Defocus Magnification using Conditional Adversarial Networks

by

Parikshit Sakurikar, Ishit Mehta, P J Narayanan

in

IEEE Winter Conf. on Applications of Computer Vision, 2019
(*WACV-2019*)

Hawaii, USA

Report No: IIIT/TR/2019/-1

Centre for Education Technology and Learning Science
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2019

Defocus Magnification using Conditional Adversarial Networks

Parikshit Sakurikar, Ishit Mehta and P. J. Narayanan
Center for Visual Information Technology, Kohli Center on Intelligent Systems,
International Institute of Information Technology, Hyderabad, India.
{parikshit.sakurikar@research.,ishit.mehta@research.,pjn@}iiit.ac.in

Figure 1: Left: A narrow-aperture portrait image [6] with low defocus in the background. Middle: The composite focus measure [24] computed over the portrait image (contrast enhanced for visualization). Right: A wide-aperture result produced using our method.

Abstract

Defocus magnification is the process of rendering a shallow depth-of-field in an image captured using a camera with a narrow aperture. Defocus magnification is a useful tool in photography for emphasis on the subject and for highlighting background bokeh. Estimating the per-pixel blur kernel or the depth-map of the scene followed by spatially-varying re-blurring is the standard approach to defocus magnification. We propose a single-step approach that directly converts a narrow-aperture image to a wide-aperture image. We use a conditional adversarial network trained on multi-aperture images created from light-fields. We use a novel loss term based on a composite focus measure to improve generalization and show high quality defocus magnification.

1. Introduction

An image captured with a finite-aperture camera is composed of focused and defocused scene points. Most portrait photographs are captured with an intent to achieve a shallow depth-of-field around the focused subject. Shallow depths-of-field however cannot be achieved using cameras that have narrow-aperture lenses. Defocus magnification is the task of selectively increasing the amount of defocus blur in order to simulate an image with a shallow depth-of-field.

Defocus magnification [2] is a post-capture image processing operation that has been well studied in computer vision. The standard approach for blur magnification is to first estimate the amount of blur at each pixel and then amplify it. Estimating the defocus blur at each pixel is analogous to estimating its relative depth in the scene, as the amount of defocus at a pixel is a direct consequence of its 3D location. Estimating the per-pixel defocus blur using a single image is a challenging task and usually works only at image locations with intensity gradients. Existing methods [2, 31, 37, 36] estimate the blur at the high-gradient pixels and propagate it to the other locations in the image. Errors in blur estimation manifest as artifacts in the synthesized wide-aperture image.

In this paper, we present an end-to-end deep neural network that takes an input image with focused and defocused pixels and selectively magnifies the blur at the defocused pixels, thereby simulating a shallow depth-of-field. We train this blur magnification network using wide aperture images created from a light-field dataset [29] and use a combination of adversarial training, perceptual loss and a novel loss term based on a focus measure to improve generalization. The end-to-end approach and the novel loss term are the main contributions of this work. We provide quantitative

and qualitative results and demonstrate high quality defocus magnification on several images. Figure 1 is an example of our single-step defocus magnification on a generic portrait image. Note that none of the images in our training data included human subjects.

2. Related Work

Defocus magnification of portrait photos is a well studied problem in computer vision. With the abundance of mobile devices equipped with high quality narrow-aperture cameras, blur magnification has become a popular tool in photo-sharing and social media applications. Several hardware modifications have also been implemented in mobile cameras to estimate per-pixel depth specifically for blur magnification.

Bae and Durand [2] proposed defocus magnification as a two-stage task of defocus estimation followed by amplification. The bluriness (defocus map) is estimated at intensity edges by fitting the best Gaussian blur kernel to pixel intensities. The blur map is then refined using bilateral filtering and homogeneously propagated to all other pixels in the image. Tai and Brown [31] propose a local contrast prior for defocus estimation which models defocus blur as the ratio between the gradient and the contrast of intensities surrounding the pixel. This is based on the observation that with increasing blur, local gradients become smaller than the local contrast due to smoothing. The estimated defocus blur is propagated through the image using an MRF optimization. Methods that are based on defocus estimation and amplification share the benefit that the depth-map and the texture-less regions in the scene need not be estimated explicitly.

