GreenGAN:Image Transfiguration using Unpaired Data

Report submitted in fulfillment of the requirements
for the Stream Project of

Eighth Semester B.Tech. (Part IV)

by

Manik Goyal

15075029

Under the guidance of

Prof. Rajeev Srivastava

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI
Varanasi 221005, India
May 2019

Dedicated to

My parents and teachers without whom this wouldn't have been possible

Declaration

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.

2. The work has not been submitted for any project.

3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi

Date:

Manik Goyal
15075029
Senior Undergraduate
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

*This is to certify that the work contained in this report entitled "**GreenGAN:Image Transfiguration using Unpaired Data**" being submitted by **Manik Goyal (Roll No. 15075029)**, carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision.*

Place: IIT (BHU) Varanasi
Date:

Prof. Rajeev Srivastava
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Acknowledgments

I would like to express our sincere gratitude to **Prof. Rajeev Srivastava** for his guidance and constant supervision. He provided us with necessary information and supported us throughout the project.

Place: IIT (BHU) Varanasi

Date: **Manik Goyal**

Contents

List of Figures

List of Tables

Abstract

Image Transfiguration implies to change the appearance or form of an image. Deep learning has helped in automating various types of Image Transfiguration processes. Extensive manual effort in producing target image, for training such deep neural nets has been a major restricting block in using these frameworks. We propose a general purpose Image Transfiguration model - GreenGAN which is able to swap/change the background of an image learning only using unpaired data(object images and background images).

The GreenGAN model acts similar to that of a green-screen used in post-video editing and is entirely trained in a weakly supervised manner with only 0/1 domain labels. The proposed Generative Adversarial Network(GAN) framework learns to separate object and background from any image and similarly combine any object and background pair to form a new image.

In this report we present a comprehensive literature review of some recent approaches which propose to improve the qualitative performance of the GAN architecture as well as improve the training aspect of such networks. We present experiment details of some of these approaches and show how they are incorporated in the basic GreenGAN model.

Chapter 1

Introduction

1.1 Overview

Image Transfiguration refers to changing the appearance or form of an image. It acts as a general term to specify any modification done over an image. Modifications can be done over an image for various reasons.

Over the recent years many researchers have worked over automating various type of Image Transfiguration processes like foreground/background segmentation [1, 2] of image, image editing to add/modify features in an image [3, 4, 5], transforming images from one domain to another domain [6, 7, 8]. With the advent of deep learning [9, 10] and computation capability most of these works have focused on utilizing a deep neural net architecture for the same. The basic framework for such transfiguration models has been to form a conversion network between the input and target image provided that both the images are available. As more and more diverse and complex image transfiguration processes are being worked on, it results in a tremendous human workload to collect/generate manually the target images for learning such conversions. The requirement of paired images(input and target image pair) for training, hinders not only the research aspect but also restricts the real-world

application of such models. Since, having trained for a particular type of images, the model requires further training for extension over other type of images.

We propose an automated general purpose Image Transfiguration model-GreenGAN which learns using only unpaired images. The model is trained to transfigure the image by changing its background, replicating the working of a green-screen video edit, where one can add any background to the object present in the frame of video. The generative model is formulated in such a manner that it requires only the new background to be provided along with the image containing object. In this manner the framework only uses weak labels-0/1, denoting the corresponding domain(object/background) of the input image, hence reducing the manual effort to a bare minimum.

Our contribution is as follows:

- We propose a general purpose Image Transfiguration model- GreenGAN, which is data independent, thus a single trained model is able to work for any type of images.
- The model learns in a weakly supervised manner using only unpaired images (i.e. unrelated object and background image pairs) thus reducing the human effort to a large extent.

In this report we provide a comprehensive literature review of some recent GAN architectures which propose to improve the qualitative performance of the GANs as well as improve the training aspect of the network, we also provide a literature review of some GAN performance metrics. We use the approaches proposed by recent GAN framework in the basic GreenGAN model to further improve the model performance and present a complete GreenGAN framework.

1.2 Organisation of the Report

The rest of this report has been organized as follows:

1. Chapter-2: **Literature review** describes the work that has been done in the field of Image Transfiguration by other researchers as well as covers the recent state-of-the-art GAN architectures. Their complete working has been presented comprehensively along with certain aspects of the models that make them perform better. We also provide a brief over the various performance metrics that can be used to quantify the model.

2. Chapter-3: **Model Formulation** goes through the step by step architecture of the proposed method. It also provides the complete mathematical details of each loss function introduced.

3. Chapter-4: **Experimental Details** describes in detail about the experimental aspects of model architecture along with the dataset used by the model.

4. Chapter-5: **Result and Discussion** showcase the comparative study of result produced by the different approaches mentioned above.

5. Chapter-5: **Conclusion and Future Work** presents in brief the conclusion of the work along with the future prospects of the model.

Chapter 2

Literature Review

2.1 Image Transfiguration

2.1.1 Image Segmentation

Image Segmentation is the task of computer vision which deals with extracting spatial information from the image by pixel-wise labelling it. The goal is to partition an image into segments which can be used for analysis. The image can be segmented to binary classify the image into foreground and background or can be used to classify particular classes. It has a wide range of application like Medical Imaging, Object Detection, Object Recognition, Autonomous driving, etc.

