Generative Adversarial Network for Image Harmonization

Yichen Jiang

Abstract—In the realm of digital image processing, image harmonization plays a crucial role in seamlessly integrating foreground objects into background scenes, thereby addressing disparities in their appearance. Traditional approaches often struggle with looking for explicit models of lighting, color, and texture of background images. This study introduces an advanced solution employing Generative Adversarial Networks (GANs) that builds upon the state-of-the-art Intrinsic Compositing algorithm. By integrating GANs for image harmonization, I aim to bridge the visual gap more effectively, ensuring a coherent outcome that provides a more realistic perceptually and visually. My approach is distinguished by the introduction of three innovative loss functions within the generative model, each designed to emulate distinct characteristics of realistic images. Experimental evaluations demonstrate that two of our models outperform the baseline set by Intrinsic Compositing algorithm alone.

Index Terms—Computational Imaging, GAN, FCN, Image Harmonization

1 Introduction

In today's visually driven society, image harmonization has emerged as a pivotal technology in oceans of domains. From integrating fictional elements into real world scenes in blockbuster movies to creating convincingly realistic landscapes in video games, image harmonization played a significant role by seamlessly incorporating a foreground object into a background scene, creating a natural and realistic composition.

However, despite its widespread, achieving harmonious integration of images remains a significant challenge. The inherent differences in lighting, perspective, and texture between foreground and background elements can disrupt the visual continuity. In the past, researchers utilized rendering based methods, like image relighting algorithms, to adjust the appearance of an image or the object in an image as lit by novel illumination. Such technology can be properly adjusted to change the appearance of the foreground image according to the background lightening. However, scientists usually achieve this goal by inferring explicit illumination condition, material properties, and 3D geometry, in which the supervision for these information is hard and expensive to acquire.

Scientists gradually move to neural network as non-rendering methods, like training a classifier, as a more reliable and efficient way to overcome the gap between the foreground and background images. Following the trend, Careaga, Miangoleh, and Aksoy proposed a state-of-the-art image harmonization algorithm, Intrinsic Harmonization for Illumination-Aware Compositing (Intrinsic Compositing), which utilized two network to derive the lightening model of the background image and to perform parameterized image edits in albedo domain [1]. When given a composite image and corresponding foreground mask, such algorithm perform well in modifying shading and color of the foreground object to match those of the background. However, the harmonized foreground objects tend to lose details to some extents, and the lightening contrast would also be lowered in general.

This project aims to further improve the quality and visual authenticity of images that are preprocessed by Intrinsic Compositing algorithm, by adopting a generative adversarial network. Apart from regular generator loss, this project also experiments with three different types of generator loss terms, which are overall content loss, mask loss, and perceptual loss, aiming to help the model emulate different features of realistic images. Both the qualitive and quantative results demonstrate that the harmonized images produced by my model restore local details and are more similar to real image perceptually.

2 Related Work

2.1 Intrinsic Harmonization for Illumination-Aware Compositing

This algorithm, proposed in 2023, aims to accurately capture the lighting inconsistencies between the foreground and background found in composited images, which are largely ignored in other image harmonization researches.

In their work, image harmonization work is done in intrinsic domain, allowing them to decompose the problem into two major sub-tasks, color harmonization and relightening.

In their harmonization pipeline, they first harmonize the color content of the foreground to match the background in the albedo domain, so the color composition of the foreground object is closer to that of the background.

The very next step is to generate a shading that reflects the illumination environment in the background image. To achieve this, the algorithm first utilized a simplified parametric illumination model to estimate the background illumination condition. Then, the algorithm generate the Lambertian shading for the foreground object and composite this shading map onto the original shading of the background. Using such shading as input to the neural network, along with the RGB composite, the model would

Fig. 1. A group of ground-truth real image (left), corresponding backward adjusted synthesized image (middle), and the harmonized image processed by Intrinsic Compositing algorithm (right). As can be observed from the images, although the algorithm works pretty well in retrieving the overall color and shading of the composite image, it still needs to be proved under harsh lightening. For example, the shadow becomes less obvious and the lightening contrast is lower when compared to ground-truth image. The processed image also lose resolutions in minor details, like the words on the box

be trained to output a new shading to the foreground object, which are realistic when putting into the background [1].

