Fig. 1). The images are captured in an artificially illuminated indoor set up so as to give them a proper flash exposure. The images in the FAID dataset are resized to 256 * 256 pixels and split into two sets: the one with facial images and the other with the rest. The dataset of facial images is further reduced to 275 images, removing images with less light exposure and misaligned flash-no-flash pairs. What follows is the foreground extraction on such images using MODNet [8]. Afterward, the dataset is augmented to generate a total of 1100 images, 935 of which are used to train the network and the remaining 165 to test it. The rationale behind removing the background from images of people is to be able to compare the accuracy with the previous work that used a uniform studio-like background with such images. The dataset with the rest of the pictures is used as such, with 300 of its 2000 images used for testing the model and the rest to train it.

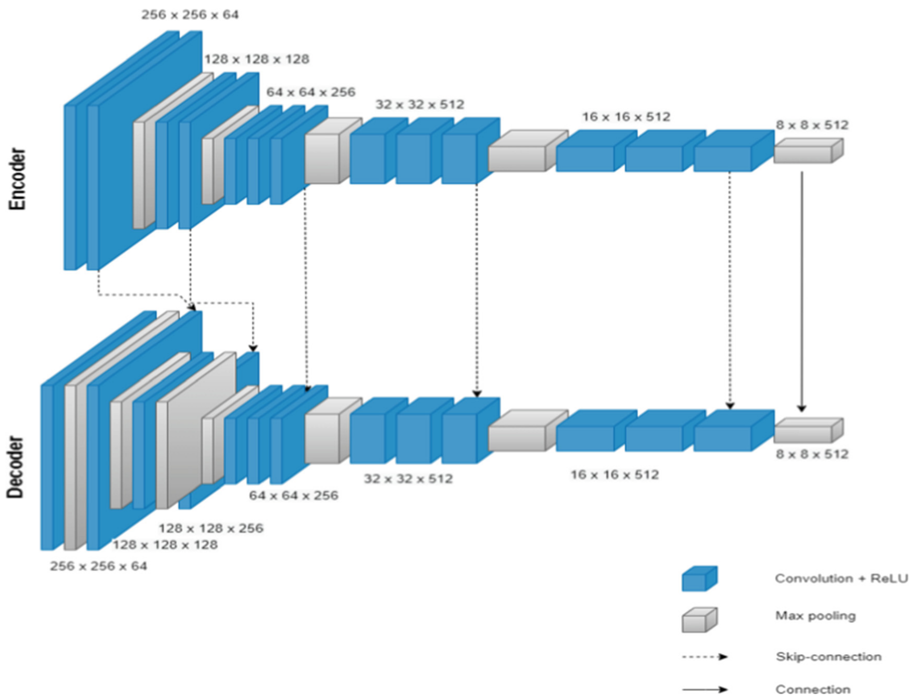

Fig. 2. The encoder-decoder architecture of the generator. The encoder takes a $256 \times 256 \times 3$ flash filtered image $Bil(I_f)$ as an input, the target being $Bil(I_f)$ - $Bil(I_a)$ normalized in the range $[0, 1]$. It gives I_o as output which is then denormalized in $[-1, 1]$.

3.2 Conditional GAN

A Conditional GAN [9] aids the creation of particular kind of images. It consists of a Generator and a Discriminator, both of which are fully convolution networks. In addition to a latent vector, a class label concatenated to it is provided as an input to the generator. This class label or data from other modalities directs the generator in terms of the image it is expected to generate. In image-to-image translation, the encoder-decoder architecture of the generator allows it to extract high-level features from an input image and construct the output close to the expected one. Afterward, the discriminator takes pairs of input and target images as inputs and tries to determine whether the target image is real or not while trying to minimize the discriminator loss. Here, fake corresponds to the image generated by the generator. The purpose served by a cGAN can be put as follows:

$$\eta_{cGAN}(\theta_G, \theta_D) = T_{a,b}[\log \theta_D(a, b)] + T_{a,c}[\log(1 - \theta_D(a, \theta_G(a, c)))] \tag{1}$$

Here, the generator θ_G works towards minimizing the objective function while the discriminator θ_D tries to maximize it i.e.

$$\theta_G^{res} = \arg \min_{\theta_G} \max_{\theta_D} \eta_{cGAN}(\theta_G, \theta_D) \tag{2}$$

