Simulating Shallow Depth of Field in Portraits Using Generative Models

Zain Nasrullah

April 21, 2018

Abstract

In response to the recent use of machine learning to create shallow depth of field images, this paper explores a generative approach to this task. While preliminary work in unpaired image translation has already explored this topic, prior methods are not able to reliably preserve the subject of an image and also have not been extended to pictures featuring people. For these reasons, this work introduces a novel portrait dataset containing images with and without a shallow depth of field. It further establishes a baseline, for visual comparison, using both traditional smoothing techniques and the prior research in generative models. To solve the issue of subject preservation, two methods are proposed that take advantage of semantic segmentation: introducing a mask loss and performing a mask overlay. The former involves computing the loss between masks of the subject in a real image and its corresponding generated output. This methods works well in terms of improving preservative performance without impacting the model's ability to smooth the background of an image. The overlay method places a mask of the real image's subject on top of the generator output. Instead of preservation, the generator focuses solely on smoothing the background while the segmentation mask preserves the subject. This method improves cycle-consistency but also introduces artifacts into the generated image. Interestingly, combining these seemingly contradictory approaches during training yields the best result in terms of subject preservation (the mask overlay is not used at test time) at the expense of generating a few artifacts. Additionally, as an unintended consequence, this combined model began generating peculiar results where backgrounds of images were stylized or contained imagined objects.

1 Introduction

In recent years, mobile device manufacturers began looking for new approaches to improve the capabilities smart phones cameras. With the goal of making the mobile photography experience comparable to that of using a Digital Single-Lens Reflex (DSLR) camera, machine learning can used to create aesthetically pleasing images without the cost and bulk of dedicated equipment. Outside of basic classification tasks, recent research has shown that deep networks are successful at tackling creative problems such as style transfer [1], image generation [2], semantic segmentation [3], color enhancement [4] and super-resolution [5].

In photography, a highly sought after effect is a shallow depth of field (DoF), where out-of-focus parts of an image are blurred [6]. With portraits, this is often desirable because it makes the subjects of a photo stand out against the background. Additionally, the shape light takes when blurred is called bokeh [7] and the aesthetic quality of various shapes is a divisive topic among photographers. However, the effect requires a lens designed with a specific aperture shape and size in mind [6]. It is difficult to implement such a design on a smart phone where the camera is limited by the size of the device and the lens cannot be readily swapped. Thus, with this hardware limitation, a need arose for software-based solutions.

In 2014, Google's research team [8] discusses how traditional computer vision techniques, such as Structure-from-Motion (SfM) and Multi-View Stereo (MVS) algorithms, can be used to efficiently estimate depth in an image. With this knowledge, it is possible to simulate a shallow DoF by blurring pixels depending on their depth and location relative to the focal plane. This effect, named Lens Blur, produces good results. With the release of Google's Pixel 2 mobile device and advancements in deep learning [9], the research team iterated on previous work by combining a depth map estimate with a person segmentation mask to simulate a shallow DoF. The feature, called Portrait Mode, became a major selling point of the device. While the final effect looks convincing, it does not achieve the same level of fidelity as a DSLR camera and the necessary

Figure 1: Images with shallow DoF (all) and bokeh (right two) in various locations and lighting conditions

algorithms must be applied at the time of taking the photograph.

The goal of this paper is to explore a generative approach for artificially creating a shallow DoF in portraits. The benefit of taking such an approach is that the effect can be applied to any picture as a post-processing step. Research in unpaired image translation [10] introduced the possibility of simulating shallow depth of field on a dataset of flowers. Extending their work to portraits of people, one can investigate how these models can be augmented to produce more realistic results. To this end, an unpaired portrait dataset is constructed using publicly available images. A flaw with the baseline CycleGAN approach is that it is trying to perform both person preservation and background smoothing simultaneously. This dual responsibility, as discussed in [11], can lead to ambiguity while training generative networks. To help address the issue, this work proposes two methods in which segmentation masks can help stabilize training. The first, taking inspiration from [12], introduces a mask loss that compares the subjects of real and generated images in an effort to encourage the network to preserve people in portraits while smoothing only the backgrounds. The second method, taking inspiration from [11], instead encourages the model to focus only on smoothing backgrounds by overlaying the real subject back on top of the generator output. It is expected that this improves cycle-consistency and discriminative performance since all modifications to subjects are always disregarded; thus, the generator attends solely to smoothing the background.

Ultimately, this research:

1. Introduces a novel, unpaired portrait dataset (with and without shallow depth of field)
2. Discusses traditional and baseline generative approaches for background smoothing
3. Implements two methods, mask loss and mask overlay, to improve person preservation in generative models
4. Validates the impact of these two methods on visual quality and relevant losses

2 Related Works

Google's Portrait Mode [8, 9] is the primary motivator for addressing the task of synthetically creating a shallow depth of field in portraits. A high-level overview of their methodology is illustrated in Figure 2. Pre-processing steps are omitted in the figure because the approaches to depth estimation and semantic segmentation are more relevant to the discussion.

(a) (b) (c) (d) (e)

Figure 2: Images taken from Google's research blog [9], (a) Input Image, (b) Segmentation Mask, (c) Depth Map, (d) Combination of Segmentation Mask and Depth Map, (e) Output Image

Depth estimation involves computing the Sum of Absolute Differences (SAD) between all RGB pixels in a stereo pair [8]. In doing so, it is possible to compare the similarities between pixels conditioned on reasonable priors; solving this optimization problem yields a depth map. While this is the approach taken with the Lens Blur feature, it is also possible to estimate a depth map using different approaches as outlined in [13] and [14]. A common point between these methods is that they all require multi-view images meaning that the camera has to be moving while taking a burst of photos. A major motivating factor for exploring a solution that does not rely on a depth map is the accessibility of this stereo data. While single image depth estimation does exist [15, 16], the results are currently not comparable to state-of-the-art using stereo. With the Pixel 2, the rear-facing camera possesses a feature called dual-pixel autofocus [17] where two photodiodes [9] are used to take distinct images by splitting the lens in half. The distance between the viewpoints is sufficient for depth estimation without moving the camera. This technique, however, requires a specialized camera design and is thus not implemented on the front-facing camera of the phone.