2.2 Generative Adversarial Network

Generative Adversarial Networks (GANs) represent a groundbreaking class of deep learning frameworks introduced by Goodfellow et al. in 2014 [2]. The core idea of GANs is to pit two neural networks against each other in a game-theoretic scenario. One network, the generator, learns to generate data (in this project, the harmonized composite images) that is indistinguishable from real data (the ground-truth background images), while the second network, the discriminator, learns to distinguish between real data and the fake data produced by the generator. Through this adversarial process, both networks iteratively improve their performance, with the generator producing increasingly realistic data, and the discriminator becoming better at detecting fakes. This innovative approach has significant implications for a wide range of applications, including image synthesis, style transfer, and the enhancement of low-resolution images, making it a pivotal development in the field of artificial intelligence and computer vision.

2.3 iHarmony4

As the first large-scale image harmonization dataset, iHarmony4 contains four sub-datasets, HCOCO,HAdobe5k, HFlickr, and Hday2night, each of which contains synthesized composite images, foreground masks of composite images and corresponding real (ground-truth) images [3].

In this project, I use HFlickr as the dataset for training and testing the model because it is one of the dataset that are generated by backward adjustment, meaning the composite images are derived by intentionally modifying color, lightening, and texture of the masked portion in the corresponding real images [4]. Such a backward adjustment image harmonization dataset is beneficial for this project since the the ground-truth images are actual real images without any edits, which allows the network to learn features of real images to the greatest extent.

Another benefit of utilizing a backward adjustment dataset like HFlickr is the model can better derive the shading model in harsh lightening, and it also eliminates the work for shadow generation when harmonizing the images.

3 Proposed Method

The core idea of this project is to first use Intrinsic Compositing scheme to preprocess the composite images, and the preprocessed images will be passed into the generative adversarial network for further harmonization. The discriminator improves its ability to distinguish harmonized composite images from their corresponding ground-truth real counterparts through backpropogation, and the generator attempted to generate high quality harmonized image that can 'fool' the discriminator. The generator updates its weights mainly through the backpropogation of two loss terms, one associated with misleading the discriminator, and the other associated with matching the processed image with the corresponding real image.

3.1 Preprocess the Composite Images

In my approach, I commence the process of image harmonization by employing the "Intrinsic Harmonization for Illumination-Aware Compositing" algorithm [1] as a foundational preprocessing step. This advanced algorithm plays a crucial role in adjusting the lighting of the foreground object to ensure it aligns with the ambient illumination conditions of the background scene. By analyzing and modifying the intrinsic color and shading of the image, it aids in creating a more realistic and natural integration of the composited elements, providing the project with a relatively high quality starting point.

However, while this preprocessing technique significantly enhances the harmonization process, it is not without its limitations. Firstly, we have observed a notable loss of details in the foreground images post-processing, like