Zhou and Sim [37] estimate the per-pixel defocus in an image by re-blurring the image with known blur radii and compute the unknown blur using the gradient ratio of the re-blurred images. They estimate sparse blur radii at intensity edges and blur values are propagated through the image using the edge-aware matting Laplacian approach of [14]. Zhu *et al*. [36] use blur-spectrum fitting to estimate a probability distribution over the scale of the point-spread-function at all pixels which is optimized using gradient-aware smoothness terms. Chen *et al*. [4] magnify the defocus in portrait images with foreground enhancement using a face prior for foreground detection. Recent work has also focused on estimating small blurs in images captured by point-and-shoot cameras. Shi *et al*. [25] estimate small scale defocus blurs using non-parametric matching and edgelet primitives which model intensity edges with varying directions, curvatures and scales. Shi *et al*. [27] estimate just-noticeable-blur using sparse reconstruction statistics based on the observation that clear and slightly blurred areas are composed of visually different dictionary patches.

Deep neural networks have also been used for estimating the per-pixel defocus kernel in a single image. Ma *et al*. [17] use a fully convolutional neural network trained on the discriminative blur detection dataset of Shi *et al*. [26]. They show state-of-the-art blur detection and estimation for several applications including defocus magnification. Park *et al*. [21] use a combination of defocus measures and deep features in order to estimate the defocus at all pixels in an image. The sparse defocus map obtained using a combined defocus measure is propagated to the full image using the matting Laplacian [14] approach.

Defocus magnification has also been shown in conjunction with methods that compute depth-maps from single images. Namboodiri and Chaudhuri [19] solve a stabilized reverse heat equation to estimate the depth-map of the scene characterized by the amount of defocus at each pixel. Barron *et al*. [3] propose a stereo based approach with disparity inference in bilateral space for high quality depth maps and consequently render shallow depth-of-field images. Srinivasan *et al*. [28] create a large dataset of aperture stacks to train a monocular depth-estimation network on images of flowers and plants. Defocus magnification is the ideally suited application for their technique. Several mobile devices are now equipped with dual cameras and primarily use stereo-based depth information to simulate a shallow depth-of-field. The Pixel 2 cameras [22, 33] use dual pixels on the sensor that generate stereo images at small baselines of $1mm$ which are useful for depth-aware defocus magnification. Our previous work [23] proposes an adversarial learning framework for scene refocusing. An image is refocused by first de-blurring it and then re-blurring it to the target focus position. We demonstrate post-capture control of the focus position in [23] and post-capture control of aperture size in this work. Contemporary to our work, Wang *et al*. [16] use deep-neural networks for post-capture control over the depth-of-field. They estimate scene depth from a single image and then apply a lens-blur network to simulate the target depth-of-field.

In this paper, we propose an end-to-end deep neural network for defocus magnification. We avoid the task of estimating the depth/defocus at each pixel and directly learn the mapping between a narrow-aperture image and a wide-aperture image. We train our neural network using aperture stacks created from a light-field dataset. We use a robust conditional adversarial network with perceptual loss as our base architecture and we implement an additional loss term based on a composite measure of focus which improves standalone performance and generalization. To the best of our knowledge, this is the first deep neural network for end-to-end defocus magnification.

3. Defocus Magnification

An image captured by an aperture-camera is a composition of light rays arriving from scene points at different

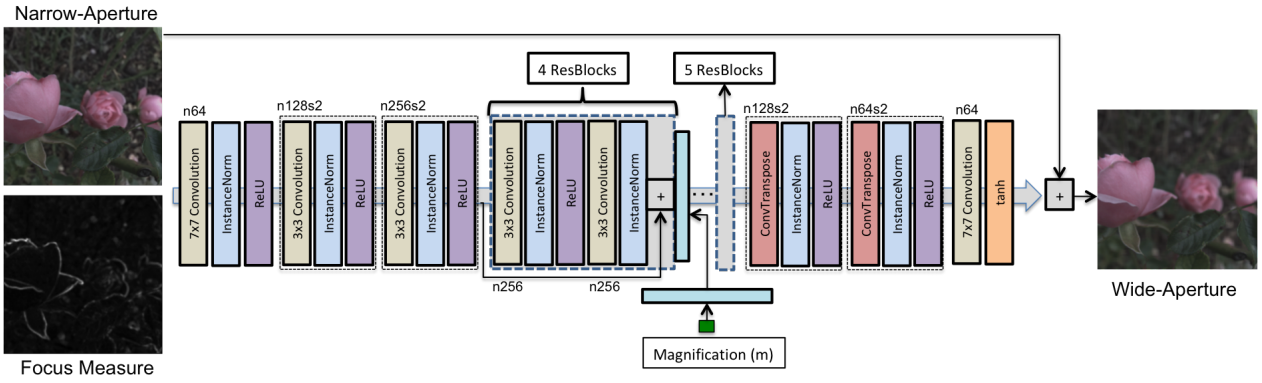

Figure 2: Our generator network $\mathcal{G}$ that produces a wide-aperture image from an input narrow-aperture image $\mathcal{I}$ along with its composite-focus-measure channel $f(\mathcal{I})$ and a magnification parameter m. The focus measure channel is appended to the RGB image to create an RBGF input to the network. The magnification parameter is converted to a 64×64 channel and appended after the fourth residual block. The structure of the network is explained in Section 3.1.