Better Foreground Segmentation Through Graph Cuts[11]

The paper is one of the early papers that present foreground segmentation approach with the help of a graph based approach. It explores the use of minimum graph cut algorithm to do the same. It proposes to build a graph of the image with each image pixel forming a vertex of the graph and two additional pixels for foreground and background. The weights between the vertices determine how well the pixel is associated with neighbouring pixels. After the formation of the graph they apply standard

graph flow methods to find an optimal cut separating foreground from background. The approach produces cleaner and accurate segmentation without training data but suffers from slow speed due to high computation for each image. The approach also enables the usage only for separation of foreground from the image discarding the background data.

Fully Convolutional Networks for Semantic Segmentation[1]

The paper by Long et al. is one of the first to explore the usage of convolutional Neural Nets for the task of Semantic Segmentation. It proposes a network with upsampling and de-convolutional layers to produce a segment of the input image. They experiment on the PASCAL-VOC[12] semantic segmentation challenge. They introduce the aspect of combining what and where by fusing the classes learnt from high level features with the spatial orientation obtained from low level features of convolution neural nets. The results showcase significant improvement from earlier approaches but suffers from excessive training data requirement and usage restricted on particular object categories on which it is trained.

A Deep Convolutional Neural Network for Background Subtraction[13]

The paper proposes the use of Convolution Neural Nets for the task of Background Separation to extract necessary and meaningful information from the input data. The paper follows a supervised learning approach where they train the network using comparison of image-background pairs. To counter the lack of training data the paper uses the approach of patch training. Since the paper focus on extracting foreground objects present in a video they use the approach of extracting background images from the video itself for training(taking into assumption that each background pixel is visible for at least a threshold). The model works by first producing patches of the input frame of video and then using the corresponding background patch to extract

the object by passing it through a trained classifier CNN that segments each pixel. After doing it for each patch they post-process to combine the image and apply spatial-median filtering to get rid of outliers in the segmentation map and to perform blob smoothing.The paper successfully proposes a model which is independent of the scene and can be applied to any scene but requires a video or continuous image frames to extract background from the image and then apply segmentation which restricts its usage to only video scenes.

2.1.2 Image-to-Image Translation

Image-to-Image Translation refers to task of translating one possible scene representation into another[6]. Many computer vision task can be defined as translation of input image into the required output domain.

Image Style Transfer Using Convolutional Neural Networks

[14] The paper by Gatys et al. is one of the earliest works that demonstrate the use of convolutional neural nets for the transfer of image from one domain to another. They use the higher level feature representation obtained through the learning phase of Convolutional Nets to pose the problem of style transfer as a texture transfer problem. They formulate the content loss as squared error loss between feature representation of two images. Similarly, they formulate the style loss between two images.

To transfer the style of an artwork onto a photograph they synthesize a new image and simultaneously match the content and style representation through the total loss.

The paper opened many new aspects with deep neural nets but suffered from a major restriction due to the fact that speed of the image synthesis process is highly dependent on the resolution of the image as convolution neural net features and dimensionality of optimization grow linearly.

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Network[15]

The paper focuses on the challenge of training Image-to-Image translation in domains where there are no paired training data available. The goal proposed is to learn a mapping G that translates images from the source domain X to the target domain Y , i.e. $G : X \rightarrow Y$ They propose a model which learns by leveraging the cycle consistency, i.e. there should also exists an inverse mapping F such that $F : Y \rightarrow X$ and $F(G(X)) \approx X$. This is done so as to prevent formulating two functions that contradict each other, since the two domains are unpaired and thus there can, exist several mapping between the two.

The paper proposes a Generative Adversarial Network model architecture to learns the two mapping functions. These mapping functions act as two Generators which try to learn the pattern of the domains and convert one into another. The model also includes two adversarial Discriminators and where the discriminators try to distinguish between the original image and new generated image for that domain. The Generators and the discriminators compete against each other in a zero-sum game framework along with the cycle consistency to learn the two mappings.

2.1.3 Image Editing

Fast Face-swap Using Convolutional Neural Networks[3]

The paper proposes to solve the problem of face swapping(where an individual's face is transformed into a target individual's face preserving pose, facial expression and lighting) using Convolution Neural Nets inspired from the work of Gatys et al[14](style transfer). It proposes a supervised learning approach where they train separate network for every target image they wish to impose on the input image using the approaches proposed by Ulyanov et al[16], Johnson et al[17] and Li[18].

The paper itself points out the limitation of their approach i.e. the need to have a large collection of images of the target identity for training. It also focuses on matching the initial pose and facial expression of the input individual which reduces the evaluation of the work to subjective perspective only. Finally the approach suffers from lack of training data which is found to be the issue of most of the paper following supervised training for generative purposes.

GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data[19]

The paper proposes a Generative Adversarial Model for the task of Object Transfiguration. Object Transfiguration involves replacing of an object in an image with another object from a secondary image or adding of an object from one image to another. The paper uses the cycleGAN as their base network and modifies the model architecture for the object transfiguration problem. The paper proposes to generate an object feature vector from a single image which can then be transplanted to other images. They propose an Encoder-Decoder architecture such that the Encoder decomposes an image with the object into two feature vectors (background feature and object feature) and a decoder that combines a background feature and an object feature to produce an image. Alongside the adversarial loss and cyclic consistency loss(reconstruction loss), the paper also adds another loss called nulling loss such that the object output of the encoder does not contain background information.

2.2 Generative Adversarial Models

The following are a few recent Generative Adversarial Model architectures that propose modifications over the basic GAN to improve the qualitative performance of the model as well improve the training aspect of the models.

Wasserstein GAN[20]

The paper proposes a new training algorithm for the Generative adversarial Networks which uses modified Wasserstein-1 or Earth-Mover distance constraints as losses to the model. In comparison to the basic GAN losses that suffer from vanishing gradient the paper proposes weight clipping technique in-order to have them lie in a constrained space after each gradient update, thus making sure that the network learns with smooth gradient everywhere.

The paper showcase their proposed algorithm soundness through mathematical proofs and also present an elaborate experimental analysis. The proposed method shows an improvement in the stability of the learning as well as removes the problem of mode collapse.

CartoonGAN: Generative Adversarial Networks for Photo Cartoonization [21]

The paper focuses on the problem of converting a real-world image into a cartoon style image. The approach provided deals with the two bottlenecks of such transformations. First the cartoon images tend to have high level simplification and abstractness in them as compared to the real-world images, second the cartoon images tend to contain clearer edges, smoother color shading and relatively simpler textures as compared to their real-world counterpart. The paper for this purpose proposes two novel loss functions

- A semantic content loss which promotes the capturing of only high level semantic aspects while ignoring the low level. Thus promoting a simplification of the image
- An edge promoting loss which prevents the loss of clear edges.

The proposed model also present an initialization phase where they train only their

Generator network by making it reconstruct the input image. This initialization phase as reported by them results in a better convergence and thus improves the qualitative aspect of the model.

WESPE: Weakly Supervised Photo Enhancer for Digital Cameras[22]

The paper proposes a Generative model which can translate photos taken by low quality cameras into DSLR quality images automatically. The proposed architecture trains using a weakly supervised method under which it only requires photos of two distinct domains. The paper constrains the translation using many loss functions: content consistency loss, adversarial color loss, adversarial texture loss and total variation loss. The model also proposes an extensive evaluation scheme where they mimic the rating of the users of Flickr through a supervised CNN.

SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis[23]

The paper propose a GAN architecture which can convert human drawn sketch images into real images. The model introduces a new network module called Masked Residual Unit(MRU) which allows a ConvNet layer to be repeatedly conditioned on input image. The complete architecture of the model is formulated in an encoder-decoder fashion with MRU blocks. The major problem the architecture suffers is the lack of human input as the complete architecture is automated.

Deep Photo Enhancer: Unpaired Learning for Image Enhancement from Photographs with GANs[24]

The paper focuses on the image translation problem of enhancing the camera clicked images in an automated fashion. The paper uses a modified U-net architecture which they call as gobal U-net. Along with this they use WGAN loss to improve the convergence of their model. They perform a user study to metricize the output images.

2.2.1 Performance Metric

Inception Score

Inception Score is one of the most widely used performance metric for GANs, It was proposed by Salimans [25] and uses the well known inception network [26] to calculate the performance metric. The inception network used is pretrained over the ImageNet dataset [9]. The score is calculated as follows:

$$IS(G) = exp(\ E_{x \sim P_g}\ KL[\ (p(y/x)\ ||\ p(y)\ ]\ ) \tag{2.1}$$

where $x \sim P_g$ implies that the image belongs to the generated images,
$KL[A||B]$ implies the KL-divergence between the two distribution A and B,
p(y/x) denotes the distribution of x as predicted by the model i.e. the conditional class distribution and $p(y) = \int_x p(y/x)P_g$ ie the marginal class distribution.

Though Inception Score is widely used as a performance metric it has some serious limitations since it fails to detect mode collapse and overfitting of the generator network which are quite common issues of the GAN training.

Fréchet Inception Distance (FID)

This metric was introduced recently by Heusel et al in [27]. The metric uses a specific layer of inception network to convert the pixel space to feature space vectors. These feature vectors are assumed as Gaussian random variables and is thus used to compute the Frechet distance with the help of calculated mean and covariance of the two distributions. The distance is given as follows:

$$FID(r, g) = ||\mu_r - \mu_g|| + Tr(Cov_r + Cov_g - 2(Cov_r Cov_g)^{\frac{1}{2}}) \tag{2.2}$$

where u_r, Cov_r is the mean and covariance of real image space, and u_g, Cov_g is the mean and covariance of generated images distribution.

FID score is better than the Inception Score(IS) since it is able to detect mode collapse and very well justify the human evaluation. Though since FID assumes Gaussian distribution it cannot be used on each distribution space [28].

