# Photobombing Photobombers

**Thibault Févry**
New York University
thibault.fevry@nyu.edu

**Mihir Rana**
New York University
ranamihir@nyu.edu

**Kenil Tanna**
New York University
kenil@nyu.edu

## 1 Introduction

Recent progress in image segmentation have led to algorithms that can reliably provide bounding boxes around objects and even label every pixel of an image. In addition, the development of generative models, most notably generative adversarial networks (GANs) has led to new approaches for image completion. In this project, our goal is to combine the two approaches to enable easy removal of undesired objects from images while reconstructing a realistic background. In short, we want to solve the issue of photo-bombing. To this end, we integrate segmentation and generation models to remove objects while leaving as few artefacts as possible. We contribute a pipeline that is able to quickly and effectively remove objects in images and explore the ethical implications of our work.

## 2 Related Work

Removing photobombers or specific objects from images is a problem that has been worked on from many years. Deterministic algorithms like [1] show very good results by using a patch based method to identify the background and remove objects. However, they do have some shortcomings, in particular when it comes to handling curved structures, sense of depth and differently structured patches.
In deep learning, variational auto encoders (VAE) [6] and generative adversarial networks (GAN) [2] have recently gained traction for their performance on specific generation tasks. In particular, a GAN architecture was used for Semantic Image Inpainting [11] which outperformed other approaches, such as DC-GAN [8] or VAE, at reconstructing faces. To this end, they use a context loss for the remaining image area and a prior loss to penalize unrealistic images. Iizuka et al. [5], by focusing on globally and locally consistent training of an adversarial network, are able to construct a network that can fill background although it falls short of filling complex objects, such as animal or people. Recently, Ulyanov et al. [9] showed that randomly-initialized neural networks used as handcrafted priors give excellent results in inpainting, and thus that the structure of a generator network is sufficient to capture low-level image statistics prior to any learning, rather than learning realistic image priors from a large number of example images. Finally, [10] propose a interesting blind inpainting solution that needs not a priori specification of the area to be inpainted, with good performance on tasks such as text or watermark removal.

## 3 Methodology

Our objective is to first perform image segmentation and remove certain objects, and then to fill the blank with the background one could expect. To this end, we integrate both approaches to provide an end-to-end pipeline where users select a pixel, the object whose pixel is selected is identified and removed and the generative model fills the image back in. This pipeline is illustrated in Fig. 1.

We train (or use pre-trained models whereby available) our models on the MS-COCO data set [7] and then integrate them on our pipeline. The two pre-trained models are trained separately as there does not seem to be any benefit from training end-to-end. Details on the individual methods are given

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

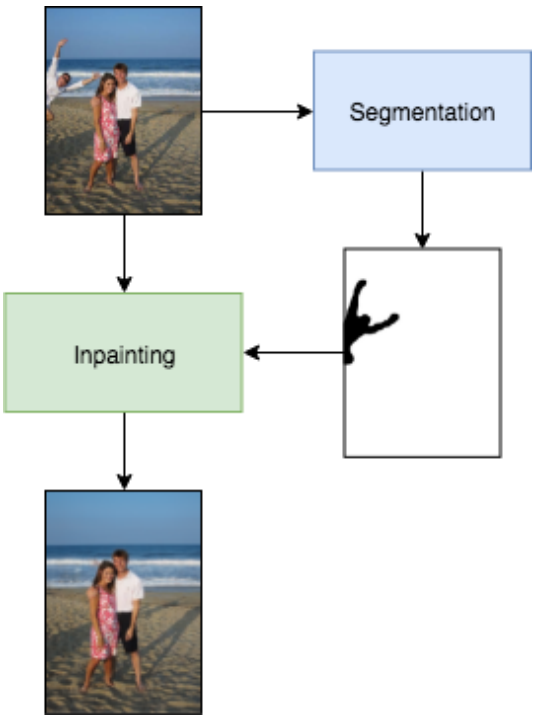


Figure 1: Proposed Pipeline

below. Evaluation for both models trained independently will follow standard evaluation metrics in related work (see below). For the end-to-end pipeline, we conduct qualitative evaluation by assessing the performance of our pipeline on various image replacement settings.

### 3.1 Segmentation

Research on segmentation methods has been very prolific, so we an abundant body of literature from which to take inspiration. Given our task, our desiderata for the model were:

- Fine-grained segmentation: for inpainting, the less the image has been altered, the easier it is to reconstruct it without noticeable artefacts. Therefore, we want methods that provide us with masks that cover the object to be removed and as little extra of the image as possible. Therefore, we want masks to be at the pixel level rather than broad bounding boxes. This excluded a body of literature focused on providing bounding boxes (which was the main focus of COCO early on [7]).
- Accurate segmentation: only removing part of the object leads the inpainting methods to reconstruct it in images, thus defeating our goal. Therefore, our methods have to be quite accurate.
- Speed: Given its plausible use, a key concern for out tool is quick finality. Therefore, we were looking for methods that were fast to run at *inference time*, with training time being less important.
- Relevant classes for our problem: Although fine-tuning networks based on their last layer's weight is always an option, we wanted the model to have been exposed in training to the kind of images we might use at inference time. Therefore we did not explore models that had been trained on narrow-domain image segmentation tasks, such as medical image segmentation.