3.3 Training

Problem Encoding. The proposed network may be used to tackle the problem in a variety of ways, the most basic of which is a model that takes a flash image I_f as an input and tries to generate its uniformly illuminated ambient equivalent I_a. The discriminator can then try to distinguish it from the expected output and learn to improvise the same. However, such a setup exhibits a decreased efficiency while learning to decouple details that have a sharp contrast from the surrounding features. The resultant image has visible artifacts and is faintly blurred. In this research, a bilateral filter [10] is used to address this problem. It is essentially a non-linear filter that smoothens an image, doing so without altering the pixel composition of the sharp image edges. Mathematically, the bilateral filter $F_{Bil}[.]$ is defined as:

$$F_{Bil}[Im]_x = \frac{1}{W_x} \sum_{y \epsilon S} K_{\sigma_s}(||x - y||) K_{\sigma_r}(|Im_x - Im_y|) Im_y \tag{3}$$

where the weighted pixel sum is ensured by W_x given as:

$$W_x = \sum_{y \epsilon S} G_{\sigma_s}(||x - y||) G_{\sigma_r}(|Im_x - Im_y|) \tag{4}$$

For an Image Im, the values of σ_s and σ_r signify the amount of filtering and G_{σ_s} and G_{σ_r} are Gaussian weightings for spatial and range intensity respectively and Im_θ signifies the intensity at pixel θ.

The result is an image whose high-frequency features are replaced with the spatially weighted average of the intensity of its surrounding pixels. The input to the network used in this work is a bilateral flash image $Bil(I_f)$, the target being the difference between the bilateral flash image $Bil(I_f)$ and the filtered no-flash image $Bil(I_a)$, normalized in the range [0, 1] (Fig. 2).

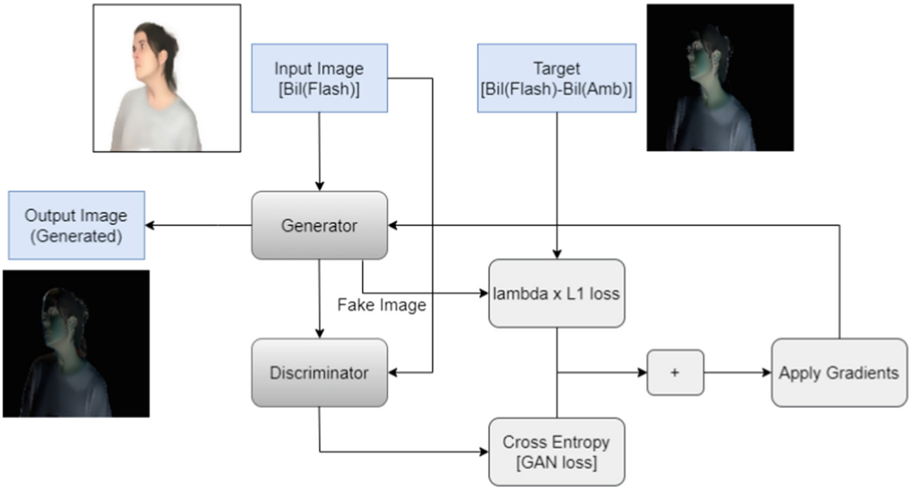

Fig. 3. Training of a generator. Here, the generator output I_o are the generated (fake) images and the L1 loss tries to minimize the difference between I_o and the target $Bil(I_f)$ - $Bil(I_a)$. The discriminator gets as input pairs of the expected and generated images ($Bil(I_f)$ - $Bil(I_a)$ and I_o), both labelled as real and uses the sigmoid cross entropy to distinguish between the two. The gradient tries to minimize both the errors as shown in Eq. (7).

Generator. The generator [9] in this work is essentially an encoder-decoder [11], with its encoder a VGG-16 network [12] and the decoder constructed accordingly. A VGG-16 is composed of 16 layers, the first 13 being the convolution layers that learn features from the input image while periodically downsampling it so that each pixel represents a larger input image context. The rest of the layers are fully connected and primarily, have a role in image classification. Here, the VGG-16 is pre-trained on the ImageNet dataset. We retain the 13 convolution layers of this network, whose weights have been updated in a way that they can classify images with high accuracy. It is called transfer learning. The activation function used is ReLU with the kernels of size 3×3 being used to pool features and a stride value of 1.