For person segmentation, DeepLabV3+ [3] with the Xception architecture and its portable alternative MobileNetv2 [18] are used. Uniquely, these architectures use atrous convolutions to learn contextual information at multiple effective fields-of-view which helps produce better segmentation results. While the research team at Google have published the majority of their work in this area, the manner in which the combined depth map and segmentation mask create the background blur has not been published. This paper will look at generative alternatives to the same task; however, the Xception model will still be used to compute any segmentation masks. It is assumed that these will be available as input to the model and, as will be discussed shortly, can be used to effectively guide training.

Unpaired Image-to-Image Translation with CycleGANs Following the success of Generative Adversarial Networks (GANs) [19] and Paired Image Translation (pix2pix) [20], Zhu et al. [10] introduced Cycle-Consistent Adversarial Networks (CycleGANs). The motivation behind this model was to enable image translation between two domains using an unpaired dataset. Consequently, for the task of transforming an ordinary portrait into one with a shallow depth of field, it is a promising starting point because paired data, in this context, involves taking identical shots using distinct cameras or lens. This would be difficult to obtain because, in addition to being an entirely manual task, it would require that the pixels containing the person in both pictures correspond to one another. A CycleGAN avoids this issue.

The CycleGAN model contains a generator and discriminator for each domain. The key distinction between prior models is a cycle consistency loss in the objective function which measures each generator's ability to recover the original image after it has been translated into another domain. This cycle consistency loss term can be written as:

$$\mathcal{L}_{cyc}(G, F) = \mathbf{E}_{x \sim p_{data}(x)}(x)[||F(G(x)) - x||_1] + \mathbf{E}_{y \sim p_{data}(y)}(y)[||G(F(y)) - y||_1] \tag{1}$$

Yielding the full objective function [10]:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) + \lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{idt}\mathcal{L}_{idt} \tag{2}$$

In Equation (2), the λ terms are scaling factors controlling the influence of their respective loss terms. The final term in the equation is identity loss which measures a generator's ability to not superfluously distort an image. It does so by measuring the L1 loss between an image y and the output of an $X \to Y$ generator; theoretically, since y is already in the Y domain, the image should not be modified. It is important to take note of this concept because it plays an important role in preserving features from the original image which is similar to the goal of this paper.

Identity-Preserving Face Recovery [12] investigates an adversarial model's ability to recover the original face of a portrait modified with style transfer. To this end, a style removal network is proposed which learns to recover the original portrait with the following MSE and identity loss functions:

$$\mathcal{L}_{MSE}(\Theta) = \mathbf{E}_{(I_s, I_r) \sim p(I_s, I_r)}[||G_\Theta(I_s) - I_r||^2] \tag{3}$$

$$\mathcal{L}_{idt}(\Theta) = \mathbf{E}_{(I_s, I_r) \sim p(I_s, I_r)}[||\Psi(G_\Theta(I_s)) - \Psi(I_r)||^2] \tag{4}$$

where I_s refers to the stylized image, I_r refers to ground truth and $\Psi(.)$ describes feature maps from the ReLU3-2 layer of a pre-trained VGG-19 network. The paper shows that these augmented

losses can produce more realistic results compared to previous works and validates that an adversarial approach outperforms a simple generative model at this task. Consequently, a modified version of these losses will be adapted for use in this paper.

Attention-GANs [11] re-frame the problem of unpaired image translation as object transfiguration. That is, there is a singular object in an image that must be transformed to another domain without modifying the rest of the scene. With this perspective, the authors propose splitting the generative network into two distinct components: an attention and a transformation network. This alleviates the burden on a single network to perform both tasks. The attention network finds the relevant object in an image while the transformation network indiscriminatingly transforms the entire image. The outputs of both networks are combined such that only the pixels corresponding to the object are transformed before being passed to the discriminator. For an input image x, generator G, attention network A_X, transformation network T_X, this takes the form:

$$G(x) = A_X(x) \odot T_X(x) + (1 - A_X(x)) \odot x \tag{5}$$

Thus, an additional attention cycle-consistency term is realized:

$$\mathcal{L}_{A_{cyc}}(A_X, A_Y) = \mathbf{E}_{x \in X}[||A_X(x) - A_Y(G(x))||_1] + \mathbf{E}_{y \in Y}[||A_Y(y) - A_X(F(y))||_1] \tag{6}$$

Yielding the full objective function:

$$\begin{aligned} \mathcal{L}(T_X, T_Y, D_X, D_Y, A_X, A_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ +\lambda_{cyc}\mathcal{L}_{cyc}(G, F) + \lambda_{A_{cyc}}\lambda\mathcal{L}_{A_{cyc}}(A_X, A_Y) + \lambda_{A_{sparse}}\mathcal{L}_{A_{sparse}}(A_X, A_Y) \end{aligned} \tag{7}$$

Equation (7) is reminiscent of Equation (2) with the inclusion of the attention cycle loss and an additional sparse loss term to encourage the network to attend to small regions in an image. The net effect of these changes is a reduction in cycle consistency loss, as the backgrounds remain untouched, allowing the model to focus solely on the attended objects. This paper will try to incorporate a similar strategy of masking the person in the original image back onto the generator output except, given segmentation masks are available, without the need of an attention network. Fundamentally, by transforming the scene and preserving the object, this work will implemented an inverted version of object transfiguration.

3 Method

3.1 Dataset

A challenge with generating a shallow depth of field in portraits is the lack of a reliable dataset. As previously mentioned, collecting paired data for such a task is difficult because it require one to manually take the same photograph with different cameras or lens in a variety of environmental conditions. As in [10], a better approach is to use publicly available unpaired images. To this end, public images were obtained from Flickr by querying for 'portrait bokeh' (shallow DoF) or 'portrait' (ordinary images) [1]. A special camera tag is passed during the query to ensure that all ordinary images come from an iPhone camera and thus do not have a shallow depth of field. The results of the query are summarized in Table 1.

Table 1: Flickr Query Summary

Portrait Type	Camera Type	Images Scraped	Images Retained
Shallow DoF	Any	4000	2157
Ordinary	iPhone 7	4000	2370

A total of 4000 images were pulled from each group. Although duplicate image URLs were to be skipped, the results still contained duplicates and images irrelevant to the search query. A cleanup activity was performed to resolve these issues.

