Realistic Bokeh effect using a generative modeling approach

Shilpa Singh
Indian Institute of Science
shilpasingh@iisc.ac.in

Sourabh Gothe
Indian Institute of Science
sourabhgothe@iisc.ac.in

Abstract

In this paper, we attempt to create a shallow depth of field image from a single, sharp image, a problem known as monocular bokeh synthesis. Due to the limitations of the mobile camera's aperture, this effect cannot be caught directly in mobile cameras in the same way it can in DSLR cameras. In the context of this, we explore multiple approaches for producing convincing monocular bokeh from a single picture. We have followed two approaches, one is depth estimation based, and another is a generative modeling approach. For the first approach, we use a convolutional neural network with transfer learning to compute a high-resolution depth map from a single RGB picture. Following a traditional encoder-decoder architecture, we initialize our encoder with features derived from high-performing pre-trained networks that result in more accurate outcomes. We experimented with image-to-image translation techniques, such as pix2pix and CycleGAN, for the second approach. We also investigated BGGAN, which generates realistic bokeh effects for photos with a high resolution.

1. Introduction

Bokeh, the blurring of the out-of-focus areas of a photograph created by a camera lens, is often regarded as one of the most significant aesthetic standards in photography. As a result of the inherent limits of mobile camera hardware, these cameras are unable to generate a strong bokeh effect naturally; hence, computer vision-based techniques are frequently utilized to achieve this effect on mobile devices. This concept grew quite popular over the last few years, and one of the most typical remedies [22] [23] [26] for this problem consists of segmenting out the primary object of interest on the shot and then blurring the backdrop. The predicted depth map [11] [25] can be acquired by the use of the parallax effect or stereo vision [3] [4]. This has led to the suggestion of an additional method for this task: blurring the picture based on the predicted depth map. Finally, in [13], a deep learning-based method and corresponding EBB! bokeh effect rendering dataset was proposed. The dataset prioritises wide-scene synthetic bokeh effect rendering. The objects in the depth-of-field region of this dataset range from portraits to traffic signs to automobiles to flowers and beyond. The bokeh effect is challenging to capture, but the rewards are worth the effort.

The EBB! Dataset has been the testing ground for a wide variety of approaches [13] [6] [14]. Many of these techniques, such as those used to locate prominent regions and calculate depth, rely on a priori information. In [6], they suggest an approach using the well-known depth estimation algorithm [19], which has excellent performance when estimating depth from a single picture. Then, a practical approach for Gaussian blur kernel application to the out-of-focus areas is developed. In our first approach, we followed the same strategy with the initial intention of developing a model that is less complicated and more amenable to training and future alterations [1]. To do this, we use transfer learning to re-purpose a high-performing pre-trained network built for image classification to serve as our deep features encoder. One significant benefit of this transfer learning-based strategy is that it paves the way for a more decoupled design, where improvements in one area may be readily applied to the depth estimation.

Based on these techniques, we may deduce that a priori information, such as a salient area detection map or depth estimate map, can enhance the resultant image's visual impression. On the other hand, the priors are strongly dependent upon those approaches, thus, when they fail in some settings, the resulting synthetic picture will have unanticipated flaws. Among the many impressive features of Generative Adversarial Networks (GANs) [9], one of their most famous is their capacity to create increasingly realistic images while still fooling human observers. GANs have been used for a variety of image-to-image translation tasks in recent years, including deblurring [18], super-resolution [20], transferring styles [16], and generating product photos. We are motivated to think about bokeh as an aspect of image-to-image translation by these studies. Three generative modeling methods—pix2pix [15], cycleGAN [27], and BGGAN [21] have been the primary focus of our explorations.

In this paper, we explored depth estimation-based bokeh generation with gaussian blur masking and generative modeling-based bokeh effect generation techniques and reported the results.

2. Related Works

2.1. Depth Based Bokeh Generation

In 2014, Eigen et al. [8] were the first to apply a Convolutional Neural Network (CNN) to the problem of image-depth estimation in which two networks were run in parallel and subsequently merged; by 2015, this had been refined to involve three parallel forks [7]. To learn global and local characteristics, Kim's [17] technique uses a concurrent two-branch structure, with one branch focusing on the entire picture. In 2021, Chen et al. [5] suggested an ACAN network that had a multibranch parallel topology and an additional content-attention module. One way to estimate the depth of an image is through a GAN, wherein the generator creates a map, and the discriminator decides whether or not it is accurate. For the input image and depth map pair, the generator in the technique developed by Islam [13] et al. in 2021 makes two false depth maps, and the discriminator decides which of the three depth map candidates is the genuine image pair with the input picture.