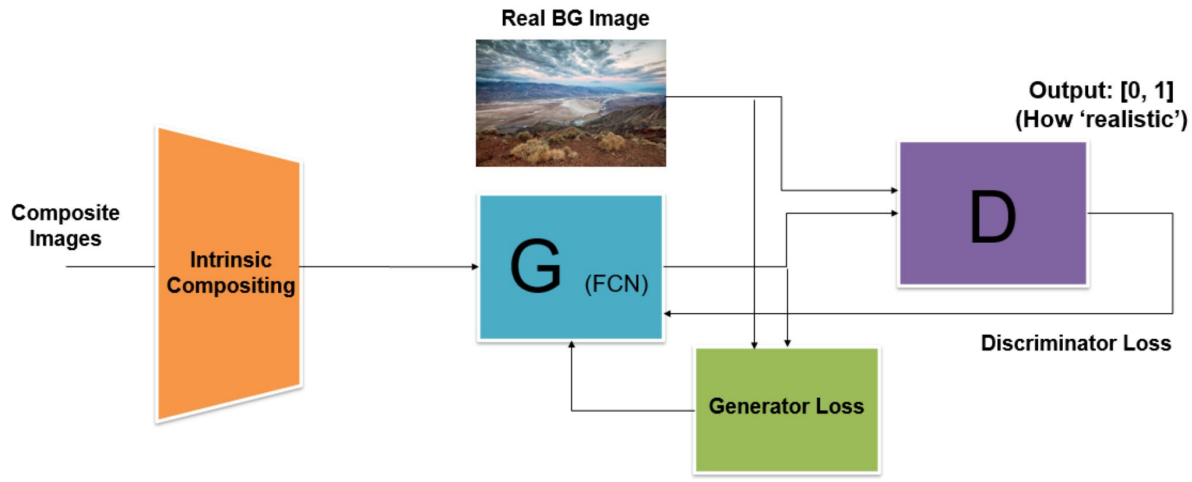

Fig. 2. The Generative Adversarial Network architecture used in this task. The original synthesized images would be preprocessed by Intrinsic Compositing algorithm first, creating semi-harmonized images with lightening and color modified to better match the background. These semi-harmonized images would be passed into GAN model proposed in this project, which include a Fully Convolutional Network (FCN) as the generator which takes in semi-harmonized images and output harmonized images. The harmonized images would be passed with their ground-truth real counterparts into a discriminator, which learns to distinguish composite images from real images. The output of the discriminator will in turn serve as a loss function term of the generator and help it produce more realistic image through backpropogation. Part of the loss function for the generator is also associated with the content difference between the generated images and the corresponding real images. In this project, I proposed three such content loss terms, pixel-wise content loss, pixel-wise content loss + mask content loss, and pixel-wise content loss + perceptual loss

the Chinese characters in the corner of the box in Fig. 1. This manifests as a blurring effect, where the crispness and clarity of the foreground elements are diminished, resulting in a less defined appearance. Such a loss of detail can detract from the overall realism and impact of the final composited image.

Secondly, the algorithm tends to alter the lighting contrast within the image. Specifically, it reduces the contrast by brightening shadowed areas and dimming regions that were originally well-lit. While this adjustment aids in blending the foreground with the new lighting conditions of the background, it inadvertently compromises the dynamic range of the foreground object. This flattening of contrast can lead to an unnatural perception of depth and material properties, especially when the foreground image was exposed to harsh lightening generated in background. For example, in Fig. 1, one of the two thin shadows in the ground-truth image is almost gone and the other one becomes less obvious when compared to the neighboring regions.

In light of these limitations, my proposed method seeks to address these challenges by introducing a Generative Adversarial Network (GAN) based framework. The objective is to retain the benefits of the "Intrinsic Harmonization for Illumination-Aware Compositing" algorithm, while mitigating its drawbacks, particularly focusing on preserving the detail and contrast of the composited images for a more realistic and visually convincing result.

3.2 GAN Model

The Generative Adversarial Network is composed of two main parts, the generator and the discriminator. The two models will improve their performance iteratively in an adversarial way, with a goal of 'defeating' each other.

A Fully Convolutional Network (FCN) is adopted as the generator due to its remarkable versatility and efficacy in image processing tasks. FCNs are favored among researchers for their ability to handle input images of various sizes, which is a requirement across all types of image-related studies such as this project, because the dimensions of samples from HFlickr is different and is hard to be predetermined. Furthermore, FCNs are famous for their ability to preserve spatial information, both globally and locally, which is favourable in the context of image harmonization problem since such a task requires the model to ensures that textures, colors, and contrasts are learned comprehensively across the whole background image. Because of the limited time to finish the project, I have to balance between the overall training time required for an model to converge and the complexity and consequent power of the model architecture. Taking such factors into consideration, the FCN used in this project consists of three downsampling layers and three upsampling layers, each followed by a ReLu function as the activation function.

The images harmonized by such a generator would be delivered into a discriminator along with the ground-truth background images. The discriminator can also handle input images of diverse dimensions, and it is also a special form of Fully Convolutional Network which outputs a single value. A sigmoid activation function would take in such value and generates a final prediction score between 0 and 1, indicating how likely any given image is real and without any modification, with 1 indicating the image is real and natural and 0 indicating the image is purely fake and synthesized.

The discriminator functions by taking in a pair of inputs: an authentic background image, which are labeled as 'real' (1), and harmonized composite images, labeled as 'fake' (0).

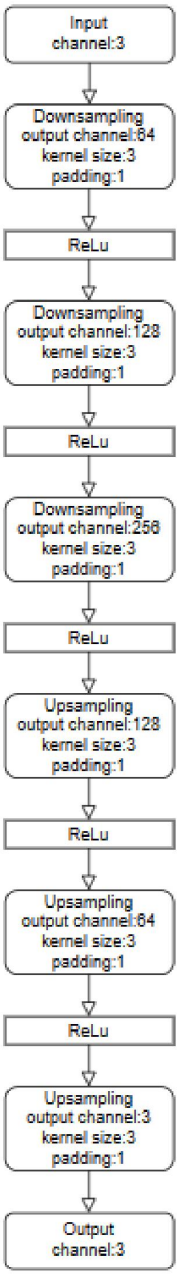

Fig. 3. Simplified FCN architecture used in the GAN model. This FCN model serves as the generator which takes in a preprocessed semi-harmonized RGB image and output a harmonized RGB image of the same dimension

As the discriminator examines each pair, it employs back-propagation to adjust its parameters, attempting to correctly classify the real images as 1 and the composite images as 0. Consequently, thru its iterative training, the discriminator improves its distinguishing prowess and becomes adept at discerning the nuances that differentiate a real image from a composite one.