- Automated removal of images that were:
  - identical in resolution and file size within each dataset (unlikely to occur by chance)

[1] The following tags were excluded from the query via Flickr's API to filter irrelevant results: blackandwhite, BW, monochrome, animal, cat, dog, flower

- Manual removal of images that were:
  - incorrectly tagged
  - unfit for their category (ex. insufficient blur for a shallow DoF image)
  - not featuring any real people (objects and virtual avatars discarded)
  - post-processed with image editing software
  - deemed inappropriate for an academic dataset

Despite best efforts, there may be residual images in each dataset that do not meet the criteria specified above. However, this would be a small proportion of the entire dataset and can be considered insignificant noise. Examples of images from the shallow DoF category are shown in Figure 1; an effort was made to obtain images with varying degrees of blur and bokeh. Ordinary images are illustrated in Figure 3. Both sets of data contain high resolution images which will be preprocessed before being input into a model; this will be discussed further in the experimental setup. As an important note, care was not taken to balance the datasets in terms of location, nationality, age and gender. It is assumed that the amount of diversity in each dataset is sufficient for the purposes of this paper.

Figure 3: Ordinary portraits from Flickr

3.2 Semantic Segmentation with Smoothing

In image processing, blurring is known as smoothing and can be considered a filtering operation. Consequently, different types of filters can create distinct blurs with the most common being Average and Gaussian filtering. Both can be interpreted as convolutional operations where a kernel K of size $N \times M$ passes over an image. As a baseline, both aforementioned smoothing techniques will be used to simulate a shallow DoF. This will occur in two steps:

1. A copy of an image will be blurred using a smoothing technique
2. People from the original image will be masked onto the copy via the segmentation mask

A weakness in this approach is that it relies heavily on the segmentation mask being accurate. Furthermore, since none of these smoothing techniques take into account depth, it is not expected to be an accurate recreation of a shallow depth of field. However, as will be seen in the results section, it still produces compelling results. The Average kernel is described as in [21]:

$$K_{avg} = \frac{1}{25}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The Gaussian kernel can similarly be generated by replacing the box filter of equal coefficients by the values generated by the following function [22]:

$$G_i = \alpha * e^{\frac{-(i-(k_{size}-1)/2)^2}{(2\sigma)^2}} \tag{8}$$

$$\sigma = 0.3(\frac{k_{size}-1}{2} - 1) + 0.8. \tag{9}$$

where α is a scaling factor ensuring $\sum G_i = 1$. An example of this technique is shown in Figure 4. As expected, the backgrounds appear to be uniformly blurred and since large kernels are used in both cases, the results are similar. This method is discussed and visualized here to establish what can be accomplished by traditional means.

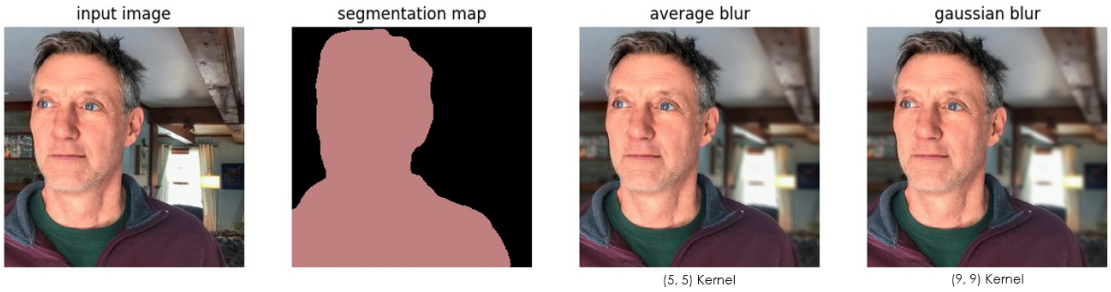

Figure 4: Semantic segmentation with average and Gaussian smoothing

3.3 Shallow Depth-of-Field with CycleGANs

A Cycle-Consistent Generative Adversarial Network (CycleGAN) is the original method for transforming unpaired data between two image domains and thus is the starting point for research. The implementation used in this paper is adapted from the original work in [10]. The generator and discriminator architectures are visualized in Figure 5. The generator convolves a large (7, 7) kernel to generate feature maps from the three channel input, down-samples with two convolutional layers, passes the intermediate results through a sequence of nine residual blocks, up-samples with convolutional layers, and then uses a final convolution layer to create a three channel output. A tanh non-linearity is used in the final layer to scale pixel intensities between -1 and 1. Pooling is not used at all during this process. Cycle and net loss, (1) and (2) respectively, were previously discussed in the related works. The discriminator possesses a sequence of convolutional layers that extract features, and then outputs a binary decision regarding whether the input image is real or fake.

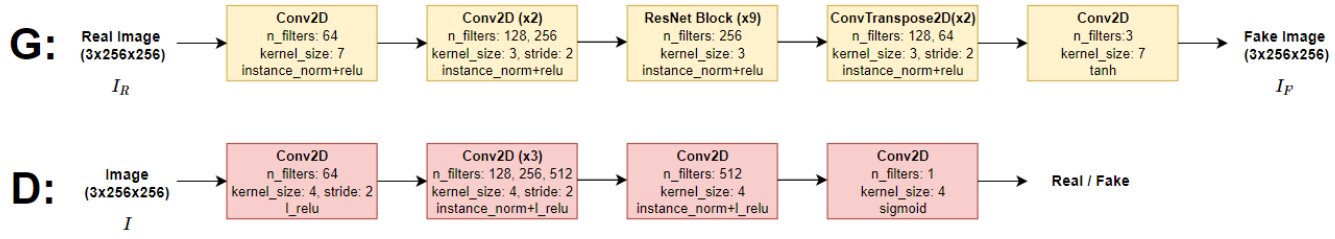

Figure 5: CycleGAN generator and discriminator architecture

It is expected that this baseline approach will produce reasonable results, but may struggle to identify where the person is in an image and thus introduce unnecessary smoothing.

3.4 Segmentation Mask for Guided CycleGAN Training

This subsection explores methods to improve the CycleGAN results by using segmentation masks to guide training. The segmentation results used in this section are generated using DeepLabV3+ [3] prior to training the generative model. Additionally, the segmentation results are saved as RGBA images where the final channel corresponds to the mask alpha M. This channel describes which pixels correspond to subject (1.0) and background (0.0) in an image.