distances from the camera. Light rays travel through the body of the lens according to the geometry dictated by the camera settings and finally fall on the sensor. All the scene points at a particular depth in the scene converge at the same distance behind the lens and contribute as focused pixels if the sensor is located at this specific distance. Scene points at other depths either converge in front of or behind this sensor position and thereby contribute to the image as defocused pixels. An image created from pixels at different depths in the scene can be modeled as:

$$\mathcal{I}(x,y) = \int\int \mathcal{H}(x,y,\delta(x,y))\ \mathcal{R}(x,y)\ dxdy, \qquad (1)$$

where $\mathcal{R}$ is the radiance of the scene point corresponding to the pixel (x,y), $\mathcal{H}$ is the per-pixel defocus kernel, δ represents the size of the kernel and $\mathcal{I}$ is the captured image. The δ parameter represents the separation in depth between the in-focus sensor position of the pixel (x,y) and the current sensor position. Therefore, δ is negligible for in-focus pixels and large for defocused pixels. The defocus kernel $\mathcal{H}$ is typically modeled as a Gaussian with a spread corresponding to δ. In-focus pixels typically correspond to scene points at the same depth in the scene.

A finite region surrounding the in-focus plane, where the defocus blurs are small enough to be imperceptible is known as the depth-of-field of the image. The size of the depth-of-field is inversely proportional to the size of the aperture. On increasing the size of the aperture opening, the deviation in incoming light rays increases thereby causing a reduction in the depth-of-field. Small/shallow depth-of-field is a useful compositional tool in photography in which the subject is usually kept in focus and other background or foreground regions are defocused. This is widely used in portrait photography, selfies and object photography. Mobile and point-and-shoot cameras are moving towards smaller form-factors and size and thus have narrow apertures. Such cameras are unable to capture small/shallow depths-of-field.

Defocus magnification or aperture magnification is the process of identifying the defocused pixels in the image and selectively blurring them further in a depth-aware manner to simulate a shallow depth-of-field. Defocus magnification simulates an increase in the size of the aperture. Ideally, enlarging the aperture would lead to a reduction in size of the depth-of-field as more pixels from the focused part of the image would now be defocused. However, existing image-processing methods keep the focused pixels intact and blur the defocused pixels further, thus simulating more blur in the background without reducing the number of focused pixels. This simulates the blurring of a shallow depth-of-field without reducing the size of the depth-of-field. Therefore the technique is more correctly referred to as defocus magnification rather than aperture magnification. It can be noted however that if small blurs can be detected accurately even for the pixels that lie within the depth-of-field, a defocus magnification method will simulate geometrically accurate aperture magnification. Generating a wide-aperture image $\mathcal{I}'$ with blur magnification in the blurred regions can be modeled as:

$$\mathcal{I}'(x,y) = \int\int h(x,y,m*k(x,y))\ \mathcal{I}(x,y)\ dxdy. \quad (2)$$

Here, $\mathcal{I}$ is the captured narrow-aperture image, h is a re-blurring kernel with a size depending on $k(x,y)$, which is a defocus measure that encodes the amount of blur at a pixel (x,y) in $\mathcal{I}$, and m is a magnification parameter that scales the amount of re-blurring. The in-focus pixels in $\mathcal{I}$ correspond to $k=0$ and will not undergo any re-blurring. The pixels that are already defocused have $k>0$ and will be re-blurred by a kernel of finite size $m*k$.

In the following section we propose an end-to-end deep neural network that magnifies the defocus of an input narrow-aperture image. We train a conditional adversarial network and use a novel loss term. We learn a residual image to convert a small aperture image to a wide-aperture image with a shallow depth-of-field.

3.1. Network Architecture

Defocus magnification is a complex image filtering operation defined in Equation 2. In-focus pixels are expected to remain in-focus while the pixels that are defocused are to be blurred further using a depth-dependent blur kernel. The challenge in producing a shallow depth-of-field image in a single-step is that the network must learn an implicit representation of focus and re-blur the pixels according to the amount of focus at each pixel. We use a conditional adversarial network for this task, drawing inspiration from [13] and [23] that use adversarial learning for the tasks of image de-blurring and scene refocusing respectively. We broadly discuss adversarial learning and consequently define our network for defocus magnification.