The generator, on the other hand, determines its loss function partially based on the output of the discriminator, which reflects how well the harmonized images can mislead the discriminator to classify them as 'real'. More information of the loss function for the generator will be explained in detail in the next subsection since it's where the main focus of the project lies on.

3.3 Loss Function of Generator

To begin with, in GAN model, one of the most important loss terms in the generator's loss function is the adversarial loss associated with the output of the discriminator. To illustrate, let's define the Generator and Discriminator as G and D, respectively. The semi-harmonized image is denoted as x_{semi}, and the real ground-truth image is denoted as X. The harmonized image generated by G is consequently x, meaning:

$$x = G(x_{semi}) \tag{1}$$

The prediction proposed by D, when inputs are harmonized image and real image, are $D(x)$ and $D(X)$. Thus, in order to teach the generator to better mislead the discriminator when taking in harmonized images, there is a loss term which is:

$$L_{adv} = BCELoss(D(x), label = 1) \tag{2}$$

where the $BCELoss()$ stands for Binary Cross Entropy loss function, measuring the distance between two probability distributions. In this case, such loss term will iteratively train the generator to output image that will be predicted as 1 (real) by the discriminator.

Apart from such regular adversarial loss term, there should be another terms responsible for making the harmonized image as close to the original image as possible since we don't want a generator that is adept at misleading the discriminator but produces image that deviates from the ground-truth image.

To do so, I experiment with three different loss terms in this project.

3.3.1 Pixel-wise Content Loss

The first intuitive thought is to directly match the harmonized image to the original real image pixel by pixel, which will keep the other regions unmodified and preserve more details of the ground-truth image. Such content loss or identity loss, as called in other image harmonization researches [5], can be defined by:

$$L_{content} = L1Loss(G(x_{semi}), X) \tag{3}$$

The overall loss for the generator is expressed by:

$$L_G = L_{adv} + \lambda_{content} * L_{content} \tag{4}$$

where $\lambda_{content}$ represents the weighting factors of the content loss term, and in this project I choose $\lambda_{content}$ to be 100.

3.3.2 Content Loss + Mask Loss

The second experiment is similar in concept to the first one, with the exception that I add another mask loss term. Apart from the regular pixel-wise content loss mentioned in the previous experiment setting, the generator in this experiment put heavier focus on matching the masked portion of the composite image and of the real image, which indicates the location and the shape of the foreground object within it.

The mask, denoted as m is preprocessed to be gray-scale with the background portion given value 0 and foreground portion given value 1, and has the same width and height as the composite image (and the real image). The mask gets loaded along with the semi-harmonized image and the ground-truth image by the data-loader but is only used in calculating the loss function, meaning the generator still process the overall semi-harmonized image and a content loss on the unmasked portion of the image is still required to keep the unmasked pixels unchanged. The corresponding loss terms are:

$$L_{mask} = L1Loss(x * m, X * m) \tag{5}$$

and

$$L_{unmask} = L1Loss(x * (1 - m), X * (1 - m)) \tag{6}$$

The overall generator loss in this experiment is defined by:

$$L_G = L_{adv} + \lambda_{unmask} * L_{unmask} + \lambda_{mask} * L_{mask} \tag{7}$$

where λ_{unmask} and λ_{mask} represent the weighting factors of the unmasked pixels and masked pixels respectively. However, such pair of weighting factors must be carefully chosen in order to make the generator focus on the foreground object as much as possible while keep the background unmodified.

3.3.3 Content Loss + Feature Loss

In the last experiment, the generator attempts to achieve harmonization in a deeper level, instead of solely matching pixel values between images. Instead, the generator manages to produce harmonized images possessing features that resemble those of real images.

Usually in computational imaging tasks, the term "feature map" is used to describe the output generated by the convolutional layers of the network. By applying filters and convolutional kernels, the network can draw out certain patterns or features from any input such as textures, edges, or specific shapes. As an image passes through successive convolutional layers of a CNN, multiple feature maps are created, each representing different aspects of the image's information. The resulting feature maps can thus capture various perceptual and visual details of an image.