3.4.1 Mask Loss

The first improvement proposed in this work involves introducing additional loss terms into the CycleGAN model. Presently, there is no direct penalization for failing to preserve the person in an image; rather, this learning is expected to occur through the min-max game played by the generator and discriminator. However, similar to the way identity loss is used in the original CycleGAN work [10], additional loss terms can be used to stabilize training. This paper employs the method used in [12] by introducing norm and VGG losses:

$$\mathcal{L}_{mask}(G, F) = \mathcal{L}_{norm}(G, F) + \mathcal{L}_{vgg}(G, F) \tag{10}$$

Here, the norm loss compares people segmented out of the real image $I_{R-Masked}$ and people segmented out of the generated fake $I_{F-Masked}$. An illustration of this is shown in Figure 6. This

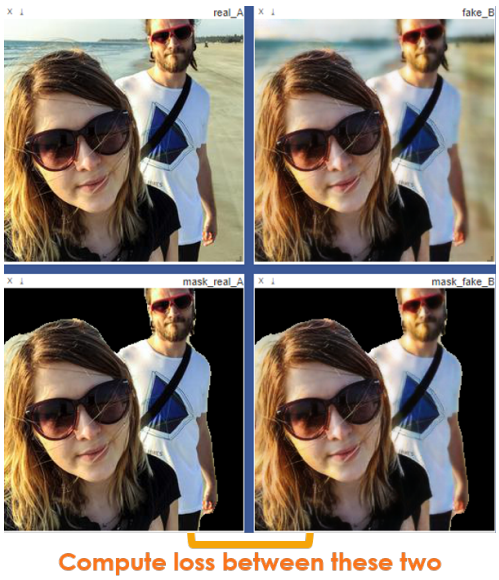

Figure 6: Visualization of real and fake image masks being compared. Top-left: Real Image, Top-right: Fake Image, Bottom-left: Real Image Masked, Bottom-right: Fake Image Masked

loss should help preserve color between the people in the real and fake images. L1 loss is selected, rather than L2 as in the original work [12], to prevent this term from dominating the total loss which would discourage the model from learning to smooth the background.

$$\mathcal{L}_{norm}(G, F) = \mathbf{E}_{I \sim p_{data}(I)}[||I_{R-Masked} - I_{F-Masked}||_1] \tag{11}$$

The VGG loss is found by passing the segmented images into the VGG-19 network pre-trained on ImageNet, taking the extracted features at the ReLU3-2 layer and then calculating the L1 loss between these features. It is expected that these feature maps can characterize structural information based on the prior work in [12] where the ReLU3-2 layer is selected through experimentation.

$$\mathcal{L}_{vgg}(G, F) = \mathbf{E}_{I \sim p_{data}(I)}[||V(I_{R-Masked}) - V(I_{F-Masked})||_1] \tag{12}$$

To ensure that Equation (10) is comparable to other losses, it is also scaled by the cycle-consistency factor λ_{cyc}.

3.4.2 Mask Overlay

An alternative for person preservation is based on the Attention-GAN paper [11]. However, rather than attending to a specific object, it is desirable for the network to attend to everything in the image excluding the specified object. Furthermore, since this work focuses on portraits where humans are the object in question, it is not necessary to train a new network for attention when a pre-trained segmentation model can provide this necessary data as input to the model. Thus, this subsection proposes removing the generator's detection responsibilities in favor of it focusing on its ability to transform. The first step in this process is to invert the segmentation mask alpha M such that any pixels containing people are turned off:

$$M_{inverse} = -1 \times (M - 1) \tag{13}$$

Once this inverted mask alpha is available, an image passing through the generator has its pixels containing people turned off. The real people are then masked back on top of the generated image. This process involves two steps which are summarized in Figure 7: an element-wise multiplication with the inverted mask alpha and then an element-wise addition of the real image mask.

This mask overlaying process, visualized with a real example in Figure 8, occurs every time an image passes through a generator ensuring that the real subjects are always overlaid prior to calculating any loss values. As a result, the network should be encouraged to focus only on backgrounds because, irrespective of how a person is transformed, loss values aren't affected. This also has the expected benefit of improving cycle-consistency because a large portion of the image is unchanged.

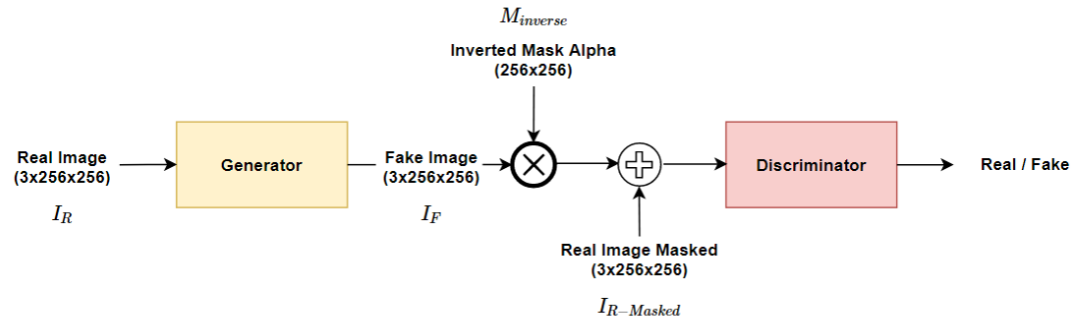

Figure 7: Block diagram for the mask overlay process

Figure 8: Example of mask overlay process

4 Results and Discussion

4.1 Experimental Setup

Experimental conditions are summarized in Table 3. Of particular note, all images are cropped and scaled to 256 x 256 prior to training. This is important because the original images used in this dataset were of a much higher resolution; thus, information is lost in the process of scaling the images down to a size that could fit into 8GB of GPU memory. This sacrifice, however, is necessary to avoid a trade-off in network complexity. Since segmentation masks are also generated prior to training rather than during, data augmentation techniques are not utilized in the experiment which would introduce an overhead of an additional 1s per image per epoch. Each model presented in the results is trained for a total of 75 epochs with the first 50 at a fixed learning rate (LR) of 0.0002 and the subsequent 25 with this rate linearly decaying to zero. While this is relatively little training for a generative model, it presented an optimal trade-off between reasonable results and training time (approx. 30 hours per model).