Generative adversarial networks (GANs) [7] are a class of deep neural networks that define the process of learning as a competition between a Generator network $\mathcal{G}$ and a Discriminator network $\mathcal{D}$. The task of the generator is to create an image parameterized by an arbitrary input which is typically a noise vector. The task of the discriminator is to distinguish between a real image and a generated image. The generator learns to create perceptually real images which can fool the discriminator. The objective function of GANs is defined as $\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}_{GAN}$, where $\mathcal{L}_{GAN}$ refers to the loss function:

$$\mathcal{L}_{GAN} = E_{y \sim p_r(y)} \left[\log \mathcal{D}(y)\right] + E_{z \sim p_z(z)} \left[\log(1 - \mathcal{D}(\mathcal{G}(z)))\right]. \quad (3)$$

Here $\mathcal{D}$ is the discriminator, z is the input noise vector to the generator $\mathcal{G}$, y is a real sample, p_r is the real distribution over target samples and p_z is a normal distribution.

Conditional adversarial networks (cGANs) are GANs with additional conditioning provided to the generator to create images of a specific kind. Conditional adversarial networks have been useful for tasks such as image reconstruction and image-to-image translation. Isola *et al*. [10] provide a comprehensive study of adversarial networks and propose a robust conditional GAN for tasks such as colorization, edge-to-photo synthesis, label-to-photo synthesis etc. Orest *et al*. [13] build on this and propose a cGAN architecture for the task of image de-blurring. Using a combination of adversarial and perceptual loss, they show high quality reconstruction of in-focus images.

Our conditional GAN for defocus magnification is similar to the cGAN proposed in [13]. The generator is adapted from the style transfer network of Johnson *et al*. [11] and consists of two strided convolution blocks with a stride of $\frac{1}{2}$ followed by four residual blocks, an additional input using the magnification parameter m, five more residual blocks and two transposed convolution blocks. Each residual block is based on the ResBlock architecture proposed by He *et al*. [9]. Each block consists of a convolution layer with dropout [30] regularization with a probability of 0.5, followed by instance normalization [32] and ReLU activation [18]. The single-valued magnification parameter m is converted to a 64×64 channel using a fully connected layer and appended to the output at the end of the fourth ResBlock. The input to the generator is the narrow-aperture RGB image with an additional input channel which encodes the amount of focus at each pixel. We use the composite-focus-measure [24] which is a statistical combination of five robust measures of focus. We compute the response of the composite measure at all pixels of the RGB image and append this as a channel to create an RGBF image, where F represents the focus channel. The focus measure channel improves performance and generalization as we shall show in our experiments. A global skip connection (ResOut) is added to accelerate learning and improve generalization [13]. The output of the generator network is a residual image which can be added to the RGB channels of the input image to generate the wide-aperture result $\mathcal{I}'$:

$$\mathcal{I}' = \mathcal{I} + \mathcal{G}_{\theta_G}\left(\mathcal{I} : f(\mathcal{I}), m\right). \quad (4)$$

Here, $\mathcal{I} : f(\mathcal{I})$ represents the RGB image appended with the F channel to create an RGBF image and m is the magnification parameter of Equation 2. The architecture of our generator network is shown in Figure 2.

The discriminator network used in our experiments is similar to [13] and is based on Wasserstein-GAN [1] with gradient penalty [8]. The discriminator is modeled as a critic network and is similar to PatchGAN [10, 15]. Except for the last layer, all convolution layers are followed by instance normalization [32] and Leaky ReLU [34] with α=0.2. We use a composite loss function for our generator. We use a combination of three different loss functions - adversarial loss from the discriminator, perceptual loss and focus-measure loss. The adversarial loss component is defined as:

$$\mathcal{L}_{cGAN} = \sum_{n=1}^{N} -\mathcal{D}_{\theta_D}(\mathcal{I} + \mathcal{G}_{\theta_G}(\mathcal{I} : f(\mathcal{I}), m)), \quad (5)$$

where $\mathcal{I} + \mathcal{G}_{\theta_G}(\mathcal{I} : f(\mathcal{I}), m)$ represents the result image created from the output of the generator.

Perceptual loss, defined in [11], is L2-loss calculated between CNN feature maps of the generated image and the target image. Differences in feature maps encode similarities between two images much better than estimating differ-

Figure 3: This figure shows the performance of our defocus magnification network on images from the test split of the light-field dataset. For each pair: Input image on the left, Defocus magnified image on the right. The left column has a magnification of $m = 2$ and the right column has a magnification of $m = 3$.

ences in the image space [35]. The perceptual loss component is computed as:

$$\mathcal{L}_X = \frac{1}{W_{ij}H_{ij}} \sum_x \sum_y (\phi_{ij}\mathcal{I}'^{gt}_{xy} - \phi_{ij}\mathcal{I}'_{xy})^2, \qquad (6)$$

where ϕ_{ij} represent the feature maps in the VGG19 network trained on ImageNet [5] after the j^{th} convolution and the i^{th} max-pooling layer and W and H are the size of the feature maps.