2.2. Generative Methods for Bokeh Rendering

It has been demonstrated that generative models [9] can produce realistic pictures with accurate textures. However, it has been demonstrated that these models frequently experience Modal Collapse. Algorithms like WGAN [2], and WGAN-GP [10] have been developed to improve training stability and hence provide a solution to this problem.

Recent research on image translation problems, such as image deblurring [18], single-image super-resolution [20], semantic segmentation, and so on, has demonstrated the effectiveness of GAN-based architectures. It has been demonstrated that Conditional Adversarial Networks [15] can generalize to a wide variety of picture translation problems. Recently, this framework has also been used for monocular bokeh synthesis [21] in enhancing the bokeh rendering quality.

3. Method

In this section, we illustrate two approaches explored, depth-based and generative modeling based. We explain the datasets we used, followed by the architectures in each.

3.1. Bokeh using Depth-Estimation

3.1.1 Dataset

The NYU Depth v2 dataset [24] contains 640x480 photos and depth maps of a variety of interior situations recorded by researchers at the university. There are 120,000 samples used for training and 654,000 for testing. As a training set, we use a subset of 50k records. Predictions from our network have a resolution of 320 by 240, which is half that of the input. We use the full-resolution input photos and downsample the ground truth depths to 320 x 240 for training purposes. Note that even though some of the input image-depth map pairings have missing pixels as a result of distortion correction preprocessing, we do not crop any of them. To assess on the pre-defined center cropping, we upsample the depth map prediction from the whole test picture by a factor of 2 to match the ground truth resolution. During testing, we calculate the final result by averaging the predictions for each picture and their mirror images.

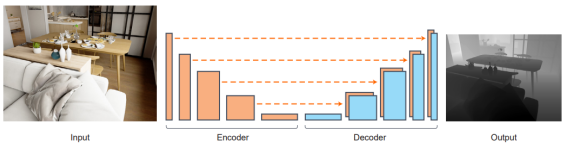

Figure 1. We employ an encoder-decoder architecture with skip connections. The encoder part is a pre-trained truncated DenseNet-169 [12]. The decoder is composed of basic blocks of convolutional layers applied on the concatenation of the 2× bilinear upsampling block in the encoder.

3.1.2 Architecture

Fig. 1 shows an overview of our encoder-decoder network for depth estimation. For our encoder, the input RGB image is encoded into a feature vector using the DenseNet-169 [12] network pre-trained on ImageNet. This vector is then fed to a successive series of up-sampling layers in order to construct the final depth map at half the input resolution. These upsampling layers and their associated skip connections form our decoder. For training our network, we define the loss L between y and $\hat{y}$ as the weighted sum of three loss functions

$$L(y, \hat{y}) = \lambda L_{depth}(y, \hat{y}) + L_{grad}(y, \hat{y}) + L_{SSIM}(y, \hat{y}) \quad (1)$$

The first loss term L_{depth} is the point-wise L1 loss defined on the depth values, and the second loss term L_{grad} is the L1 loss defined over the image gradient g of the depth image. Lastly, L_{SSIM} uses the Structural Similarity (SSIM) term, a commonly-used metric for image reconstruction tasks. It has been recently shown to be a good loss term for depth estimating CNNs.

3.2. Bokeh using Generative Modelling

3.2.1 Dataset

AIM 2020 Bokeh Effect Rendering Challenge has made available the EBB! dataset [13]. There are 5,000 pairs of shallow and wide depth-of-field images that make up the

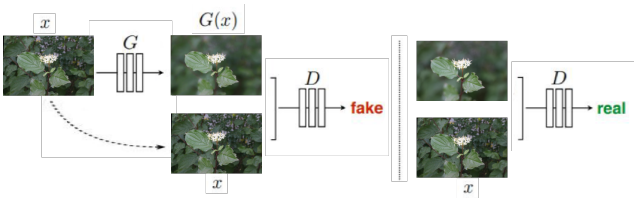

Figure 2. Training a conditional GAN to map edges→photo. The discriminator, D, learns to classify between fake (synthesized by the generator) and real bokeh, original tuples

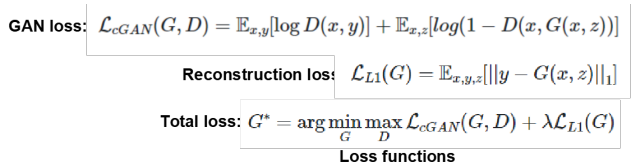

Figure 3. Loss functions in pix2pix architecture

dataset. In all, there are 4600 picture pairs in the training dataset, 200 in the validation dataset, and 100 in the test dataset. The non-bokeh image in each pair is shot at f/16, while the bokeh image is taken at f/1.8, the widest aperture possible. The photos are captured in a variety of places with automatic mode.