Table 2: Experimental Setup

Parameter	Value
Dataset	Flickr
Train Set	Ordinary (2270) / Shallow DoF(2061)
Test Set	Ordinary (100)
Image Resolution	3 x 256 x 256
Batch Size	1
Training Epochs	75
LR 1 (50 epochs)	0.0002
LR 2 (25 epochs)	Linear Decay to Zero
Hardware	Nvidia GTX 1080

Additionally, segmentation masks are not passed through any convolutional layers and are only used to turn pixels on/off via the alpha channel or for performing the mask overlay.

4.2 Model Comparison

During experimentation, combining the mask loss and mask overlay produced interesting results. Conceptually, these approaches should be contradictory with one method encouraging the network to preserve people while the other disregards anything that is not part of the background. To explore further, an additional model is trained mask overlay and mask loss scaled down by a factor of 1/2. Curiously, this approach did a better job of preserving the person in the image at test time without applying the overlay. That is, when only mask loss is used to preserve the individual and cycle-consistency exclusively attends to the background, the model's ability to preserve the person improved. This is discussed further in Section 4.3.

Figure 9: Comparison of generative models (b)-(e); (f) the result using conventional smoothing

A general comparison of the generative models is shown in Figure 9. These generated images contain artifacts and checkerboard patterns that are clear indicators of them being synthetic. This is most prominent in (9d) and is indicative of a problem with the mask overlay method. Otherwise, the results seem to be fairly consistent across the models with minor variances. Of the generated images, only the mask overlay (9d) and manual smoothing (9f) overlay the real segmented person back on top of the fake image. For (9e), the mask overlay is only used during training and not at test time. In terms of aesthetic quality, the manual smoothing approach seems to be the most natural but the generated images also possess unique stylistic properties.

The remaining results presented in this section can be categorized into one of four groups summarized in Table 3. Each figure representing these groups contains two rows highlighting a distinct instance of that group. The manually smoothed images are also provided as a reference.

Table 3: Result Groups

Group	**Description**
Successes	One of the generative models improves upon the CycleGAN result
Failures	None of the generative models produced compelling results
Imagined	Cases where the generative model appears to have created an interesting background
Stylized	Cases where the generative model appears to have stylized an image

Figures 10 and 11 contain successful cases that improve upon the CycleGAN in one regard (smoothing the background or preserving the person) without compromising on another.

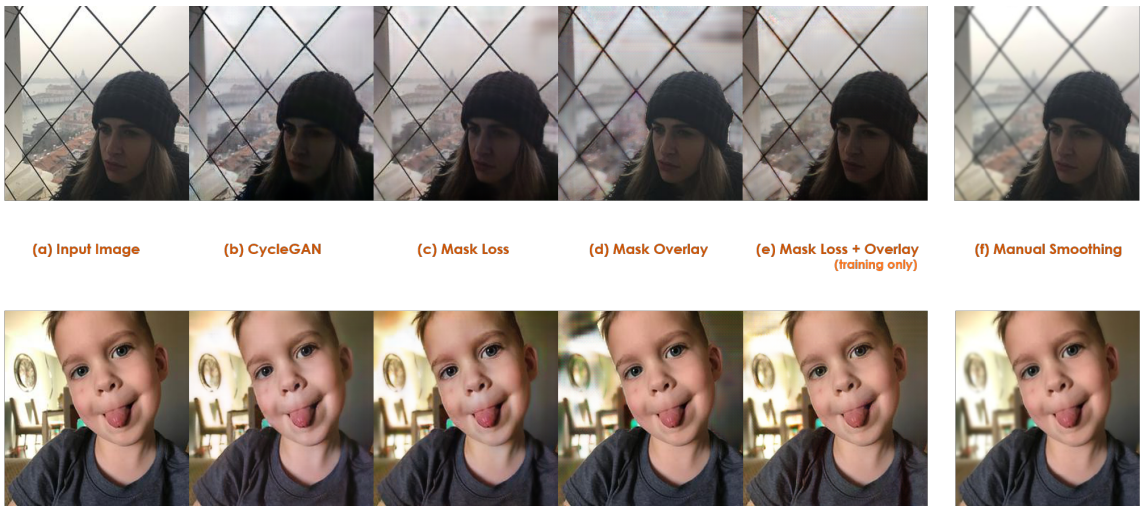

Figure 10: Images where one or more of the models were successful. Compared to the CycleGAN, the top row succeeds at preserving the individual while the bottom row does a better job at smoothing the background.

In general, the inclusion of mask loss yields a model that does a better job at preserving people while still smoothing the background. Since the mask overlay doesn't have to consider preservation, as it masks the real people back onto the generated image, it is able to produce more interesting backgrounds. However, these backgrounds are easy to distinguish as synthetic. Combining the two methods produces a result that strikes a balance between their results.

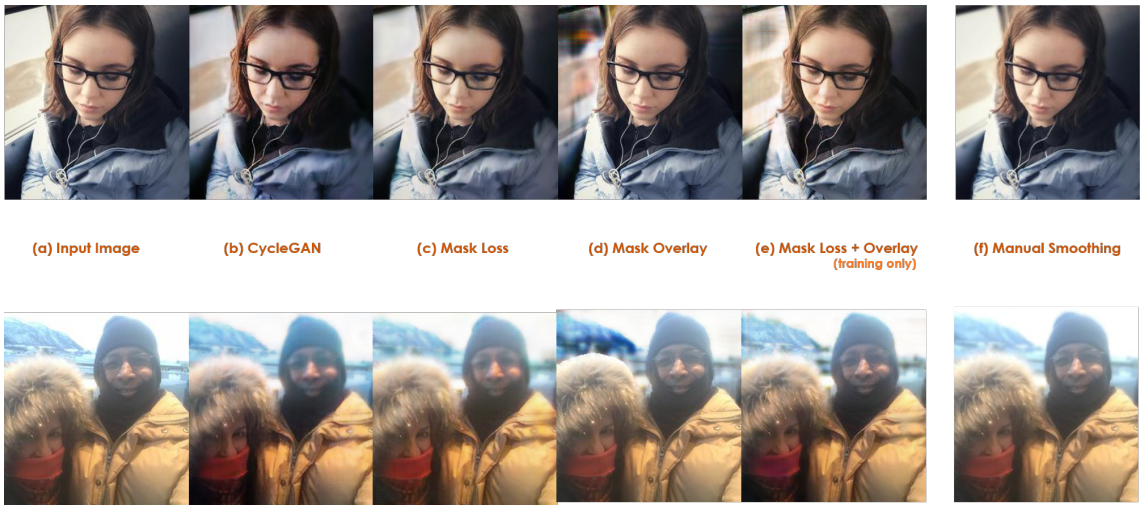

Figure 11: Additional cases where one or more of the models were successful. Compared to the CycleGAN, the top row does a better job at smoothing while the bottom row does a better job at preserving.

