Class Project Report on

Image Harmonization using Generative Adversarial Networks

Kranti Kumari: 202SP011
Nikhil Bobate: 202SP017

Under the guidance of
Dr. A. V. Narasimhadhan

Department of Electronics and Communication Engineering
National Institute of Technology, Surathkal
Karnataka-575025, India

Date of Submission: 24-01-2020

in partial fulfillment for the award of the degree
of

Master of Technology
In
Signal Processing and Machine Learning
At

Department of Electronics and Communication Engineering
National Institute of Technology Karnataka, Surathkal

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

Department of Electronics and Communication Engineering

CERTIFICATE

This is to certify that the project entitled **'Image Harmonization using Generative Adversarial Networks', submitted by Kranti Kumari: 202SP011 and Nikhil Bobate: 202SP017** is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of M.Tech in Signal Processing and Machine Learning at National Institute of Technology Karnataka,Surathkal.

Dr. A. V. Narasimhadhan
Assistant Professor
Department of ECE
NITK, Surathkal

ACKNOWLEDGEMENT

With immense pleasure we are presenting the ”Image Harmonization using Generative Adversarial Networks” Project report as a part of the curriculum of ”EC875 - Probabilistic Models in Machine Learning” under the department of “Electronics and Communication Engineering, National Institute of Technology, Karnataka”. We wish to thank all people who gave us the unending support. We express our profound thanks to our Professor, Dr. A. V. Narasimhadhan, and all those who have indirectly guided and helped us in the preparation of this project.

Contents

1. Abstract

Compositing a realistic image is a challenging task and usually requires considerable human supervision using professional image editing software. The process of improving the realism of composite results is often called harmonization. To generate realistic composites, the appearances of foreground and background need to be adjusted to make them compatible. Previous approaches to harmonize composites have focused on learning statistical relationships between hand-crafted appearance features of the foreground and background, which is unreliable especially when the contents in the two layers are vastly different. In this project, we train a convolutional neural network in an adversarial way for image harmonization.

2. Introduction

Generative adversarial network (GAN) was first proposed by Goodfellow et al. [3] to address the problem of realistic image generation from input noise variables. The key idea of GAN is to train a generator and a discriminator in an adversarial fashion. While the discriminator is trained to distinguish fake images from real ones, the generator is trained to deceive the discriminator and generate images as realistic as possible.

The GAN architecture is comprised of a generator model for outputting new plausible synthetic images, and a discriminator model that classifies images as real (from the dataset) or fake (generated). The discriminator model is updated directly, whereas the generator model is updated via the discriminator model. As such, the two models are trained simultaneously in an adversarial process where the generator seeks to better fool the discriminator and the discriminator seeks to better identify the counterfeit images.

Compositing is one of the most common operations in image editing. To generate a composite image, a foreground region in one image is extracted and combined with the background of another image. However, the appearances of the extracted foreground region may not be consistent with the new background, making the composite image unrealistic. Therefore, it is essential to adjust the appearances of the foreground region to make it compatible with the new background. Previous techniques improve the realism of composite images by transferring statistics of handcrafted features, including color and texture, between the foreground and background regions. However, these techniques do not take the contents of the composite images into account, leading to unreliable results when appearances of the foreground and background regions are vastly different.

In this project, we are training a convolutional neural network in an adversarial way for image harmonization. Given a composite image and a foreground mask as the input, our model directly outputs a harmonized image, where the contents are the same as the input but with adjusted appearances on the foreground region. The foreground appearances can be adjusted accordingly to generate a realistic composite image. Toward this end, we train two neural networks in an adversarial way to capture the context of the input image and to reconstruct the harmonized image using the learned representations.

Training an end-to-end deep CNN requires a large-scale training set including various and high-quality samples. However, unlike other image editing tasks such as image colorization and inpainting where unlimited amount of training data can be easily generated, it is relatively difficult to collect a large-scale training set for image harmonization, as generating composite images and ground truth harmonized output requires professional editing skills and a considerable amount of time.

To this end, we propose a simple yet effective way to build up a synthetic dataset which satisfies this demand which facilitates the learning process. After applying colour transformation to the ground truth images, we generate composite images which doubles the size of our dataset.