After a further convolution operation, a decoder almost symmetric to the encoder is developed. The input to the decoder is essentially this output from the last convolution operation. Here, the aim is to reconstruct the output I_o similar to $Bil(I_f)$ - $Bil(I_a)$. So, the data is upsampled in a way so that at each step, it is symmetric to the corresponding encoding layer. Also, batch normalization is performed so that the output at each layer is a valid input to the next. To make the gradient non-zero, the activation tensor thus obtained goes through LeakyReLU [13], a non-saturating nonlinear activation function. Lastly, skip connections are introduced to address the degradation and vanishing gradient problems as the proposed network has many layers. Pertinent to mention is that the Generator was trained on pairs of images to map its performance without a discriminator, the output of which has been mentioned in the results section (Fig. 3).

Loss Function. This work uses the Adam Optimizer, which is an extension of the stochastic gradient descent to minimise the loss function. Another major objective is to minimize the difference between the low contrast frequencies of $Bil(I_f)$ and $Bil(I_a)$ which we expect to retrieve later. This objective function can be put as:

$$O(p,t) = \frac{4}{3N}\sum_{o}((t_o - I_o) + E|I_o - t_o|)^2 \tag{5}$$

where

$$t_o = Bil(I_f) - Bil(I_a) \tag{6}$$

and I_o is the predicted output. This output is then denormalized in the range $[-1, 1]$ and later subtracted from I_f to get the reconstructed ambient image.

While trying to minimize the model loss, using both the mean absolute error $L1$ and the Conditional GAN loss in the ratio of 100:1 gives reasonably good results. Therefore, the combined loss function is given by:

$$\theta_{FlashGAN} = \arg\min_{\theta_G}\max_{\theta_D}\eta_{cGAN}(\theta_G, \theta_D) + \lambda\eta_{L1}(\theta_G) \tag{7}$$

where $\lambda = 100$ and $\eta_{L1}(\theta_G)$ is the L1 loss given as:

$$\eta_{L1}(\theta_G) = \mathrm{T_{a,b,c}[||b - \theta_G(a,b)||_1]} \tag{8}$$

Discriminator. A discriminator [14] tries to distinguish an image developed by the generator from the ground truth expected in the output. In this paper, a PatchGAN [15] has been used as a discriminator. As opposed to classifying an entire image, a PatchGAN discriminator convolves on an image and attempts to determine whether each NxN patch in a picture is real or not. In the output, it averages all such responses. Each activation in the output layer is a projection of an area from the input image described by the receptive field [16] R_f. A 70×70 PatchGAN is employed that convolves over each 70×70 patch of the input image of size $256 \times 256 \times 3$. This PatchGAN produces sharp outputs, both spatially and in color. All the ReLUs used are leaky with a slope of 0.2. The given discriminators architecture is:

$$C64 - C128 - C256 - C512$$

where the receptive field increases from four in the output convolution layer to seventy in the first one (C64). This receptive field maps an area from the input image to a final activation.