Figure 12 describes failure cases. In the top row, the CycleGAN fails to properly smooth the background while in the bottom row it fails to preserve the individual. The proposed models do not improve in this regard and instead introduce artifacts or bizarre coloring. The manual smoothing example shown in the top row is also a failure case because a second individual, at some further depth in the image, is picked up by the segmentation mask. Consequently, this second individual is not considered part of the background highlighting the weakness of not considering depth.

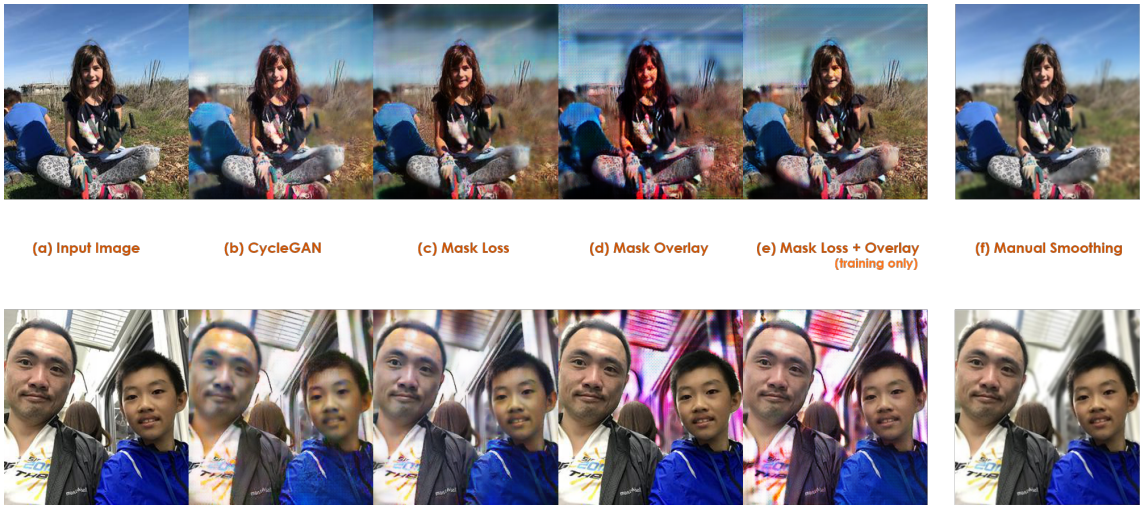

Figure 12: Images where all models were unsuccessful. In both the top and bottom row, if the person is preserved, the model introduces strong distortions or incorrect colorizations into the image.

Figure 13 contains images where the model seems to be imagining things. Both cases (top and bottom row) show people standing before a bland background. Rather than apply smoothing, the proposed models created plausible backgrounds that could be mistaken for real images. This shows that the model hasn't simply learned to smooth but rather is transforming backgrounds, stylistically, to the domain of shallow depth of field images.

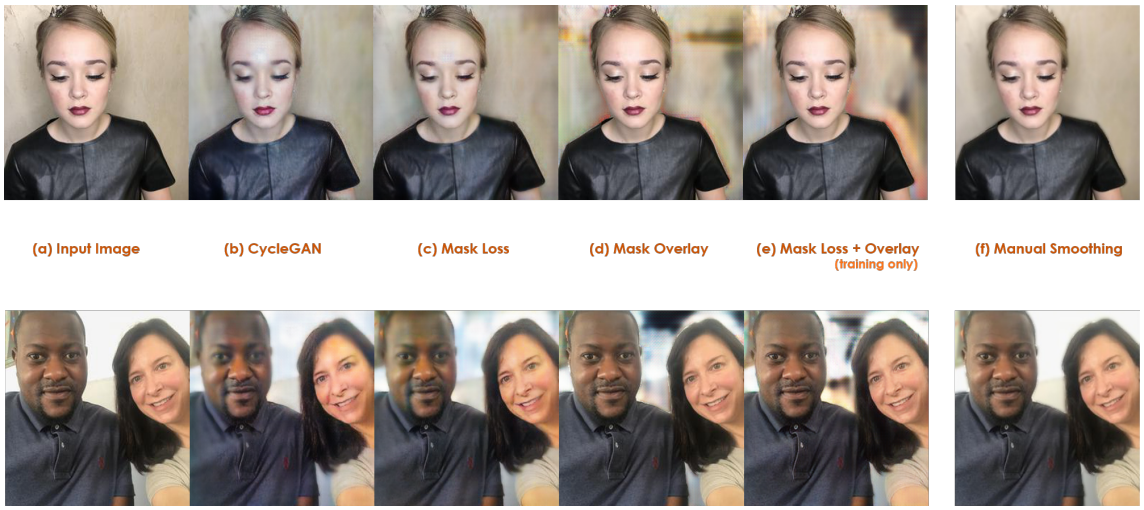

Figure 13: Cases where the model seems to be imagining a background

Figure 14 corroborates this hypothesis illustrating stylized versions of the input image. While these could still be considered shallow depth of field images, their backgrounds look more like the output of a style transfer model. It is not clear what exactly is causing this, but it is certainly an interesting result.

