UNPAIRED IMAGE-TO-IMAGE TRANSLATION USING CYCLE-CONSISTENT ADVERSARIAL NETWORKS

Arushi Arora, Chandana Thimmalapura Jagadeeshaiah, Pallabi Chandra
New York University
{aa10350, ct3002, pc3131}@nyu.edu

ABSTRACT

In this project, we aim to solve the problem of image-to-image translation, where usually a mapping is learned between an input image and an output image in a supervised manner. However, in many cases, such paired image data may not be available. We propose a robust solution which will work on non-corresponding pairs of data. Our goal is to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an adversarial loss. Because this mapping is highly under-constrained, we couple it with an inverse mapping $F : Y \rightarrow X$ and introduce a cycle consistency loss to enforce $F(G(X)) \approx X$ (and vice versa). We then use the method to demonstrate the results on various applications such as style transfer, object transfiguration, etc. The code for the experiments can be found here.

1 Introduction

The problem of image-to-image translation is extremely important because its applications are extremely versatile. These include the following- Style transfer, Object Transfiguration, Season transfer, Photo generation from painting, Photo enhancement (eg: low-light enhancement or superresolution), Semantic segmentation, etc.

In case paired images are available, then a straightforward one-way neural network (eg. pix2pix) [1] is sufficient to learn a mapping $G : X \rightarrow Y$ such that the distribution of images from $G(X)$ is indistinguishable from the distribution Y using an adversarial loss. However, such paired images are difficult to obtain and very expensive to annotate involving a large amount of time and effort. It may also require artistic authoring, and the outputs may also be subjective and not very well defined. Thus, it is important to solve the problem of image-to-image translation when paired images are not available.

Although we lack supervision in the form of paired examples, we can exploit supervision at the level of sets: we are given one set of images in domain X and a different set in domain Y . We may train a mapping $G : X \rightarrow Y$ such that the output $\widehat{y} = G(x), x \in X$, is indistinguishable from images $y \in Y$ by an adversary trained to classify $\widehat{y}$ apart from y. However, since the mapping may not be learned in a meaningful way, we impose a cycle consistency constraint in the learning process. If we have a translator $G : X \rightarrow Y$ and another translator $F : Y \rightarrow X$, then G and F should be inverses of each other, and both mappings should be bijections. We apply this structural assumption by training both the mapping G and F simultaneously, and adding a cycle consistency loss that encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$. Combining this loss with adversarial losses on domains X and Y yields our full objective for unpaired image-to-image translation.

2 Method

We solve the problem of unpaired image-to-image translation as described earlier using a conditional generative adversarial neural network (GAN) [2] . It consists of two pairs of networks consisting of the following parts- a generator, which is an autoencoder that generates the target image from the given input image and a discriminator, which classifies whether the generated image is close enough to the target image or not. Since the images are unpaired, the GAN used in this paper are cycle consistent which means both pairs of networks (each containing a generator and discriminator) G

and F are trained simultaneously, where the mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$ are learnt such that G and F are inverses of each other.

2.1 Loss Functions/ Objective

In this section, we will describe the objective function used to train the cycle consistent GAN. It consists of the following parts:

2.1.1 Adversarial Loss

The adversarial loss function is the log loss for classification of the generated image as to whether it is able to fool the discriminator into believing that the generated image can replace the target image. It can be denoted by the following equation for mapping G -

$$L_{GAN}(G, D_Y, X, Y) = E_{y \sim p_{data}(y)}[logD_Y(y)] + E_{x \sim p_{data}(x)}[log(1 - D_Y(G(x))]$$

Where, G denotes the mapping $X \rightarrow Y$ and D_Y is the discriminator for $G(x)$ and y.

Conversely, the adversarial loss for the mapping F -

$$L_{GAN}(F, D_X, Y, X) = E_{x \sim p_{data}(x)}[logD_X(x)] + E_{y \sim p_{data}(y)}[log(1 - D_X(F(y))]$$

where, F denotes the mapping $Y \rightarrow X$ and D_X is the discriminator for $F(y)$ and x.