In this project, I use VGG-19 model as the feature extractor, which is a popular choice in similar tasks. The harmonized image x and its corresponding ground-truth image X get passed into the feature extractor, which is denoted as F_x, and the two corresponding feature maps will be outputted. Part of the task for the generator in each training epoch is to update its parameters to better fit the feature map of the harmonized image to that of a real image. The feature loss, or perceptual loss, is thus:

$$L_{feature} = MSELoss(F_x(x), F_x(X)) \tag{8}$$

where MSELoss, which stands for Mean Squared Error Loss, calculates the averaged squared difference between two feature maps. $F_x(x)$ and $F_x(X)$ represent the feature maps of x and X extracted, respectively.

Despite such loss term striving to match feature maps between composite and real images, the content loss is still needed because the background regions in harmonized images are expected to be unmodified and the details in the original images can be preserved to the largest extent.

The overall loss in this experiment is defined by:

$$L_{feature} = L_{adv} + \lambda_{content} * L_{content} + \lambda_{feature} * L_{feature} \tag{9}$$

where $\lambda_{content}$ still represents the weighting factor of the content loss, and $\lambda_{feature}$ represents the weighting factor of the feature loss term. In this project, I set $\lambda_{feature}$ to be 10 and $\lambda_{c}ontent$ to be 100.

4 Experimental Results

In this project, I use 8000 shuffled samples from HFlickr dataset as training set and the rest 500 samples as test set. The first two experiments are trained for 50 epochs and the third experiments are trained for 40 epochs.

The project uses two quantitative metrics to evaluate the performance of the harmonization model, which are both widely favored in image harmonization research [4]. Peak Signal-to-Noise Ratio (PSNR) calculates the ratio between the maximum possible power of a signal (in this case, the original image) and the power of corrupting noise that affects the fidelity of its representation (the harmonized image), and it's usually expressed in logarithmic decibel scale. Generally speaking, higher PSNR represents greater harmonization quality, like better restoring of local details.

Another metric used is Learned Perceptual Image Patch Similarity (LPIPS), which is used to assess the perceptual similarity between two images. Unlike PSNR, which is based on pixel-wise differences, LPIPS uses deep learning features to more closely align with human visual and perceptual perception. A lower LPIPS score indicates greater perceptual similarity as judged by the learned model, and ideally, by human observers. LPIPS is a valuable tool for evaluating image processing tasks where visual fidelity from a human perspective is crucial. In image harmonization task, where the primary goal is to create realistic images synthesized indistinguishable by humans, such metric is a significant benchmark.

One representative group of results are shown in Figure 4, and the corresponding quantitative results are shown in Table 1. The foreground object is the strawberry in the left upper corner. When comparing the semi-harmonized

Fig. 4. The first row displays a group of harmonized images generated from GAN model, along with the ground-truth and backward adjusted composite image from HFlickr. The foreground object is the left-upper strawberry. The second row shows corresponding zoomed-in images of the foreground. All three methods restore the local details and the lightening contrast of the strawberry to different extent. However, the second experiment (GAN + Mask loss) shows the worst quantitative results among the three experiments, possibly because of the sub-optimal weighting factor chosen in training

TABLE 1
Quantitative Results of the Harmonized Images

Metric	Composite Image	Intrinsic Compositing	GAN + Content Loss	GAN + Mask Loss	GAN + Feature Loss
PSNR	37.9	38.6	38.7	37.4	38.7
LPIPS Score	0.00321	0.00243	0.00221	0.00252	0.00209

image generated by Intrinsic Compositing algorithm [1] to the original real image, one can see that the local details are somehow lost (like the seeds in the strawberry), so the foreground becomes slightly "blurry". The lightening contrast is also alleviated as one can tell from the image that the lower part of the strawberry, which is in shadow in real image, becomes bright in the harmonized foreground generated by Intrinsic Compositing algorithm.

When using content loss to update the generator parameters, the local details are restored and the foreground contrast is retrieved, which can be seen from the resulting graph. The image also has slightly higher PSNR than the image generated by Intrinsic Compositing algorithm, and it also shows lower LPIPS score than Intrinsic Compositing output, indicating that the harmonized image is perceptually and visually more realistic.