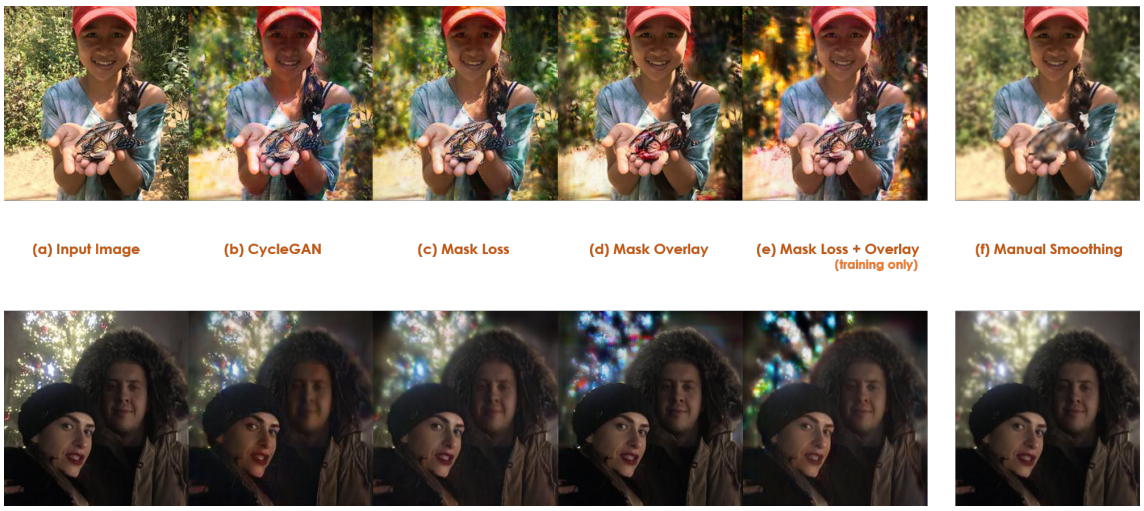

Figure 14: Cases where the output is highly stylized

4.3 Loss Curves

Having visually inspected the model outputs, it is possible to gain further insight into performance through the analysis of training loss curves. Firstly, 15 validates that the mask overlay improves cycle-consistency. This is not a surprisingly result since the inclusion of an overlay reverts a significant portion of the generated image back to the input, but the benefit of doing so is that the model can focus on attending to the necessary parts of an image. The task of maintaining cycle-consistency in this specific problem is difficult because it involves transforming an image with a smooth background into one with a clear background. This is akin to super resolution without having a paired reference for doing so. For this reason, the mask overlay essentially reduces the workload of the generator and also makes the discriminator attend specifically to image backgrounds.

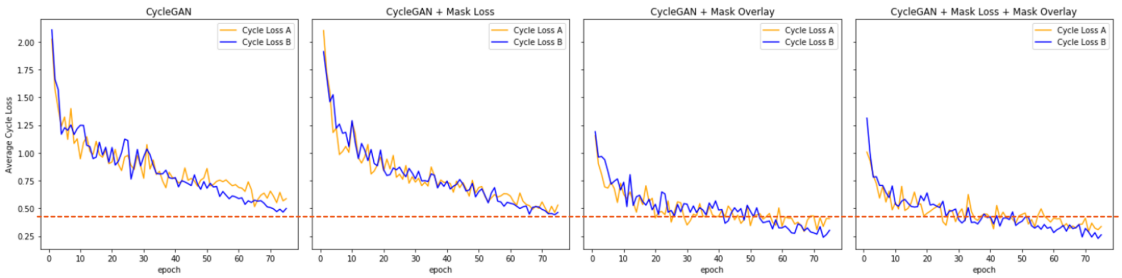

Figure 15: Comparison of cycle loss in generative models. Inclusion of the mask overlay reduces cycle loss in both cases where it is applied.

A more interesting result is the effect of the overlay on mask loss which is shown in Figure 16. Both the norm and VGG loss achieved lower values through the inclusion of the mask overlay during training. This occurred in spite of reducing mask loss by a factor of 1/2 which should theoretically place less emphasis on preserving the people presented in the images. This may be the result of allowing the generator to focus on a single task during training. As mentioned in [11], ambiguity regarding which task to improve during training can result in the model performing sub-optimally in both. By ensuring that cycle-consistency loss is not directly tied to a detection (or in this case preservation) task, one can get around this issue. As a result, the model trained using the mask loss and overlay (during training only) performs better at preserving the person at test time. Figure 17 shows an example of this.

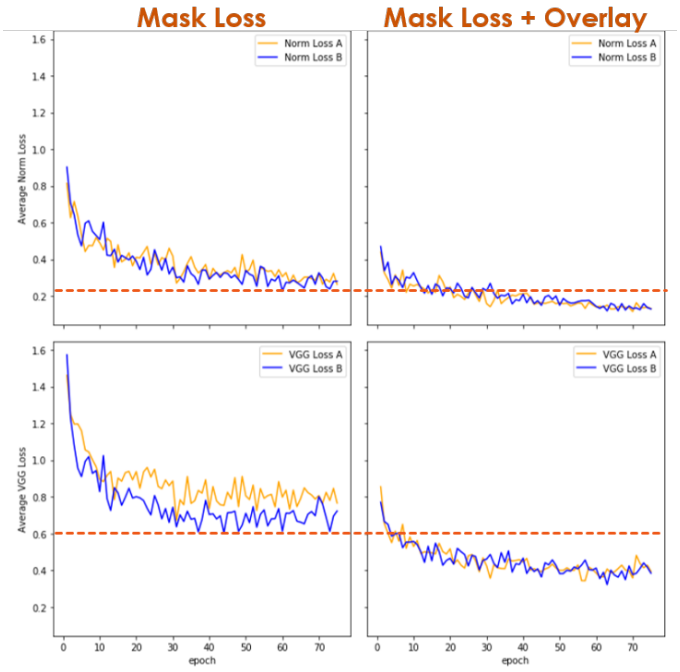

Figure 16: Effect of overlay on mask loss. Losses are much higher when only using the mask loss(left) than when the overlay is also included during training (right)

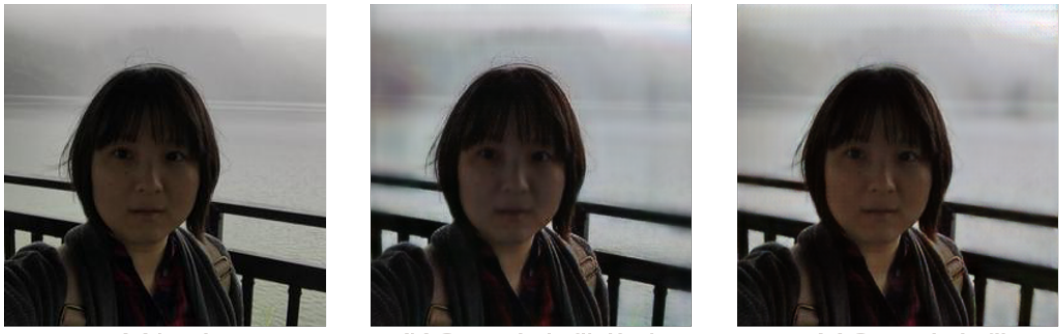

Figure 17: Visual impact of mask loss and overlay. Including the overlay improves the models ability to preserve the individual in the image.