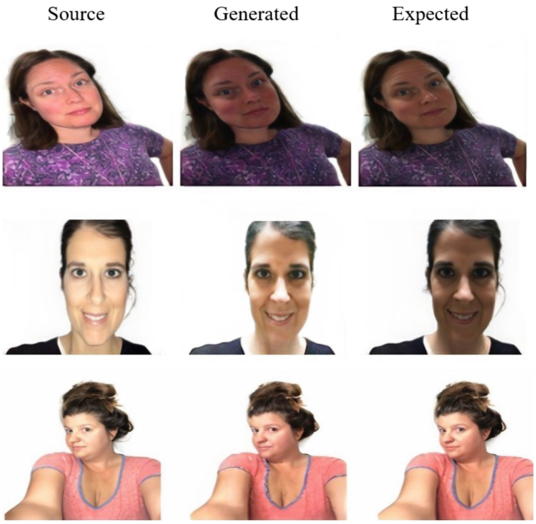

Fig. 4. Results on Segmented facial image dataset

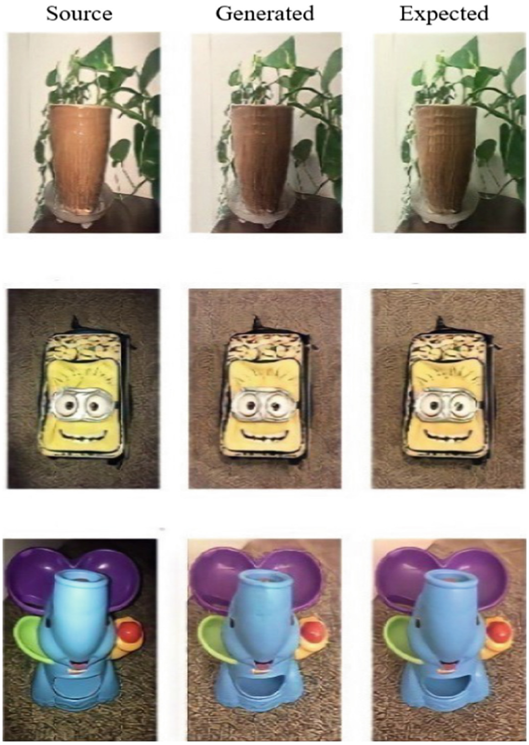

Fig. 5. Results on the generic dataset

4 Results

FlashGAN was trained on pairs of images of resolution 256 × 256 on an Nvidia GeForce GTX 1050 Ti GPU for about 17 h. At this point, the value of Structural Similarity Index (SSIM) [17] was high indicating that the generated image is highly similar to the ground truth.

Figures 4, 5 and 6 show the output generated from the test images after the training was stopped. The generated images here are the FlashGAN outputs denormalized in the range $[-1, 1]$ and then subtracted from I_f. SSIM and Peak signal-to-noise ratio (PSNR) [18] are the metrics used to evaluate the efficiency of our model. SSIM measures the similarity between the ground truth I_a and the generated image denormalized and subtracted from the flash image $I_f - I_o$.

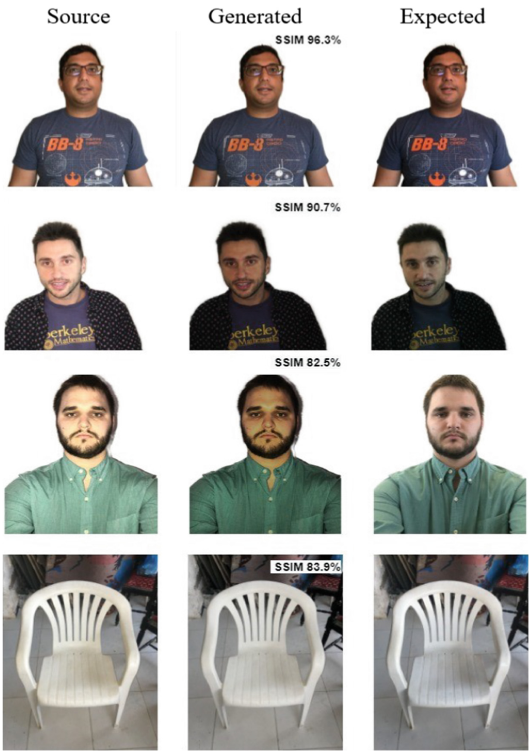

Fig. 6. Generated images along with their SSIM.

Both the segmented and the generic dataset were evaluated individually using the given metrics. The model performed equally well on images with multiple faces. The average SSIM and PSNR value on a sample of 70 images from the test set of facial images is 0.951 and 29.37 dB respectively. On a similar sample for the generic images, these values are 0.926 and 23.18 dB. A high value of PSNR signifies a better-reconstructed output. The results also indicate the proposed model is significantly better than the state-of-the-art as shown in Table 1. In this table, the SSIM and PSNR for FlashGAN is an average on a mixed sample of facial and generic images. Also, when the generator (encoder-decoder) was trained on the same sample, the SSIM and PSNR values obtained were 0.894 and 21.89 dB indicating that pairing it with a discriminator to make an adversarial network significantly improves the overall performance (Fig. 7).