However, both the qualitative and quantitative results of the second experiment, in which the generator uses a mask loss term to put more weights on learning differences between foreground regions of the two images, doesn't quite meet the expectation. One plausible reason is the poor weighting factor of the mask loss term chosen, forcing the generator to put too much focus on the foreground so the generator fails to keep the background unchanged. The harmonized image becomes generally brighter when compared to the ground-truth which is often undesirable in image harmonization task, although the foreground indeed restores local details and lightening contrast.

The third GAN model turns to be the optimal among all three experiments. As in previous two experiments, the generator including feature loss term in its loss function processes foreground object to resemble the ground-truth foreground object in both local details and lightening contrast. It also has the greatest PSNR and the lowest LPIPS distance among all the harmonized images. Such result shows that the adversarial model utilizing feature loss term produces a high quality image that provides the most visual and perceptual similarities to a natural image.

5 Conclusion

In this project, I use a generative adversarial network (GAN) model to further harmonize images that are produced by Intrinsic Compositing algorithm, a state-of-the-art technology used in image harmonization [1]. The project adapts a Fully Convolutional Network as the generator to generate harmonized image and discriminator to predict how likely any input is real and non-synthesized image. GAN works by iteratively updating the generator and the discriminator, with one of the two striving to beat the other one.

Besides the regular adversarial loss, the project also undertakes a exploration of image harmonization through the lens of three distinct experimental settings, with each setting designed to illuminate one approach of creating realistic harmonized images. The first experiment setting attempts to enforce the harmonized image to approach the real image in pixel-wise level, and the second experiment deals with the harmonization task from a more reasonable angle by adding another loss term that strategically emphasizes the masked portions of the image, thereby instructing the generator to prioritize these areas during the image harmonization process. In the last experimental setting, a pivotal adjustment was introduced to the generative model's loss function, designed to refine its focus on feature maps within the images, which are extracted by a VGG-19 feature extractor, in order to achieve a higher degree of perceptual harmony. By comparing these feature maps, the experiment sought to guide the generation process towards producing harmonized images, wherein the features closely resemble those of ground-truth images.

The ultimate results, both quantitative and qualitative, demonstrate that the GAN model succeeds in lifting the harmonization level of the composite image to some extent. However, the model under the second setting performs relatively poorly when compared to the benchmark set by Intrinsic Compositing algorithm, possibly because of the sub-optimal weighting factor chosen for mask loss term. The results of the last experiment were promising, as evidenced by the achieved higher Peak Signal-to-Noise Ratio (PSNR) and the lowest Learned Perceptual Image Patch Similarity (LPIPS) score among the tested settings, indicating an enhancement in restoring the intrinsic details of the image and in producing harmonized outputs that exhibit a high degree of visual realism.

The limitation still exists, since the harmonized images still looks fake under harsh lightening. Although the model optimizes the appearances of foreground objects by increasing its lightening contrast, the contrast level is too low in harsh lightening, especially when the foreground object has sophisticated shape and requires inclusion of shape-aware model to generate corresponding shadow within it. The model is also trained on a backward adjustment dataset so it needs not to worry about the overall shadow projected by foreground object on the background, which is a complicated and important research subject in other image harmonization tasks when using a forward adjustment dataset [4].

The future work lies on employing model of more complex and powerful architecture, like adding the skip structure in FCN or deepening the network by introducing more neural layers. Training with a larger dataset like HCOCO (containing roughly 40k training samples) [3] and for more epochs may also facilitate in optimizing the performance of the model. There are also large space in tuning the hyper-parameters (learning rate, early stopping point, and weighting factors etc.), and utilizing an optimal combination of these hyper-parameters is potentially promising in generating realistic images.

ACKNOWLEDGMENTS

The author would like to thank Professor Gordon and Cindy, who serves as my mentor in this project, for their highly helpful suggestions and all supports. Their expert guidance and insightful suggestions were invaluable to the development and completion of this project.

REFERENCES

[1] C. Careaga, S. M. H. Miangoleh, and Y. Aksoy, "Intrinsic harmonization for illumination-aware compositing," in *Proc. SIGGRAPH Asia*, 2023.

[2] I. J. 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*, 2014.

[3] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang, "Dovenet: Deep image harmonization via domain verification," in *CVPR*, 2020.

[4] L. Niu, W. Cong, L. Liu, Y. Hong, B. Zhang, J. Liang, and L. Zhang, "Making images real again: A comprehensive survey on deep image composition," 06 2021.

[5] F. Zhan, S. Lu, C. Zhang, F. Ma, and X. Xie, *Adversarial Image Composition with Auxiliary Illumination*, 02 2021, pp. 234–250.