We also use a loss term based on the response of the composite focus measure [24] over the input image $\mathcal{I}$. We use this term to improve generalization as it encodes the amount of focus at a pixel and is independent of image category. The focus measure loss component is computed as:

$$\mathcal{L}_f = \sum_x \sum_y f(\mathcal{I}_{xy}) \cdot ||\mathcal{I}'_{xy} - \mathcal{I}_{xy}||_1, \qquad (7)$$

where $f(\mathcal{I})$ is the normalized response of the composite focus measure evaluated over the input image $\mathcal{I}$ and $\mathcal{I}'$ is the result image from the generator. At the in-focus locations in the input image where the value of f is high, the result and the input are expected to be identical and this is enforced by the focus measure loss term. The overall loss for the generator is a combination of the three loss terms:

$$\mathcal{L} = \mathcal{L}_{cGAN} + \lambda_X \mathcal{L}_X + \lambda_f \mathcal{L}_f. \qquad (8)$$

The λ_X parameter is set to 100 as defined in [13] and the λ_f parameter is empirically set to 50.

3.2. Training Details

To train our network for defocus magnification we render narrow-aperture and wide-aperture images from a light-field dataset. Srinivasan *et al*. [29] captured a large dataset of 3343 light-fields of similar scenes consisting of flowers and plants. The images were captured using a Lytro Illum camera [20] and consist of 14×14 lenslet images at a spatial resolution of 376×541 pixels. Typically, only the central 10×10 or smaller grid is useful because the lenslet samples towards the corners suffer from heavy clipping. We simulate narrow-aperture images by applying the light-field rendering equation of [20] to the central 3×3 lenslet grid. To simulate wide-aperture images with more blurring we apply the rendering equation to larger grids of 5×5, 7×7 and 9×9 lenslet images. These images notationally correspond to a magnification parameter of $m = \{2, 3, 4\}$ respectively. The images created with larger grids naturally have higher blurring in the defocused pixel locations. All grids are centered around the same lenslet image and the shift-sum parameter is fixed across all renderings. The rendered images thereby correspond to increasing aperture sizes for the same view and serve as ideal training samples for blur magnification.

We divide the light-field dataset into training and test splits of 3000 and 343 light-fields each. Each rendered finite-aperture image is cropped to a square aspect ratio and rescaled to a size of 256×256 pixels before computing the composite focus measure over it. The input image to the network is a $256\times256\times4$ sized RGBF image. The magnification parameter m is provided as an input based on the target wide-aperture image. We augment the training

Config	Loss Function for $\mathcal{G}$	Train PSNR (dB)	Train SSIM	Test PSNR (dB)	Test SSIM
1	Without $\mathcal{L}_{cGAN}$	38.373	0.947	38.170	0.945
2	Without $\mathcal{L}_X$	35.825	0.960	35.946	0.961
3	Without $\mathcal{L}_f$	42.467	0.987	43.782	0.989
4	With $\mathcal{L}_{cGAN}, \mathcal{L}_X, \mathcal{L}_f$	**43.622**	**0.990**	**45.112**	**0.991**

Table 1: Our experiments on training and test splits of the light-field dataset. PSNR and SSIM values are computed between ground truth wide-aperture images and generated images from the network. The configuration 4 network, using all three loss functions, works best.