Mean Opinion Score

Mean Opinion Score (MOS) is a way to measure the qualitative performance of a generative model using human evaluation. Each person scores the output on a range on 1-5 describing the output quality as bad, poor, fair, good and excellent. The score is then averaged to get a Mean opinion score. Since GANs are generative in nature using human evaluation as metric is quite common, but this methods suffers from issues like biasness, difficulty in reproduction and comparison and high resource requirement.

1-Nearest Neighbour Classifier

1-Nearest Neighbour Classifier (1NN)[29] can help to evaluate the model performance by having two different distributions $X_r \sim P_r^m$ and $X_g \sim P_g^n$ and computing Leave one out (LOO) accuracy of 1NN classifier by assuming one distribution as positive class and the other as negative class. In perfect mapping scenario where the generated samples are able to well match the real sample distribution the classier results in the accuracy of 0.5. In case of overfitting the model results in lower accuracy value than 0.5 since each real image finds more generated samples in its surrounding than real samples. In worst case, when then generated distribution exactly encompass the real distribution point by point this performance metric will tend to zero. While for generated images since the samples surrounding it are majorly other generated images it tend to have a higher accuracy value. In case of mode collapse this value will tend

towards one

We thus compute two different 1NN scores one for generated images and one for real images to capture if overfitting or mode collapse occur.

Chapter 3

Formulation

3.1 Basic GreenGAN

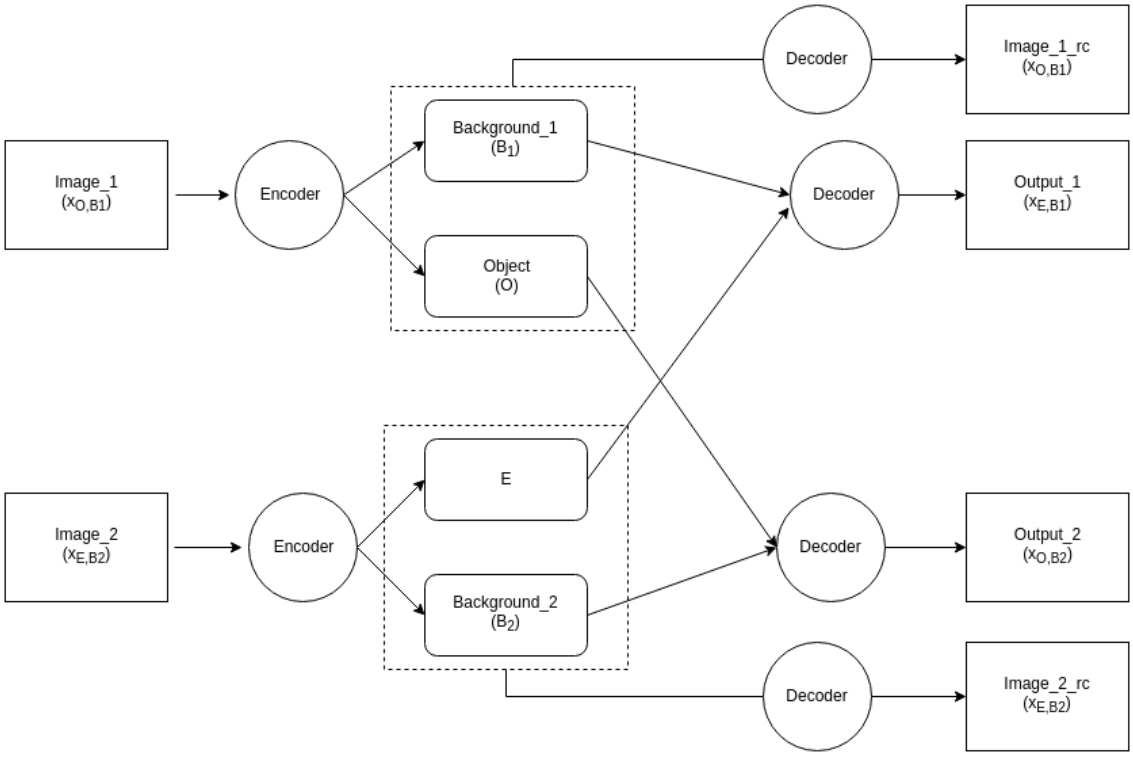

Figure 3.1 Block Diagram of Proposed **GreenGAN** model

The proposed model GreenGAN is based on the simple image transfiguration

assumption that image is comprised of object and background pair which can be separated out through mathematical computation[30, 31]. The model is formulated in such a manner that it learns to separate and combine the object and background pair as two functions which are inverse of one other as shown in the Equation 3.1. Once we are able to learn to separate out the object and background pairs present in an image, the model automatically learns to combine any two object and background pairs that are obtained from two different images and thus generate new images.

$$C(E(I)) = I \text{ and } E(C(O, B)) = O, B \tag{3.1}$$

where E and C are the functions corresponding to separation and combination of object(O) background(B) pair and I refers to the image.

We use the approach developed by Zhou et al. [19] to separate the object and background pair from the image by using an encoder-decoder type generative adversarial model. The encoder first encodes the image in an object and background pair $E : I \rightarrow O, B$.