3. Related Work

Our goal is to harmonize a composite image by adjusting its foreground appearances while keeping the same background region. Generating realistic composite images requires a good match for both the appearances and contents between foreground and background regions. Existing methods use color and tone matching techniques to ensure consistent appearances, such as transferring global statistics, applying gradient domain methods, matching multi-scale statistics or utilizing semantic information. While these methods directly match appearances to generate realistic composite images, realism of the image is not considered. Different from the previous methods, our neural networks directly learn from pairs of a composite image as the input and a ground truth image, which ensures the realism of the output results.

3.1. Harmonization network

The harmonization network takes one composite image and a foreground mask as input, and performs appearance adjustments on the foreground while keeping the background unchanged. To further enhance the realism of the harmonized results, the harmonization network is trained in an adversarial way with a discriminator network, which distinguishes the disharmonious image from the harmonious ones. The Harmonization network uses U-net Architecture.

Figure 1 depicts the Architecture of our generator model:

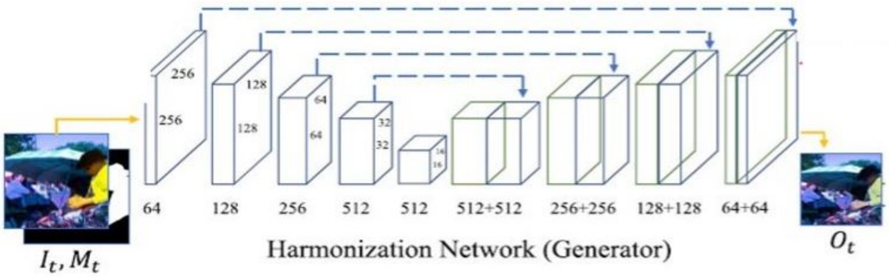


Figure 1: Architecture of the Generator model

3.2. Discriminator network

The objective of Discriminator network is to classify harmonious image from disharmonious ones, respectively, while keeping the parameters of G fixed. Both the input image and the harmonized result can be taken as the fake samples, while the ground-truth images in the dataset are taken as the real samples. The Discriminator network is a PatchGAN discriminator.

Figure 2 depicts the Architecture of our Discriminator model:

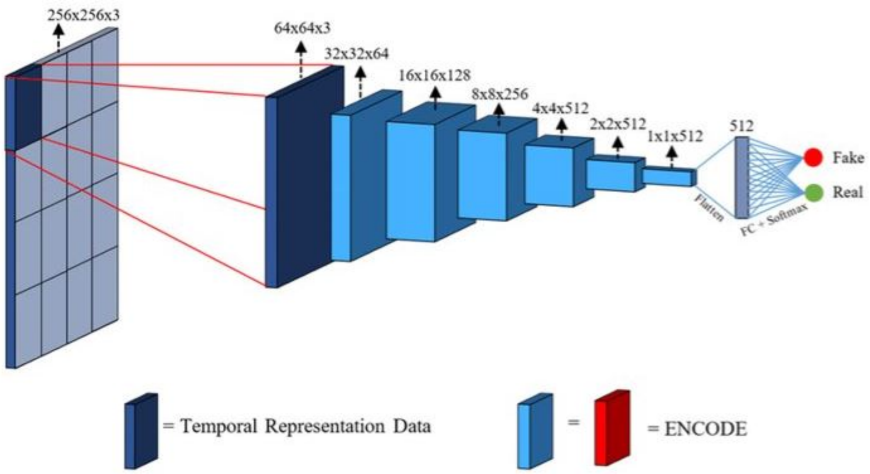


Figure 2: Architecture of the Discriminator model

3.3. A Synthetic Training Dataset

For the image harmonization task, we have collected images from the MSCOCO dataset which have ground-truth foreground masks, and then applied color transfer between random foreground pairs with the same semantic labels. While the image after a foreground adjustment is used as the input, the original image is used as the ground-truth. In this project, we simply focus on images containing people to avoid a class bias. It is easy to transfer the method to other kinds of foregrounds. For training, we have used 1000 images and for testing, we have used 10 images.

3.4. Adversarial Training

Our network contains two parts, a harmonization network G behaving as a generator and a discriminator network D. The harmonization network G processes a composite image to generate a harmonized image. At each time step t, G takes a composite image I_t and a foreground mask M_t as input and generates a harmonized output frame O_t = G(I_t , M_t), which adjusts the appearance of the foreground to make it look more natural in the background. To acquire more realistic harmonized results, we use a discriminator network D to play against G by telling disharmonious image from harmonious ones.