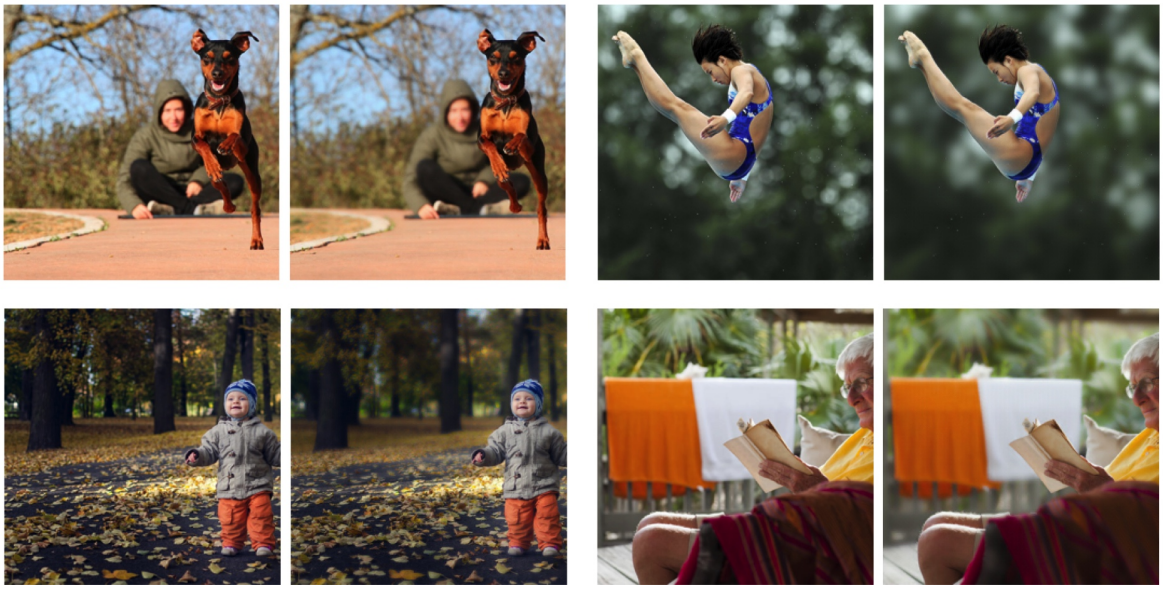

Figure 4: This figure shows the generalization of our defocus magnification network on images selected from the portrait dataset [6] and the blur detection dataset [26]. For each pair: Input image on the left, Defocus magnified image ($m = 2$) on the right. Note that these images are significantly different from images in the training split of the light-field dataset.

data by rotating the square input images by $90°$, $180°$ and $270°$. This improves the generalization of our network on images that were not a part of the training data. Our training experiments are performed on an Nvidia GTX 1080Ti and the generator network is trained for 30 epochs. We use the Adam solver [12] for gradient descent. The learning rate is initially set to 10^{-4} and is linearly decreased to zero during the second half of the training stage. To test the quality of generalization of our network, we sample a set of portrait images from the portrait dataset of [6]. The images from the portrait dataset include human and other subjects which were not a part of the training data. The additional loss term based on the composite focus measure along with adversarial training enables our network to generalize to unseen data, which is demonstrated in Figure 4.

4. Experiments

We perform quantitative evaluation of our network by analysing the performance of defocus magnification over the training and test splits of the light-field dataset. We report the PSNR and the SSIM between ground truth and generated wide-aperture images in Table 1. The table shows four configurations of our generator network. The first configuration does not use $\mathcal{L}_{cGAN}$ during training, the second does not use $\mathcal{L}_X$, the third configuration does not use $\mathcal{L}_f$ and the fourth configuration, which is our proposed network, uses all three loss terms while training. It can be seen that $\mathcal{L}_{cGAN}$ and $\mathcal{L}_X$ have a higher contribution than $\mathcal{L}_f$, however the improvement in both training and test performance on using $\mathcal{L}_f$ suggests an overall benefit when the focus measure loss term is used.

In Figures 3 and 4 we show the qualitative performance of our defocus magnification network on narrow-aperture images. Image pairs in Figure 3 demonstrate blur magnification on narrow-aperture images from the test split of the light-field dataset with magnification parameters of $m = 2$ (left column) and $m = 3$ (right column). The focused pixels remain in-focus and the blur at all other pixel locations

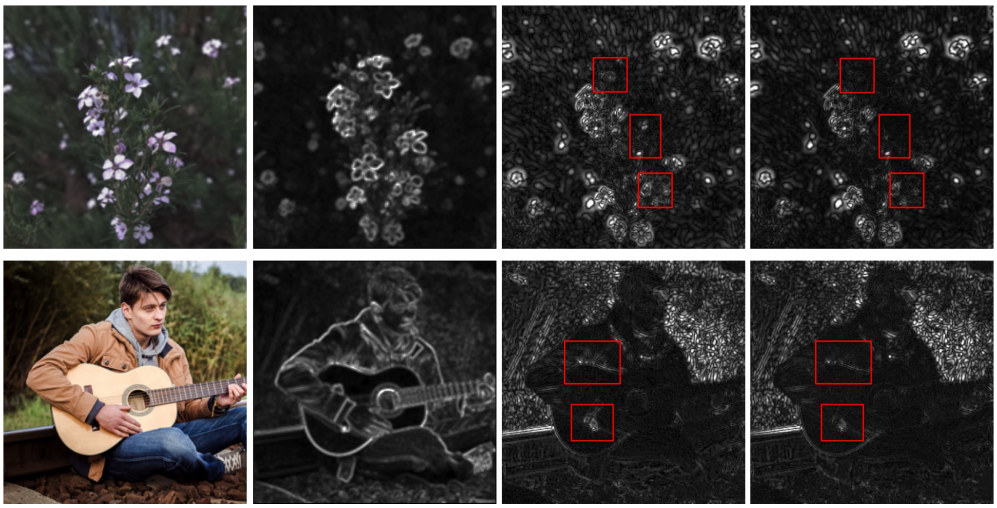

Figure 5: Comparison of our network configurations 3 and 4 from Table 1. Left-to-right: Input image, Focus Measure Response, difference image $|\mathcal{I} - \mathcal{I}'|$ for network 3, difference image $|\mathcal{I} - \mathcal{I}'|$ for network 4. Difference images are expected to be zero at in-focus pixels and high at other pixels. It can be seen that our proposed network 4 is better for both in-focus and defocused pixels. Images are best viewed in the electronic version.