The decoder then combines the object and background pair $C : O, B \rightarrow I$ obtained both from the same image and from two separate images. Using this approach the model generates two original images and two newer images. These images are then compared with the original images to train the network. The block diagram of the proposed model is illustrated in Figure 3.1.

To train the model using only the domain labels (object/background) of the image, we use the approach of [15] and use cyclic reconstruction loss (Equation 3.2) which reduce the discrepancy between the original and regenerated images. Alongside this, standard GAN [32] losses (Equation 3.3 and Equation 3.4) are used to measure the quality of the image generated and produce perfect transfiguration. Finally, nulling

loss(Equation 3.5) is used to ensure that images from background domain d_2 does not constitute any object while applying encoding.

All the loss functions used for training the Encoder (E) and Decoder (C) are as follows:

$$L_{reconstruction} = ||I_{generated} - I_{original}||_1 \tag{3.2}$$

$$\begin{aligned} L^0_{GAN} &= -E_{z\sim P_0}||logD(x_{\epsilon B_1}, z)|| \\ L^{\neq 0}_{GAN} &= -E_{z\sim P_{\neq 0}}||logD(x_{OB_2}, z)|| \end{aligned} \tag{3.3}$$

$$\begin{aligned} L^{\neq 0}_D &= -\big\{E_{z\sim P_{\neq 0}}||logD(x_{OB_1}, z)|| + E_{z\sim P_{\neq 0}}||log(1 - D(x_{OB_2}, z))||\big\} \\ L^0_D &= -\big\{E_{z\sim P_0}||logD(x_{\epsilon B_2}, z)|| + E_{z\sim P_0}||log(1 - D(x_{\epsilon B_1}, z))||\big\} \end{aligned} \tag{3.4}$$

$$L_0 = ||\epsilon||_1 \tag{3.5}$$

where:

$I_{generated}$ *is the generated image*

$I_{original}$ *refers to the original image*

P_0 *refers to the domain* d_2 *containing no object and* $P_{\neq 0}$ *refers to domain* d_1

B_1 *and* B_2 *refers to background obtained from domains* d_1 *and* d_2 *respectively*

x_{OB} *refers to image generated by combining object(*O*) and background(*B*)*

D *refers to the Discriminator*

3.2 GreenGAN + WGAN

We update our GAN loss function to WGAN[20] losses in-order to have a faster convergence and better stability. The updated losses are as follows:

$$\begin{aligned} L^{0}_{GAN} &= -E_{z\sim P_0}||D(x_{\epsilon B_1}, z)|| \\ L^{\neq 0}_{GAN} &= -E_{z\sim P_{\neq 0}}||D(x_{OB_2}, z)|| \end{aligned} \tag{3.6}$$

$$\begin{aligned} L^{\neq 0}_{D} &= -\big\{E_{z\sim P_{\neq 0}}||D(x_{OB_1}, z)|| - E_{z\sim P_{\neq 0}}||D(x_{OB_2}, z)||\big\} \\ L^{0}_{D} &= -\big\{E_{z\sim P_0}||logD(x_{\epsilon B_2}, z)|| - E_{z\sim P_0}||D(x_{\epsilon B_1}, z)||\big\} \end{aligned} \tag{3.7}$$

In order to enforce the constraint to have smooth gradient the WGAN applies the clipping of weights as follows after every gradient update

$$w = clip(w, -c, c) \tag{3.8}$$

where:

w *are the weights of the Discriminator* D*,*

c *constraint hyper-parameter*

3.3 Initialization Phase

Generative Adversarial Models suffer from sub-optimal results on training with randomly initialized weights. We use the approach proposed by the authors of CartoonGAN[21] where they add an initialization phase for the generator network. The generator network is thus pre-trained alone for a few epochs in the beginning to reproduce the given original image after converting it into a sample space vector using only the L_{GAN} losses, this reconstruction allows the network to converge faster and to produce images that are of better quality than that produced without initialization as shown in results.

3.4 Post-Processing

To improve the quality of the images generated using our proposed GreenGAN model we also add an image de-noising post-processing method. We use Non-local De-noising technique where we try to find window patches that are similar to our primary window patch, then we average the pixel values over these window patches and replace our primary pixel values with this average pixel value found. The idea is similar to de-noising a complete image using a set of similar images and averaging the pixel values.

Chapter 4

Experimental Details

4.1 GreenGAN

4.1.1 Dataset Used

The dataset used has been divided into two completely independent domains:

1. d_1: Object centred images

2. d_2: Images containing only background/Images containing a scene.

The domains being independent implies that the background/scene images obtained from domain d_2 do not directly correspond/match with the background/scene present in the images obtained from the domain d_1. This is done in order to remove any human effort while collecting the dataset for training. Figure 4.1 and Figure 4.2 showcase the two domains of images being used. The dataset has been accumulated from [33], [34] and [35]. A total of 1.4k object and 1.4k background/scene images are grouped together.