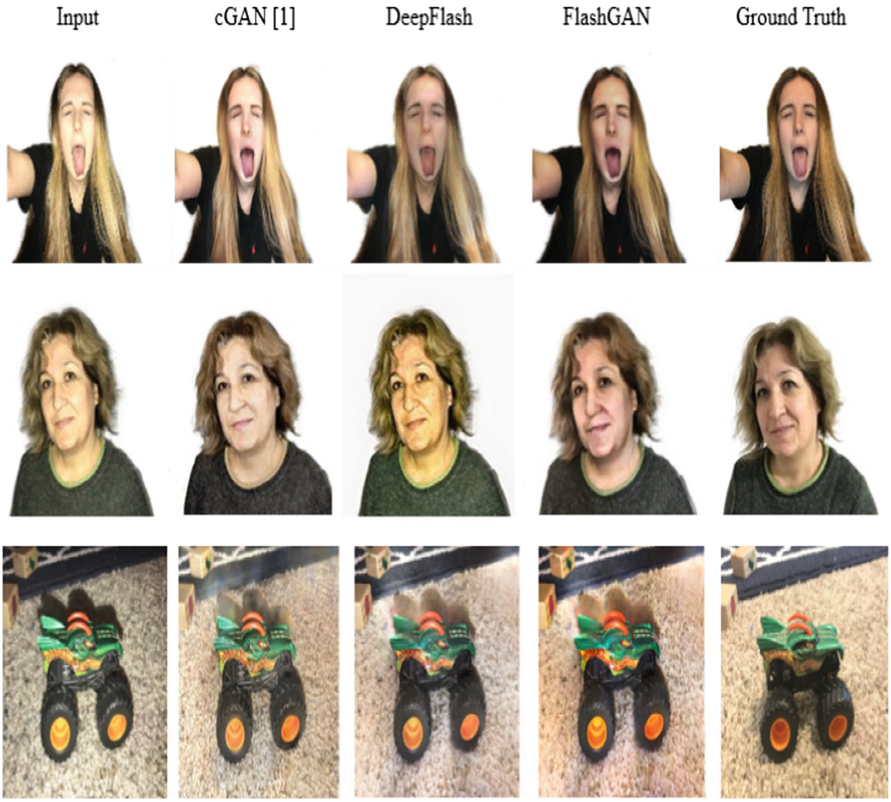

Fig. 7. A comparison between Guided cGAN, DeepFlash and the proposed network (FlashGAN).

Table 1. Mean SSIM and PSNR of DeepFlash [2], guided cGAN by J Chávez [1] and FlashGAN (proposed network).

Method	SSIM	PSNR
DeepFlash	0.8878	20.58 dB
Guided cGAN	0.684	15.67 dB
FlashGAN	0.937	25.14 dB

5 Conclusion and Future Scope

The results achieved by the proposed architecture suggest that conditional GANs outperform the other methods of image-to-image translation. In this work, it was pivotal in reconstructing the overexposed areas, generating the exact skin tone, and removing artifacts from the flash images with high efficiency.

Despite this, the models' accuracy dropped when it was trained without applying a bilateral filter on the input and target images. Filtering the images before they are passed through FlashGAN results in increased complexity. The attempt to reconstruct portrait images with heterogeneous background resulted in a decreased accuracy too. Also, an attempt to generate the ambient image as an output of the decoder (as opposed to I_o, the actual output) resulted in a decreased accuracy. In some cases, a low SSIM score was attributed to misaligned images. Models to address these problems can be part of future work. The model must be further optimized to remove or reconstruct the background without the shadow artifacts generated because of the flash light.

References

1. Chávez, J., Mora, R., Cayllahua-Cahuina, E.: Ambient lighting generation for flash images with guided conditional adversarial networks. arXiv preprint arXiv:1912.08813 (2019)
2. Capece, N., Banterle, F., Cignoni, P., Ganovelli, F., Scopigno, R., Erra, U.: DeepFlash: turning a flash selfie into a studio portrait. Sig. Process. Image Commun. **77**, 28–39 (2019)
3. Aksoy, Y., et al.: A dataset of flash and ambient illumination pairs from the crowd. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 644–660. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_39
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
5. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
6. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

7. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
8. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: MODNet: real-time trimap-free portrait matting via objective decomposition. In: AAAI (2022)
9. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
10. Elad, M.: On the origin of the bilateral filter and ways to improve it. IEEE Trans. Image Process. **11**(10), 1141–1151 (2002)
11. Zhou, S., Nie, D., Adeli, E., Yin, J., Lian, J., Shen, D.: High-resolution encoder-decoder networks for low-contrast medical image segmentation. IEEE Trans. Image Process. **29**, 461–475 (2019)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
13. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
14. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
15. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 702–716. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_43
16. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
17. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
18. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: 2010 20th International Conference on Pattern Recognition, pp. 2366–2369. IEEE, August 2010