These considerations lead us to consider several models. After experimenting on images, we found Mask R-CNN [3] to best fulfill our requirements. A quick overview of the architecture can be seen on Figure 2.

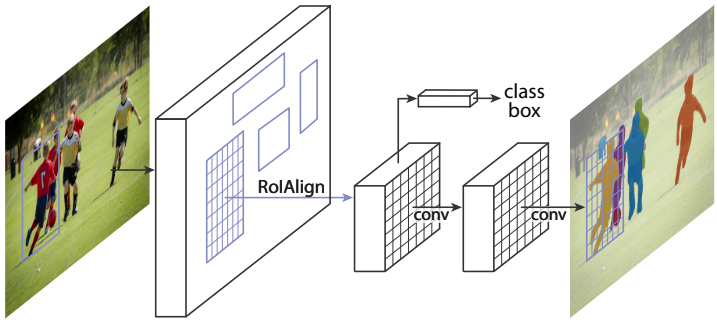


Figure 2: Overview of Mask R-CNN

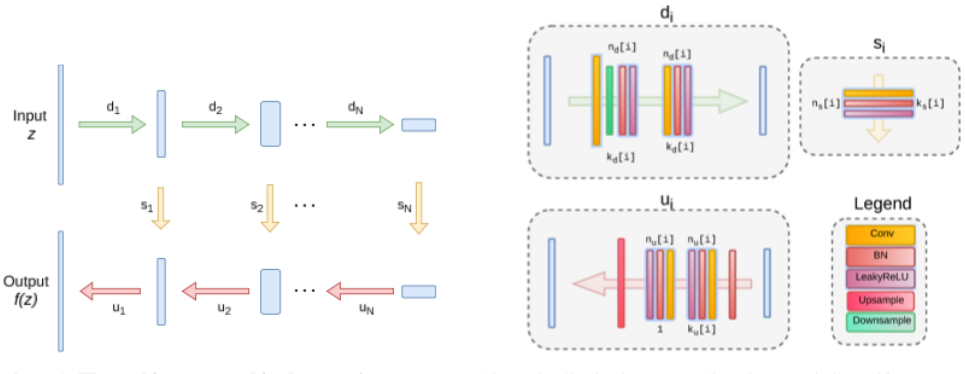


Figure 3: Encoder-Decoder based architecture with skip connections

### 3.2 Inpainting

Our focus was on comparing two methods for the task of image in-painting. One is an unsupervised method, which does not use any pre-trained networks and the other is a supervised method, trained on the ImageNet data set.

#### 3.2.1 Deep Image Prior [9] (DIP)

This method takes inspiration from denoising auto-encoders where a corrupted version of the image is generated, and a network is trained to generate a better image from that corrupted version. Let $x$ be the image such that $x \in \mathbf{R}^{3\times H\times W}$ and $z \in \mathbf{R}^{C^{'}\times H\times W}$ be the noise, then we can define , $x = f_\theta(z)$, where $f_\theta$ is a convolutional network with randomly initialized parameters. We find the parameters $\theta$ such that:

$$\theta = arg\min_{\theta} ||(x - x_0) \odot m||^2 \tag{1}$$

Here $x_0$ is the image with the blacked out portion where the mask is and $m$ is a binary mask corresponding to $m \in \{0, 1\}^{H\times W}$. The architecture used for $f_\theta$ can be seen in Fig. 3

Intuitively, the network learns to model the image aside from the mask due to the MSE loss given above and then fills the mask by using the learned features. Hence, this method is likely to perform better when background represents the majority of the image.

### 3.3 WGAN with Contextual Attention [12] (GIICA)

Here they improve upon ideas of various different methods and use tricks for efficient architecture training. Their improved architecture is illustrated in Fig. 4. Then, a contextual attention layer is

added to the image to leverage features from distant spatial locations. The contextual attention layer and the added contextual features can be seen in Fig. 5 and its visualization in Fig. 6.