For generator Network, loss function to minimize is:

$$L_G = \frac{1}{N} * \parallel O_t - X_t \parallel^2 + \frac{\lambda}{N} * \parallel D(O_t) \parallel^2$$

where,

N = Total number of pixels in the input image
O_t = Harmonized frames
X_t = Ground Truth realistic image
D(O_t) = Discriminator output for Harmonized frames
lambda = Learning rate

For Discriminator Network, loss function to minimize is Binary cross entropy and a weighting is used so that updates to the model have half (0.5) the usual effect.

4. Methodology

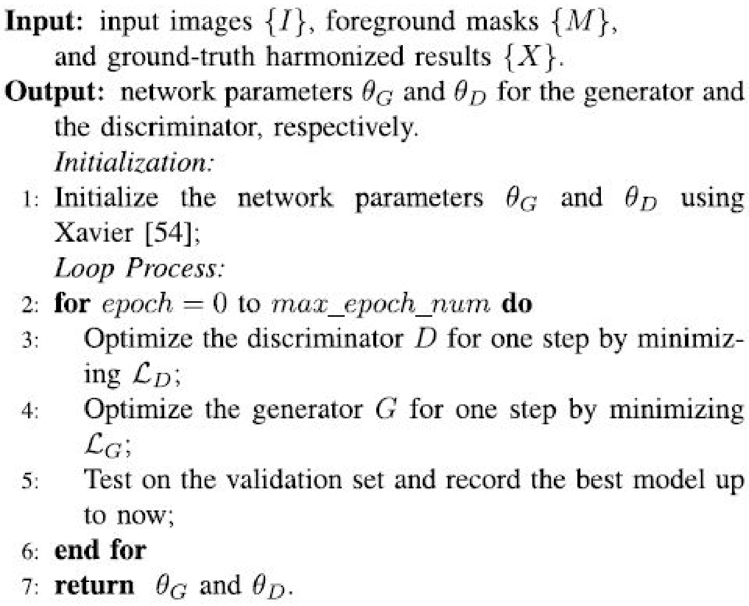
The architecture is comprised of two models: the discriminator and the generator.

Steps involved in the implementaion of our model is:

Step 1: Filter the MSCOCO dataset by altering the annotations to get only the images containing people and apply colour transform to them to get obtain composite images.

Step 2: The generator is an encoder-decoder model using a U-Net architecture. The model takes a source image and generates a target image. It does this by first downsampling or encoding the input image down to a bottleneck layer, then upsampling or decoding the bottleneck representation to the size of the output image. The U-Net architecture means that skip-connections are added between the encoding layers and the corresponding decoding layers, forming a U-shape.
The define_generator() function below implements the U-Net encoder-decoder generator model. It uses the define_encoder_block() helper function to create blocks of layers for the encoder and the decoder_block() function to create blocks of layers for the decoder. The tanh activation function is used in the output layer, meaning that pixel values in the generated image will be in the range [-1,1].

Step 3: The discriminator design is based on the effective receptive field of the model, which defines the relationship between one output of the model to the number of pixels in the input image. This is called a PatchGAN model and is carefully designed so that each output prediction of the model maps to a 70×70 square or patch of the input image. The define_discriminator() function implements the 70×70 PatchGAN discriminator. The model takes two input images that are concatenated together and predicts a patch output of predictions. The model is optimized using binary cross entropy, and a weighting is used so that updates to the model have half (0.5) the usual effect.



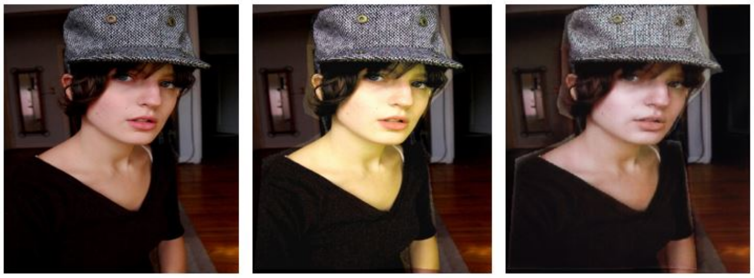
```
Input: input images $\{I\}$, foreground masks $\{M\}$,
    and ground-truth harmonized results $\{X\}$.
Output: network parameters $\theta_G$ and $\theta_D$ for the generator and
    the discriminator, respectively.
    Initialization:
    Initialize the network parameters $\theta_G$ and $\theta_D$ using
    Xavier [54];
    Loop Process:
    for $epoch = 0$ to $max\_epoch\_num$ do
        Optimize the discriminator $D$ for one step by minimizing $\mathcal{L}_D$;
        Optimize the generator $G$ for one step by minimizing $\mathcal{L}_G$;
        Test on the validation set and record the best model up
        to now;
    end for
    return $\theta_G$ and $\theta_D$.
```