4.4 Limitations and Future Work

The limitations associated with the work presented in this paper can be represented by three categories.

4.4.1 Artifacts, Distortions and Discoloration

The generated images in many cases contain distortions or checkerboard patterns which make them appear unrealistic. This is especially prevalent in models with the mask overlay and prevents the discriminative network from learning because it is not challenging to distinguish true and fake images. This eliminates the adversarial nature of the model and may be responsible for mode collapses. There are a number of different approaches to tackle this issue:

1. Training the network for longer and experimenting with the learning rate
2. Investigating different network architectures
3. Replacing deconvolutions with nearest-neighbor interpolation [23]
4. Using data augmentation or collecting more data

These techniques can be explored in future work to build upon the results of this paper.

4.4.2 Reliance on Segmentation Mask

There are strong limitations with relying solely on the segmentation mask to guide the training of a network designed to perform a task involving depth. Firstly, the segmentation masks identify all people in an image irrespective of depth or focus. As a result, the model will not learn to properly smooth backgrounds that contain people. Furthermore, segmentation results aren't guaranteed to be accurate and may mislead the model in some cases.

It is thus advisable to combine the depth map and segmentation mask to, at the very least, identify which subject in the image is considered to be the target. This can be accomplished, similar to Google's approach in [9], by taking the subject that's closest to the camera. While this may not work in all cases, it would significantly reduce noise. It may also be worthwhile to investigate whether an attention-based model, as suggested in [11], could effectively learn which subject in an image should be attended to.

Another interesting direction may be to include the depth map estimate as an additional channel during training. This would increase the complexity of the model, but may encourage the network to learn the amount of blur to apply. However, it is unclear whether a depth map estimate of a shallow depth of field image would be accurate or provide additional information over the segmentation mask.

4.4.3 Resource Requirements of GANs

One of the biggest problems with taking a generative approach, over using traditional techniques for smoothing, is the resource requirements of the network. This is particularly problematic when one considers the amount of information lost while scaling down the images from their original resolution to 256 x 256. With the hardware used in this experiment, even a slightly larger resolution of 512 x 512 could not be used without significant impact to the network's architecture. A possible solution to this may be to split full-sized images up into square blocks and then train on these individual blocks. While the model will lose the ability to see a subject in its entirety, it may still be able to differentiate the unique characteristics associated with skin and clothing.

In addition to this issue of resolution, a generative model can take a long time to train and still suddenly collapse. In this paper, only 50 epochs are performed at the initial learning rate which, with the introduction of additional loss terms, is only sufficient for producing acceptable results and was not expected to be realistic. To ensure that results are fairly compared, each model must be trained for the same number of epochs which made it difficult to justify a longer training time without significant overhead. The total training length of 75 epochs required approximately 30 hours to train (per model) and this also introduced challenges with respect to tuning hyper-parameters.

5 Conclusion

The ultimate goal of this work was to explore a generative approach to Google's Portrait Mode and build upon the shallow depth of field framework established in the CycleGAN paper. To facilitate this generative approach, a novel dataset consisting of ordinary and shallow depth of field portraits is constructed using publicly available images on Flickr. Traditional smoothing techniques and the original CycleGAN model are first used to establish a visual baseline. This work then proposes to solve the problem of subject-preservation through the use of segmentation masks and two methods: mask loss and mask overlay. The former involves comparing the segmentation masks between the real and fake images thus encouraging the network to preserve subjects in the image. This method is shown to visually improve preservative performance without negatively impacting smoothing. Mask overlay, on the other hand, attempts to make the network focus solely on smoothing by overlaying the segmentation mask at the output of the generator. Despite improving cycle-consistency, this method ended up generating distortions and artifacts in the fake image due to a lack of training or mode collapse. Interestingly, combining these two methods during training and then removing the overlay at test time greatly improved preservative performance. This improvement came at the expense of introducing artifacts into the smoothed background but also gave the model the unique ability to imagine and stylize backgrounds. In future work, it is recommended to incorporate a depth map estimate into the model and address the issue of artifacts by using methods to stabilize GAN training.

References

[1] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[2] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv:1802.02611*, 2018.

[4] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):118, 2017.

[5] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.

[6] Nikon. *DSLR Camera Basics*, 2018. http://imaging.nikon.com/lineup/dslr/basics/19/05.htm.

[7] Nasim Mansurov. *What is Bokeh?*, 2018. https://photographylife.com/what-is-bokeh.

[8] Carlos Hernández. *Lens Blur in the new Google Camera app*, 2017. https://research.googleblog.com/2014/04/lens-blur-in-new-google-camera-app.html.

[9] Marc Levoy. *Portrait Mode on Pixel 2 and Pixel 2 XL*, 2017. https://research.googleblog.com/2017/10/portrait-mode-on-pixel-2-and-pixel-2-xl.html.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

[11] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. *arXiv preprint arXiv:1803.06798*, 2018.

[12] Fatemeh Shiri, Xin Yu, Fatih Porikli, Richard Hartley, and Piotr Koniusz. Identity-preserving face recovery from portraits. *arXiv preprint arXiv:1801.02279*, 2018.

[13] Jiahao Pang, Wenxiu Sun, JS Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *International Conf. on Computer Vision-Workshop on Geometry Meets Deep Learning (ICCVW 2017)*, volume 3, 2017.

[14] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[15] Bo Li, Yuchao Dai, Huahui Chen, and Mingyi He. Single image depth estimation by dilated deep residual convolutional neural network and soft-weight-sum inference. *arXiv preprint arXiv:1705.00534*, 2017.

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, volume 2, page 7, 2017.

[17] Nikon. *An Introduction to Dual Pixel Autofocus (DPAF)*, 2017. http://learn.usa.canon.com/resources/articles/2017/intro-to-dual-pixel-autofocus.shtml.

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. In *CVPR*, 2018.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[21] OpenCV Documentation Team. *Smoothing Images*, 2013. http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_filtering/py_filtering.html.

[22] OpenCV Documentation Team. *Image Filtering*, 2014. https://docs.opencv.org/3.0-beta/modules/imgproc/doc/filtering.html.

[23] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.