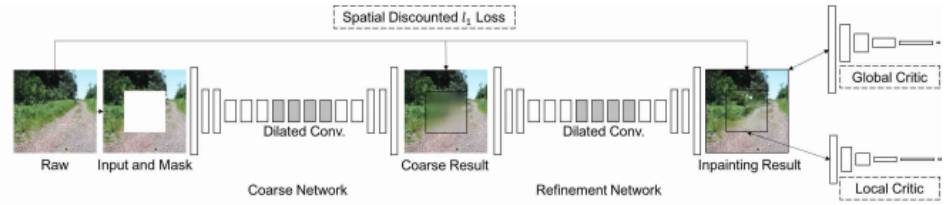


Figure 4: Improved Architecture

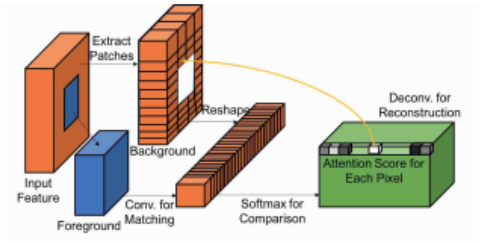


Figure 5: Contextual Attention Layer

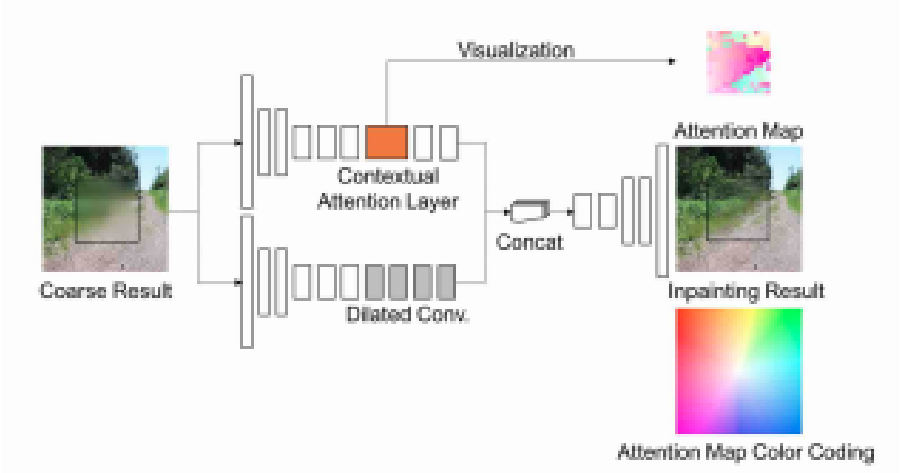


Figure 6: The Architecture with Contextual Attention Added

## 4 Results

Our experimental setup was a Intel(R) Broadwell @ 2.60GHz, and NVIDIA P40 (24GB Memory). The running time for the Deep Image Prior method was $\sim$ 5 minutes per image. An example of the results for the mask and final result from Deep Image Prior is shown in Fig. 7.

Figure 7: Original image (left), Mask Obtained by Segmentation (middle), and Inpainted Image (right)

Next, we compare the different inpainting methods in Fig. 8. We find that Deep Image Prior generally outperforms GIICA but we notice that in the last image both perform poorly. Nevertheless, GIICA, due to its supervised training regime, performs slightly better background replacement. We suspect both our models, because they have an image prior, tend to have trouble when the removed objects are somewhere where there would generally be an object rather than background.

Figure 8: Comparison between DIP (middle) and GIICA (right)

We also compare the architectures in Fig. 9 and find that using deep encoder-decoder architecture with skip connections generally works best, leading to sharper and better inpaiting. The comparison between different encoder-decoder architecture can be seen in Fig. 10.

Figure 9: Comparison between different architectures, U-Net, ResNet, Encoder-decoder architecture respectively.

Figure 10: Comparison between different depths for the architectures: 2,4 and 6 respectively

We also observed that training for too long led the model to overfit and the loss function to diverge as shown in Fig. 11. We see that after 5000 iterations it converges but running it for longer leads to divergence. The plot of loss vs iterations is shown in Fig. 12.

Figure 11: Comparison between running it for 5000 iterations (middle) vs 50000 iterations (last)

As emphasized earlier, our method also fails on some images, as shown in Fig. 13. We find that when background does not represent a large enough portion of the original image our model fails to inpaint realistically.