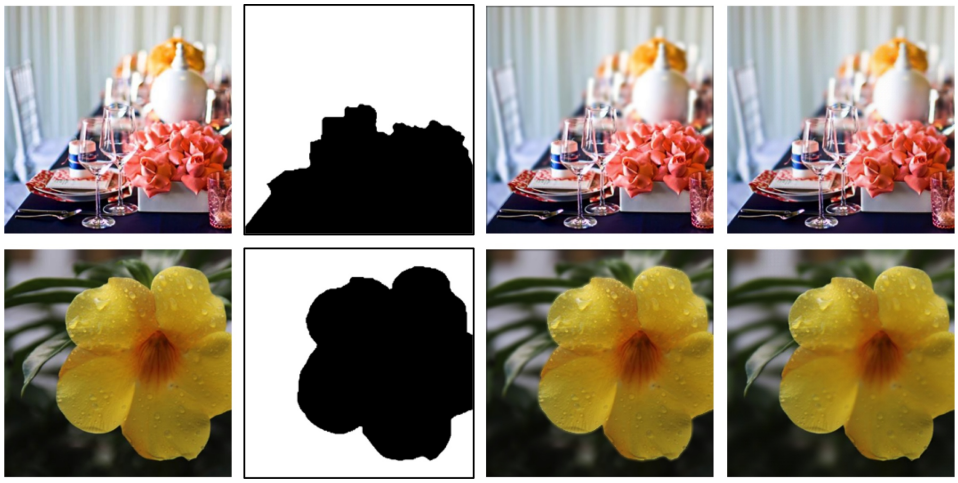

Figure 6: Left to Right: Input Image, ground truth defocus map, depth-aware defocus magnification (generated by re-blurring all pixels from defocused regions of the ground truth defocus map), our defocus magnified image with $m = 1$. Our network produces defocus magnification that is very close to ground truth re-blurring, while carefully preserving the in-focus pixels.

is magnified. Figure 4 demonstrates defocus magnification with magnification parameter of $m = 2$ on images selected from the portrait dataset [6] and blur detection dataset [26]. These images are substantially different from our training images which do not consist of any human subjects. Our network magnifies the blur even in slightly defocused foreground segments, thus truly simulating a shallow depth-of-field.

To study the qualitative improvement provided by our focus measure loss term, we compare the blur magnification of the network configurations 3 and 4 from Table 1. We generate the target wide-aperture image using both networks and study the difference between the input and the output images. The difference is ideally expected to be zero at the in-focus pixels and high at the defocused pixels. The difference images are shown in Figure 5. The difference images corresponding to configuration 3 are shown in the third column and those from configuration 4 are shown in the fourth column of Figure 5. Higher errors in the in-focus regions of configuration 3 suggest that the focus measure loss term is useful for keeping the in-focus regions intact and blurring other regions. We compare our method with state-of-the-art methods [17] and [21] in Figure 7. Our approach keeps in-focus pixels intact, specifically at the boundaries

Figure 7: First Row: Input image, defocus magnification of Ma *et al*. [17], our defocus magnification. Our method keeps the in-focus pixels intact at depth-edges along the hair. Second Row: Input image, defocus magnification of Park *et al*. [21], our defocus magnification. Our approach keeps the in-focus pixels intact near depth edges along the hair and the face.

Figure 8: Our network can be used to iteratively increase the amount of defocus. From left to right: consecutive blur magnification of $m = 2$ on the input image (leftmost column). Integrity of foreground pixels is preserved even after multiple iterations of blur magnification.

of the foreground subject. The methods [17] and [21] fail at depth-edges because of erroneous labeling in the estimated defocus maps. Our blur magnification network can also be used iteratively till the desired blurring in the background is achieved. In Figure 8, we show the effect of iteratively applying a magnification of $m = 2$ on the input image on the left. In Figure 6, we show how our network generates close to ground-truth defocus magnification by comparing our synthesized images with depth-aware defocus rendering using the defocus-maps of [26]. All images in the paper are best viewed in the electronic version.