3.2.2 Pix2Pix

Generative Adversarial Networks (GANs) are generative models that learn a mapping, G: z -¿ y [9], between a random noise vector z and an output picture y. In contrast, conditional GANs learn a mapping from observed image x and random noise vector z, to y, G: x, z → y. The generator G is trained to produce outputs that cannot be distinguished from "real" images by an adversarially trained discriminator, D, which is trained to do as well as possible at detecting the generator's "fakes." As shown in Fig. 2 we adapt our generator and discriminator architectures. Both the generator and discriminator use modules of the form convolution-BatchNorm-ReLu. Specific loss functions are given in Fig. 3.

3.2.3 CycleGAN

The architecture is depicted in Figure 4. In principle, adversarial training has the potential to learn mappings G and F that provide outputs that are identically distributed as target domains Y and X, respectively. However, with large enough capacity, a network can map the same set of input images to any random permutation of images in the target domain, where any of the learned mappings can induce an output distribution that matches the target distribution. Thus, adversarial losses alone cannot guarantee that the learned function can map an individual input x_i to a desired output y_i. Here, the Cycle consistency loss helps along with regular GAN loss. All the loss functions are shown in Figure 5.

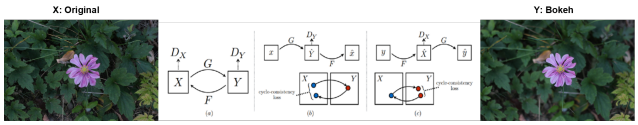

Figure 4. Training a conditional GAN to map edges→photo. The discriminator, D, learns to classify between fake (synthesized by the generator) and real bokeh, original tuples

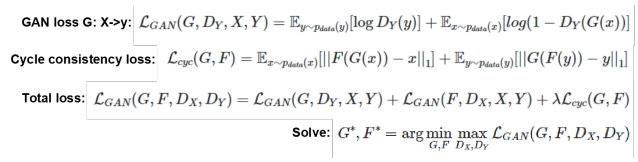

Figure 5. Loss functions in CycleGAN architecture

3.2.4 BGGAN

The Glass-Net and Multi-receptive-field Discriminator are combined to construct BGGAN. Glass-Net is an end-to-end network that takes an image as an input and produces the result with the bokeh effect. The name is given as Glass net as the network resembles a pair of glasses. it's a two-stage network, the first network learns the mapping of input image I(i.e., image without bokeh) and the residual R between input image I and ground truth O i.e, R = I - O, that infers I-R to be a rough bokeh result. The second stage of the Glass-Net refines the rough bokeh results to generate more realistic bokeh effects.

he loss function for fine-tuning Glass net consists of L_1 reconstruction loss, negative SSIM loss L_{SSIM}, The perceptual loss that computes the euclidean loss on feature maps of VGG19 L_{VGG} is given by below equation,

$$L_{VGG} = \frac{1}{HWC} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{C} \|F(G(I)_{i,j,k} - F(C_{i,j,k}\|_1 \quad (2)$$

where F(.) is the 34th layer's feature map of the VGG network which is pre-trained on ImageNet, $G(I_{i,j})$ indicates the image which is produced by Glass-Net, and C is the ground truth. And the adversarial loss L_{adv} for the glass-net generator is given as,

$$L_{adv} = -\frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} D(G(I)_{i,j}) \quad (3)$$

Here D(.) indicates the output of the discriminator. Finally, the four losses are incorporated into a hybrid loss for fine-tuning Glass-Net with appropriate weights, the hybrid loss is defined below the equation.

Figure 6. Qualitative results of Depth Based methods

$$L_{hybrid} = 0.5 \times L_1 + 0.05 \times L_{SSIM} + 0.1 \times L_{VGG} + L_{adv} \quad (4)$$

Here L_{adv} has a larger factor as adversarial training plays a major role in optimization.

4. Experiments and Results

We built our own test set from the EBB! training data since the ground truth for the EBB test set is not made available to the public. We picked 100 photos to not be used in the methods' training processes. All the experiments have been evaluated on the same set.

4.1. Depth Estimation based bokeh

Depth estimation is done on NYU dataset, Unet based model is employed for Depth estimation. The train accuracy obtained is around 80.42%, validation accuracy is around 80.47% and test accuracy obtained is around 76.01%. The trained depth model is used to generate depth map for EBB dataset and guassian blur is applied on the obtained depth map based on threshold values 0.2, 0.25 and 0.3. Figure 6 describes the qualitative results for the depth-based methods. We experimented with three different values for the focal plane which we call thresholds. It can be observed that a lower threshold value (0.2, 0.25) is making most of the part of the image blurred and 0.3 produces reasonably best results.

In spite of the fact that the PSNR values for lower thresholds are greater and on par with BG-GAN, we found that the bokeh effect was erroneously applied in the vast majority of situations when inspected manually. We can set the threshold based on empirical results. However, it is challenging to generalize to the majority class of images, which does not

Figure 7. Qualitative results of Generative Modeling methods

occur in the case of our other approach.