Finally, one more potential hindrance to achieving an end-to-end robust system is that fact that Mask R-CNN, as it stands today, does not identify objects associated one primary object – like the leash of

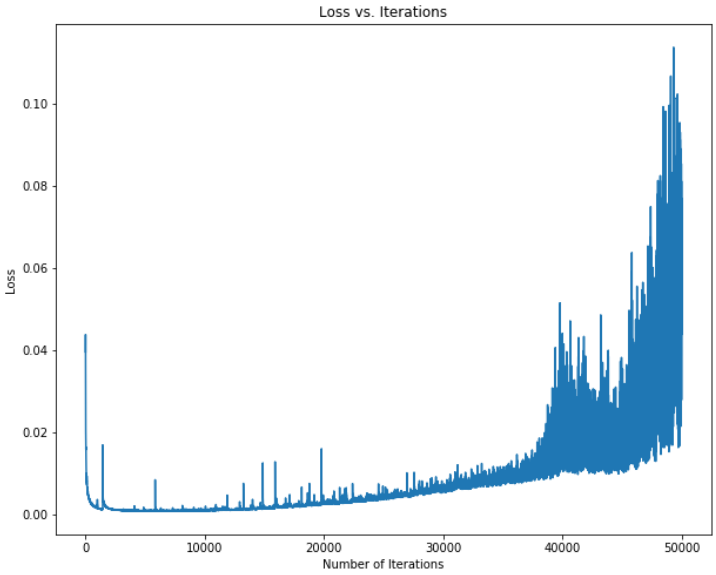


Figure 12: Plot of Loss vs. Iterations for the *Beach* example


Figure 13: Failure cases for deep image prior

a dog, the shadow of a person, etc. This, in turn, leads to cases such as those shown in Fig. **??**, where we are not able to remove the leash of the dog (top), and the shadow and the stick of the person in the snow (bottom). We believe this is a major limitation of our system and will investigate ways to associate masks together in future work.

In Fig. 15, we perform the segmentation of the person to the extreme-right, and try to inpaint it with the background. A cursory look would lead us to believe that DIP is not work well, since the resulting image has a portion which is filled with "nonsense". However, upon carefully analyzing the result, an interesting insight is revealed – even though a major portion of the image is taken up humans – DIP

Figure 14: Potential Hindrance to Inpainting

is smart enough to inpaint it with the background. The resulting inpainted part comprises portions of the grass, leaves, etc., and partial, obscure trunk of a tree, which is commendable given the context.

## 5 Ethics Statement

Although we emphasize the use of our model to remove photobombers, our work is more general and can be used to remove any object in a still image to replace it with a still background, which may be used for malevolent purposes, such as automated censorship, manipulation, etc. Indeed, such feats have been carried out long digital images were around, with the most prominent examples probably being image censorship in USSR (see Figure 16 for a notorious example). More recently, commercial software like Adobe PhotoShop has facilitated such feats, even providing features to remove objects and replace them with background.

What deep learning methods such as the ones highlighted in our paper bring is a way to automate such methods and run them on a massive scale. With the advances in facial recognition, such as those by startup SenseTime, it is now feasible to remove a specific person from thousands of photos in just a few hours. More engineering efforts are likely to enable real-time removal and replacement.

In order to avoid such uses of these technologies, we believe it is important to bring awareness of what they enable and what their current state is. Thus, in this work we have chosen a relatively harmless

Figure 15: Another potential hindrance

Figure 16: Removal of Nikolai Yezhov, executed in 1940, by USSR's censorship service. Such manipulation was common in USSR

use-case to illustrate these risks and showcase what is currently possible with openly available software and research.

## 6 Conclusion

We explored the integration of image segmentation and generative models to replace inpainters with generated background in a coherent way. To this end, we tune and compare different architectures on a wide variety of pictures. We find that a combination of Mask R-CNN and Deep Image Prior provides high-quality results in under 5 minute per frame, with the replacement being hardly noticeable in most generated images.

## 7 Future Work

In the future, we would like to explore ways to speed up inpainting and automatically find hyper-parameters for image inpainting depending on image structure and mask. This should enable us to gather large data sets of photobombed images so that we can automatically detect instances of photobombs instead of having the user select them manually. Also, it would allow us to integrate the full pipeline on a website demo, so that it can be easily used and demonstrated.

Moreover, automatically detecting the optimal architecture for inpainting is also an important succeeding task. This could, for example, be achieved by analyzing the image context and masks (and their size with respect to the entire image).

Finally, as we showed before, detecting masks associated with an object (dog leash, shadows, etc.) during segmentation and inpainting them too (and not just the object itself) is imperative to building a robust system. This could be done by modifying the existing architecture of Mask R-CNN to segment associated objects (like shadows, etc.) linked to a primary object.

We also wish to explore solutions to the two main drawbacks of our method, including removing objects associated with a removed mask (i.e shadows, dog leashes, etc.) and being able to handle more objects (in fact, any object, following the works of [4]).

## References


[1] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, Sept 2004.

[2] 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.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[4] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. *arXiv preprint arXiv:1711.10370*, 2017.

[5] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, July 2017.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[8] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[9] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Deep image prior. *CoRR*, abs/1711.10925, 2017.

[10] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 341–349. Curran Associates, Inc., 2012.

[11] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6882–6890, 2017.

[12] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. *CoRR*, abs/1801.07892, 2018.