Figure 3: Algorithm for image harmonization model

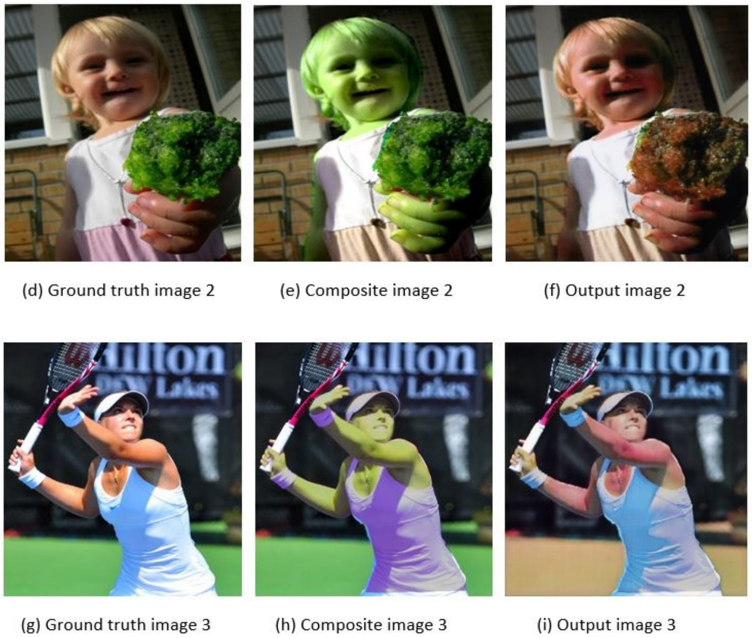
5. Results

The main goal of our project is to maintain generate Harmonized Images using Adversarial Network.

Figure 3 depicts the results obtained using our method.



(a) Ground truth image 1 (b) Composite image 1 (c) Output image 1



(d) Ground truth image 2

(e) Composite image 2

(f) Output image 2

(g) Ground truth image 3

(h) Composite image 3

(i) Output image 3

Figure 3: Output generated by our network

PSNR and MSE scores on three ground truth images is shown in table 1 below:

Ground truth Image	**MSE**	**PSNR**
Ground truth image 1	0.65	49.98
Ground truth image 2	0.53	50.88
Ground truth image 3	0.48	51.27

6. Conclusion

The proposed network contains two parts: a generator (i.e., the harmonization network) and a discriminator. The Harmonization Network outputs a harmonized image that looks more realistic. The Discriminator network distinguish disharmonious image from the harmonious ones. The two networks capture both the context and semantic information for image harmonization. To facilitate the training process, we develop an efficient method to collect large-scale and high-quality training pairs. Experimental results show that our method performs favorably on both the synthesized datasets and real composite images against other state-of the-art algorithms.

7. Future Work

We were trying to implement Video Harmonization but we were facing issues to implement the Pixel-wise discriminator which is different from Global Discriminator. It classifies harmonious pixels from disharmonious one. So, in future, we are planning to take one step further to attack the problem of video harmonization. We train a convolutional neural network in an adversarial way, exploiting a pixel-wise disharmony discriminator to achieve more realistic harmonized results and introducing a temporal loss to increase temporal consistency between consecutive harmonized frames.

8. References

[1] H. -Z. Huang, S. -Z. Xu, J. -X. Cai, W. Liu and S. -M. Hu, "Temporally Coherent Video Harmonization Using Adversarial Networks," in IEEE Transactions on Image Processing, vol. 29, pp. 214-224, 2020, doi: 10.1109/TIP.2019.2925550.

[2] Y. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu and M. Yang, "Deep Image Harmonization," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 2799-2807, doi: 10.1109/CVPR.2017.299.

[3] I. Goodfellow et al., “Generative adversarial nets,” in Proc. NIPS, 2014, pp. 2672–2680.