4.2. Generative modeling based bokeh

We employed conditional Adversarial training to perform direct image-to-image translation to generate bokeh images from original images, for this purpose we experimented with two models CycleGAN and Pix2Pix.

4.2.1 Pix2Pix

To train Pix2pix on EBB! dataset, we created a dataset of a pair of images based on the original image and its bokeh version. We build the unet style-based generator model and patch GAN-based discriminator model for 70x70 patch size. We trained the discriminator and generator alternatively based on the loss functions given in the method section and post-training we evaluated PSNR and SSIM on generated bokeh images. for progressive and smoother training we employed model checkpointing every few epochs.

4.2.2 CycleGAN

To train CycleGAN on EBB! dataset, we employed similar architecture of generator and discriminator as in pix2pix. Datasets for original and bokeh images were created separately. we trained a generator and discriminator for the original to bokeh and also trained one more generator and discriminator for bokeh to the original based on loss functions given in the method section, post-training we generated bokeh images based on original images and evaluated PSNR and SSIM based on ground truth and generated images. for progressive and smoother training we employed model checkpointing every few epochs.

Method	PSNR	SSIM
Depth Estimation + Bokeh(0.2)	**29.086**	0.805
Depth Estimation + Bokeh(0.25)	29.106	0.790
Depth Estimation + Bokeh(0.3)	29.127	0.775
BGGAN	30.967	**0.897**
Pix2Pix	16.743	0.441
CycleGAN	21.545	0.741

Table 1. Quantitative Results on EBB Testset

Figure. 7 shows the qualitative results of all three generative models. It can be observed that pix2pix is capturing the saliency of the image appropriately, however it is masking with dark pixels instead of blur. Whereas CycleGAN is producing overall blurred images and may improve with more epochs of training. We Utilized pre-trained BGGAN and evaluated PSNR and SSIM. The combined quantitative results are shown in Table. 1, we observed that BGGAN has the best results among all and it was proven true when manually inspecting the bokeh effect on test images.

5. Conclusion

In this paper, we experimented with two approaches to solve the problem of bokeh generation. First, based on depth estimation we tried out with multiple focal planes and found the 0.3 threshold to be working better empirically. Second, we experimented with three different generative models pix2pix, cycleGAN, and BGGAN, and observed that BGGAN achieves the best PSNR and produces more realistic bokeh effects than other methods. However BGGAN mostly resembles CycleGAN architecture, we believe we can further improve cycleGAN and pix2pix to achieve similar results. We also observed that PSNR does not indicate the overall performance, as we saw that bokeh images generated from the depth method with 0.2 thresholds were erroneous even with high PSNR value.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. 1

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[3] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4474, 2015. 1

[4] Fausto Tapia Benavides, Andrey Ignatov, and Radu Timofte. Phonedepth: A dataset for monocular depth estimation on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3049–3056, 2022. 1

[5] Yuru Chen, Haitao Zhao, Zhengwei Hu, and Jingchao Peng. Attention-based context aggregation network for monocular depth estimation. *International Journal of Machine Learning and Cybernetics*, 12(6):1583–1596, 2021. 2

[6] Saikat Dutta. Depth-aware blending of smoothed images for bokeh effect generation. *Journal of Visual Communication and Image Representation*, 77:103089, 2021. 1

[7] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 2

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1, 2, 3

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 2

[11] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5413–5421, 2016. 1

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[13] Andrey Ignatov, Jagruti Patel, and Radu Timofte. Rendering natural camera bokeh effect with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 418–419, 2020. 1, 2

[14] Andrey Ignatov, Jagruti Patel, Radu Timofte, Bolun Zheng, Xin Ye, Li Huang, Xiang Tian, Saikat Dutta, Kuldeep Purohit, Praveen Kandula, et al. Aim 2019 challenge on bokeh effect synthesis: Methods and results. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3591–3598. IEEE, 2019. 1

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[17] Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via in-

tegration of global and local predictions. *IEEE transactions on Image Processing*, 27(8):4131–4144, 2018. 2

[18] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 1, 2

[19] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[20] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7769–7778, 2020. 1, 2

[21] Ming Qian, Congyu Qiao, Jiamin Lin, Zhenyu Guo, Chenghua Li, Cong Leng, and Jian Cheng. Bggan: Bokeh-glass generative adversarial network for rendering realistic bokeh. In *European Conference on Computer Vision*, pages 229–244. Springer, 2020. 1, 2

[22] Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102. Wiley Online Library, 2016. 1

[23] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European conference on computer vision*, pages 92–107. Springer, 2016. 1

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 2

[25] Fisher Yu and David Gallup. 3d reconstruction from accidental motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3986–3993, 2014. 1

[26] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 297–305, 2017. 1

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1