5. Conclusion

We propose an end-to-end deep neural network for the task of defocus magnification. We use adversarial training with perceptual and focus based loss terms to comprehensively learn blur magnification in a single step. We train our network using aperture stacks created from a large light-field dataset of flowers and plants. To improve generalization we provide an additional input encoding the amount of focus at each pixel and use a focus based loss term during training. We show high quality results using our network on test images from the light-field dataset as well as generic portrait images that are substantially different from the training data. Our single-step approach to blur magnification is ideally suited to photo-sharing applications on mobile devices where the task is analogous to a filter applied on the input image. The performance of our network deteriorates slightly on images of very different scenes such as indoor environments with variable color schemes. Domain adaptation between training images and representative target images may be useful in such cases. This is a direction we intend to pursue in the future.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

[2] S. Bae and F. Durand. Defocus magnification. In *The Computer Graphics Forum*, volume 26-3, pages 571–579, 2007.

[3] J. T. Barron, A. Adams, Y. Shih, and C. Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4474, 2015.

[4] W. Chen, F. Kou, C. Wen, and Z. Li. Automatic synthetic background defocus for a single portrait image. *IEEE Transactions on Consumer Electronics*, 63(3):234–242, 2017.

[5] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[6] F. Farhat, M. M. Kamani, S. Mishra, and J. Z. Wang. Intelligent portrait composition assistance: Integrating deep-learned models and photography idea retrieval. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 17–25. ACM, 2017.

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

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5767–5777, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017.

[11] J. Johnson, A. Alahi, L. Fei-Fei, C. Li, Y. W. Li, and F. fei Li. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[13] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8183–8192, 2018.

[14] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2):228–242, 2008.

[15] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 702–716, 2016.

[16] W. Lijun, S. Xiaohui, Z. Jianming, W. Oliver, H. Chih-Yao, K. Sarah, and L. Huchuan. Deeplens: Shallow depth of field from a single image. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)*, 37(6):6:1–6:11, 2018.

[17] K. Ma, H. Fu, T. Liu, Z. Wang, and D. Tao. Deep blur mapping: Exploiting high-level semantics by deep neural networks. *IEEE Transactions on Image Processing (TIP*, 27-10:5155–5166, 2018.

[18] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

[19] V. P. Namboodiri and S. Chaudhuri. Recovery of relative depth from a single observation using an uncalibrated (real-aperture) camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6. IEEE, 2008.

[20] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report 2(11) 1-11*, 2005.

[21] J. Park, Y.-W. Tai, D. Cho, and I. S. Kweon. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2017.

[22] Pixel 2 cameras. https://research.googleblog.com/2017/10/portrait-mode-on-pixel-2-and-pixel-2-xl.html.

[23] P. Sakurikar, I. Mehta, V. N. Balasubramanian, and P. J. Narayanan. Refocusgan: Scene refocusing using a single image. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[24] P. Sakurikar and P. J. Narayanan. Composite focus measure for high quality depth maps. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1623–1631, 2017.

[25] J. Shi, X. Tao, L. Xu, and J. Jia. Break ames room illusion: depth from general single images. *ACM Transactions on Graphics (TOG)*, 34(6):225, 2015.

[26] J. Shi, L. Xu, and J. Jia. Discriminative blur detection features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2965–2972. IEEE, 2014.

[27] J. Shi, L. Xu, and J. Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 657–665, 2015.

[28] P. P. Srinivasan, R. Garg, N. Wadhwa, R. Ng, and J. T. Barron. Aperture supervision for monocular depth estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng. Learning to synthesize a 4d RGBD light field from a single image. In *IEEE International Conference on Computer Vision, ICCV*, pages 2262–2270, 2017.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res. (JMLR)*, 15(1):1929–1958, 2014.

[31] Y.-W. Tai and M. S. Brown. Single image defocus map estimation using local contrast prior. In *IEEE International Conference on Image Processing (ICIP)*, pages 1797–1800. IEEE, 2009.

[32] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

[33] N. Wadhwa, R. Garg, D. E. Jacobs, B. E. Feldman, N. Kanazawa, R. Carroll, Y. Movshovitz-Attias, J. T. Barron, Y. Pritch, and M. Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Transactions on Graphics (TOG)*, 37(4):64, 2018.

[34] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.

[35] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[36] X. Zhu, S. Cohen, S. Schiller, and P. Milanfar. Estimating spatially varying defocus blur from a single image. *IEEE Transactions on image processing (TIP)*, 22(12):4879–4891, 2013.

[37] S. Zhuo and T. Sim. Defocus map estimation from a single image. *Pattern Recognition*, 44(9):1852–1858, 2011.