4.1.2 Architecture Details

The whole Encoder-Decoder Generative Adversarial model is based on the experimental details from [19] and [15]. The encoder is a 3 layered Convolutional Neural

Figure 4.1 Domain d_1: Object centred images

Figure 4.2 Domain d_2: Images containing only background/Images containing a scene

Network with leaky ReLU [36] non-linearity activation after every layer. The layers are stacked as follows:

- conv1 : 4x4 filter, 128 feature maps, 2x2 stride
- conv2 : 4x4 filter, 256 filter maps, 2x2 stride with batch normalization
- conv3 : 4x4 filter, 512 filter maps, 2x2 stride with batch normalization

The object 'O' and background 'B' are separated from the last layer by simple threshold of pixel position with a fixed value of 0.5. The decoder is made completely opposite of encoder with fractional stride so that it can mimic the inverse behaviour correctly. The discriminator used for adversarial training is also a Convolutional Neural Net with 4 layers and leaky ReLU non-linearity activation after every layer. The complete layer structure of the same is as follows:

- conv1 : 4x4 filter, 128 feature maps, 2x2 stride
- conv2 : 4x4 filter, 256 feature maps, 2x2 stride with batch normalization
- conv3 : 4x4 filter, 256 feature maps, 2x2 stride with batch normalization
- conv4 : 4x4 filter, 256 feature maps, 2x2 stride with batch normalization
- fc : a fully connected layer with sigmoid activation

The model takes images with shape (64,64,3) and are trained in batches of 64 with a learning rate of 5e-5 and a weight decay of 5e-5 for both generator and discriminator throughout the learning phase using RMSProp optimization [37].

Before beginning the complete training of the network architecture, the initialization phase[21] pre-trains the generator of the model for 50 epochs. This pre-training helps to initialize the weights of the generator which thus results in better convergence and optimal results.

For post-processing using non-local image de-noising we fix the template primary window size as 5 pixels and the search window size as 11 pixels. We use an opencv implementation of the same and adjust the hyper-parameter after ablation study.

4.2 Environment

The entire procedure is implemented in Python 2.7 on the institute server with the help of Tesla P4 GPU in order to facilitate quick computations, specially those related to deep learning. For deep learning related implementations, tensorflow along with pytorch has been used.

Chapter 5

Result and Discussion

In order to quantify the results produced by the various approaches we use various different evaluation techniques to compare the model

- **Training Time**, this metric is quantified by the number of epochs the model requires to train over the complete dataset and reach a global optimum.
- **Mean Opinion Score (MOS)**, where a group of 20 individuals are asked to score the output images on a scale of 1 to 5. 1 being poor and 5 representing high quality. And then taking the average of this score. For this score we provide each individual with randomly selected 50 output images of both the approaches.
- **1-Nearest Neighbour Classifier**, As suggested by [29] we use the 1-NN classifier as a metric evaluation method to compare the generated images. 1-NN tries to see if the generated images lies within the pixel-space of real images or not. The score lies in the range of [0,1] with an unconventional perfect score of 0.5 achieved if the two distribution of real and generated images match with each other.

On applying WGAN losses[20] as compared to the original GAN loss, the Green-

GAN model is able to converge faster and with better stability as seen with the final loss convergence. Though the training time reduces, the output generated by the two loss functions are qualitatively similar in nature which is verified by the computed MOS score of the images produced. Also the 1NN score showcase that the images generated are able to encompass well in the pixel space of the real images ie the two distributions are very close to each other.

With introduction of the initialization phase the network converges to an optimal solution and results in a better 1NN score, the MOS evaluation of the model with and without initialization phase is very close to each other. The post-processing de-noising technique helps remove the bluriness from the generated images, thus improving the quality of the image not much affecting the 1NN score.

The Figure 5.1 shows the GreenGAN output ie Base GreenGAN with WGAN loss and Figure 5.2 showcase comparison of images generated after using initialization phase against the baseline GreenGAN with WGAN losses. The Table 5.1 presents the comparison of the training epochs required to converge the approaches to an optimum and Table 5.2 presents the MOS score obtained for the images. Table 5.3 show the 1-NN score Leave one out accuracy values for random samples of images taken. Since GAN architectures tend to suffer from training issues like mode collapse and slow learning, WGAN loss thus provide a great approach to solve this issue. Initialization Phase help achieve much better 1-NN score and with added post-processing de-noising we are able to reduce the burliness introduced without affecting the 1-NN score.

Model	**Training Epochs**
Base GreenGAN	2000
GreenGAN	1200
GreenGAN + Initialization Phase	1200

Table 5.1 Comparison of training epochs for the different approaches

Model	**MOS**
Base GreenGAN	3.09
GreenGAN	3.01
GreenGAN + Initialization Phase	3.25

Table 5.2 Comparison of Mean opinion score (MOS) for the different approaches

Model	**1-NN (real)**	**1-NN (generated)**
Base GreenGAN	0.8	1
Base GreenGAN + WGAN	0.6875	1
GreenGAN + Initialization Phase	**0.5625**	**0.875**
GreenGAN + Initialization + denoising	**0.5625**	**0.875**

Table 5.3 Comparison of 1-Nearest Neighbour Leave one out accuracy for the different approaches (Value closer to 0.5 is preferred)

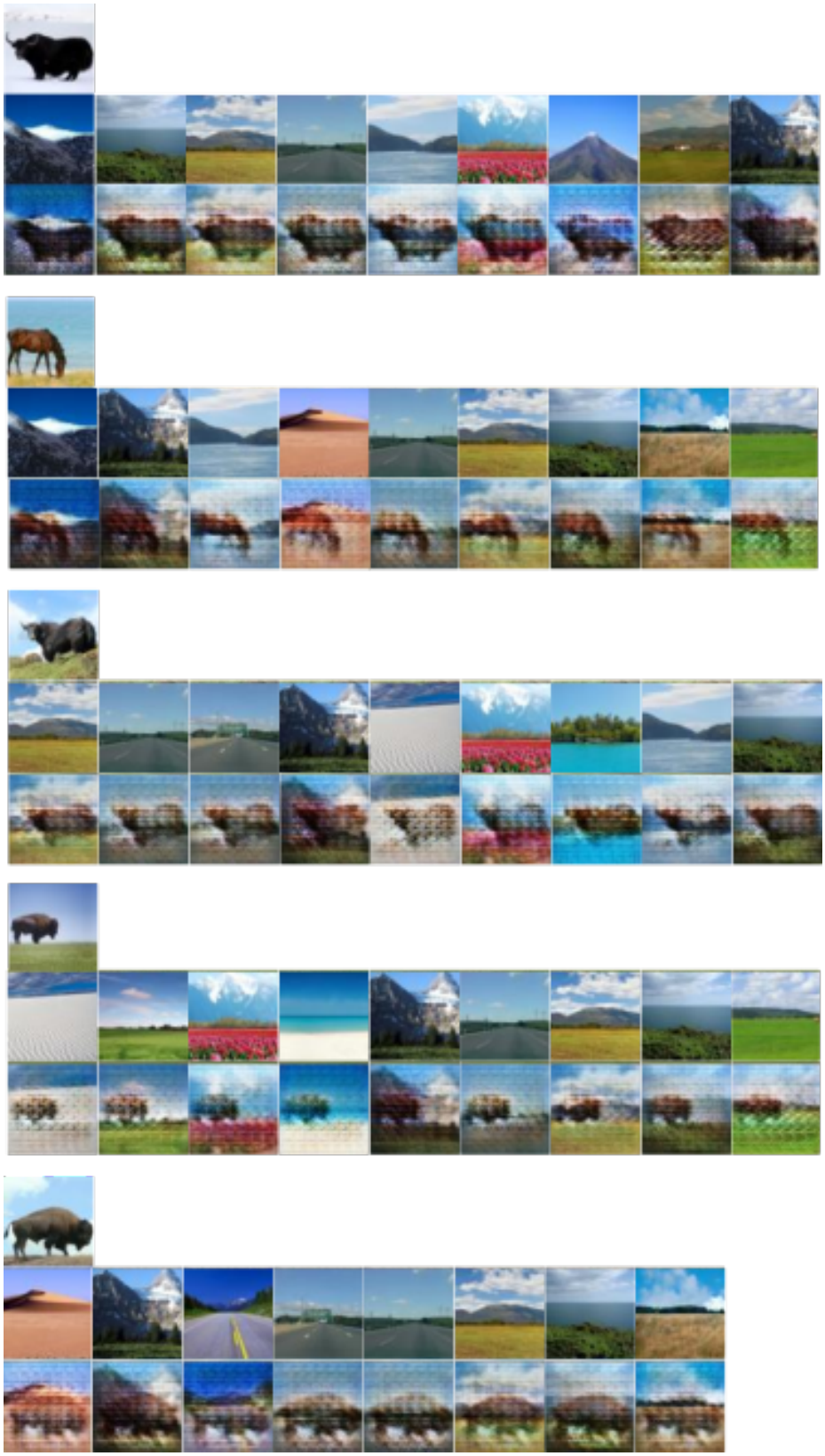

Figure 5.1 GreenGAN model: Base GreenGAN with WGAN loss

Figure 5.2 Comparison of the three different approaches a)GreenGAN + WGAN loss b)GreenGAN with initialization c)GreenGAN with initialization and de-noising

Chapter 6

Conclusions and Discussion

We propose a Generative Adversarial Network architecture, GreenGAN which acts similar to a green-screen used in post-video editing and trains entirely in a weakly supervised fashion. We propose to use WGAN [20] losses as compared to the standard GAN losses which may face mode collapse during training. We also introduce an initialization phase for generator network as suggested in [21] in order to have optimal results and help improve the performance. We present three different methods to see the performance of the different approaches, we provide both qualitative and quantitative analysis of our model and showcase that GreenGAN is able to well reproduce the real image pixel space.

6.1 Future Work

We present a weakly supervised trained model which is able to very well encompass the real image space. Since training in weakly supervised fashion has its own disadvantages like poor overall quality, one can think of using a supervised model architecture to prevent this. Another future direction that can be explored is to use a CNN for post-processing the image, one can also introduce a complete end-to-end network which includes post-processed layer.

Bibliography

[1] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1605.06211, 2016. [Online]. Available: http://arxiv.org/abs/1605.06211

[2] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for background subtraction," *CoRR*, vol. abs/1702.01731, 2017. [Online]. Available: http://arxiv.org/abs/1702.01731

[3] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," *CoRR*, vol. abs/1611.09577, 2016. [Online]. Available: http://arxiv.org/abs/1611.09577

[4] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Neural photo editing with introspective adversarial networks," *CoRR*, vol. abs/1609.07093, 2016. [Online]. Available: http://arxiv.org/abs/1609.07093

[5] J. R. Gardner, M. J. Kusner, Y. Li, P. Upchurch, K. Q. Weinberger, and J. E. Hopcroft, "Deep manifold traversal: Changing labels with convolutional features," *CoRR*, vol. abs/1511.06421, 2015. [Online]. Available: http://arxiv.org/abs/1511.06421

[6] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016. [Online]. Available: http://arxiv.org/abs/1611.07004

[7] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *CoRR*, vol. abs/1603.08511, 2016. [Online]. Available: http://arxiv.org/abs/1603.08511

[8] R. Zhang, J. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros, "Real-time user-guided image colorization with learned deep priors," *CoRR*, vol. abs/1705.02999, 2017. [Online]. Available: http://arxiv.org/abs/1705.02999

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[11] N. R. Howe and A. Deschamps, "Better foreground segmentation through graph cuts," *CoRR*, vol. cs.CV/0401017, 2004. [Online]. Available: http://arxiv.org/abs/cs.CV/0401017

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[13] M. Babaee, D. T. Dinh, and G. Rigoll, "A deep convolutional neural network for background subtraction," *CoRR*, vol. abs/1702.01731, 2017. [Online]. Available: http://arxiv.org/abs/1702.01731

[14] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015. [Online]. Available: http://arxiv.org/abs/1508.06576

[15] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017. [Online]. Available: http://arxiv.org/abs/1703.10593

[16] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," *CoRR*, vol. abs/1603.03417, 2016. [Online]. Available: http://arxiv.org/abs/1603.03417

[17] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016. [Online]. Available: http://arxiv.org/abs/1603.08155

[18] C. Li and M. Wand, "Combining markov random fields and convolutional neural networks for image synthesis," *CoRR*, vol. abs/1601.04589, 2016. [Online]. Available: http://arxiv.org/abs/1601.04589

[19] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He, "Genegan: Learning object transfiguration and attribute subspace from unpaired data," *CoRR*, vol. abs/1705.04932, 2017. [Online]. Available: http://arxiv.org/abs/1705.04932

[20] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *ArXiv e-prints*, Jan. 2017.

[21] Y. Chen, Y. Lai, and Y. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," pp. 9465–9474, 2018.

[22] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. V. Gool, "WESPE: weakly supervised photo enhancer for digital cameras," *CoRR*, vol. abs/1709.01118, 2017. [Online]. Available: http://arxiv.org/abs/1709.01118

[23] W. Chen and J. Hays, "Sketchygan: Towards diverse and realistic sketch to image synthesis," *CoRR*, vol. abs/1801.02753, 2018. [Online]. Available: http://arxiv.org/abs/1801.02753

[24] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, June 2018, pp. 6306–6314.

[25] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *CoRR*, vol. abs/1606.03498, 2016. [Online]. Available: http://arxiv.org/abs/1606.03498

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017. [Online]. Available: http://arxiv.org/abs/1706.08500

[28] A. Borji, "Pros and cons of GAN evaluation measures," *CoRR*, vol. abs/1802.03446, 2018. [Online]. Available: http://arxiv.org/abs/1802.03446

[29] Q. Xu, G. Huang, Y. Yuan, C. Guo, Y. Sun, F. Wu, and K. Q. Weinberger, "An empirical study on evaluation metrics of generative adversarial networks," *CoRR*, vol. abs/1806.07755, 2018. [Online]. Available: http://arxiv.org/abs/1806.07755

[30] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015. [Online]. Available: http://arxiv.org/abs/1511.06434

[31] P. Upchurch, J. R. Gardner, K. Bala, R. Pless, N. Snavely, and K. Q. Weinberger, "Deep feature interpolation for image content changes," *CoRR*, vol. abs/1611.05507, 2016. [Online]. Available: http://arxiv.org/abs/1611.05507

[32] 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 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[33] Caltech, "Caltech archive dataset." [Online]. Available: http://www.vision.caltech.edu/html-files/archive.html

[34] S. Dev, Y. H. Lee, and S. Winkler, "Categorization of cloud image patches using an improved texton-based approach," in *2015 IEEE International Conference on Image Processing, ICIP 2015,Quebec City, QC, Canada, September 27-30, 2015*, 2015, pp. 422–426.

[35] S. Wang, J. Joo, Y. Wang, and S. Zhu, "Weakly supervised learning for attribute localization in outdoor scenes," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3111–3118.

[36] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *CoRR*, vol. abs/1505.00853, 2015. [Online]. Available: http://arxiv.org/abs/1505.00853

[37] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. III–1139–III–1147. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3043064