2.1.2 Cycle Consistency Loss

Since the adversarial loss is insufficient to learn a two-way mapping between permutations of images, we use a cycle consistency loss. For each image x from domain X, the image translation cycle should be able to bring x back to the original image, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. We call this forward cycle consistency. Similarly, for each image y from domain Y , G and F should also satisfy backward cycle consistency: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. The equation is given by -

$$L_{cyc}(G, F) = E_{x \sim p_{data}(x)}[\| F(G(x)) - x \|_1] + E_{y \sim p_{data}(y)}[\| G(F(y)) - y \|_1]$$

2.1.3 Final Loss Function

The final loss function is a summation of the two adversarial losses and the cycle consistency loss. The equation is given by-

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda\, L_{cyc}(G, F),$$

where λ acts as a weighing parameter to decide how important each of the two losses are.

2.2 Network Architecture

In this section, we'll describe the model architecture. It consists of the following parts:

2.2.1 Generator

For the generator in each network pair G and F, we use an architecture consisting of 3 x 3 convolutions, several residual blocks, two fractionally-strided convolutions with stride 2, and one convolution that maps features to RGB. The number of blocks used depends on the size of input patches taken for training.

If the input size is 256 x 256, the generator down samples them, then up samples them back to 256 x 256 creating the generated image. The Generator Architecture is shown in 2.2.1.

2.2.2 Discriminator

The discriminator has the architecture as shown in 2.2.2

Each neuron (value) of the output tensor holds the classification result for a 70 x 70 area of the input image. Usually, the discriminator of GANs outputs one value to indicate the classification result of the input image. By returning a tensor of size 30 x 30, the discriminator checks if every 70 x 70 area — these areas overlap each other — of the input image

Layer	Activation Size
Input	3x256x256
64x7x7 conv, stride 1, pad 3	64x256x256
128x3x3 conv, stride 2, pad 1	128x128x128
256x3x3 conv, stride 2, pad 1	256x64x64
9 consecutive Residual Blocks	256x64x64
128x3x3 convTranspose, stride 2, pad 1, out_pad 1	128x128x128
64x3x3 convTranspose, stride 2, pad 1, out_pad 1	64x256x256
3x7x7 conv, stride 1, pad 3	3x256x256

Table 1: Generator Architecture

Layer	Activation Size
Input	3x256x256
64x4x4 conv, stride 2, pad 1	64x128x128
128x4x4 conv, stride 2, pad 1	128x64x64
256x4x4 conv, stride 2, pad 1	256x32x32
512x4x4 conv, stride 1, pad 1	512x31x31
1x4x4 conv, stride 1, pad 1	1x30x30

Table 2: Discriminator Architecture

seems real or fake. Doing so is equivalent to manually selecting each of these 70x70 areas and having the discriminator examine them iteratively. Finally, the classification result on the whole image is the average of classification results on the 30 x 30 values.

3 Experiments

In this section, we discuss implementation details of the training experiments. The code for this training experiment is contained within the following repository - repo

The two networks are trained pair wise using the objective loss function as described above.

In this study, we conducted experiments on several datasets, namely Summer2Winter Yosemite, Cityscapes, Facades, and Monet2Photo. Specifically, for the Summer2Winter Yosemite dataset, we performed a series of experiments using varying hyperparameters. These experiments involved modifying the learning rate, lambda values, and beta values for the optimizers.

We observe that the model performs best on the following hyperparameters -

Hyperparameter	Value
Lambda (λ)	10
Batch size	1
Optimizer	Adam
Learning Rate	0.0002
Learning Rate decay	Yes

Table 3: Hyperparameter Summary

4 Results & Conclusions

In the following figures, we demonstrate results on a wider range of applications.

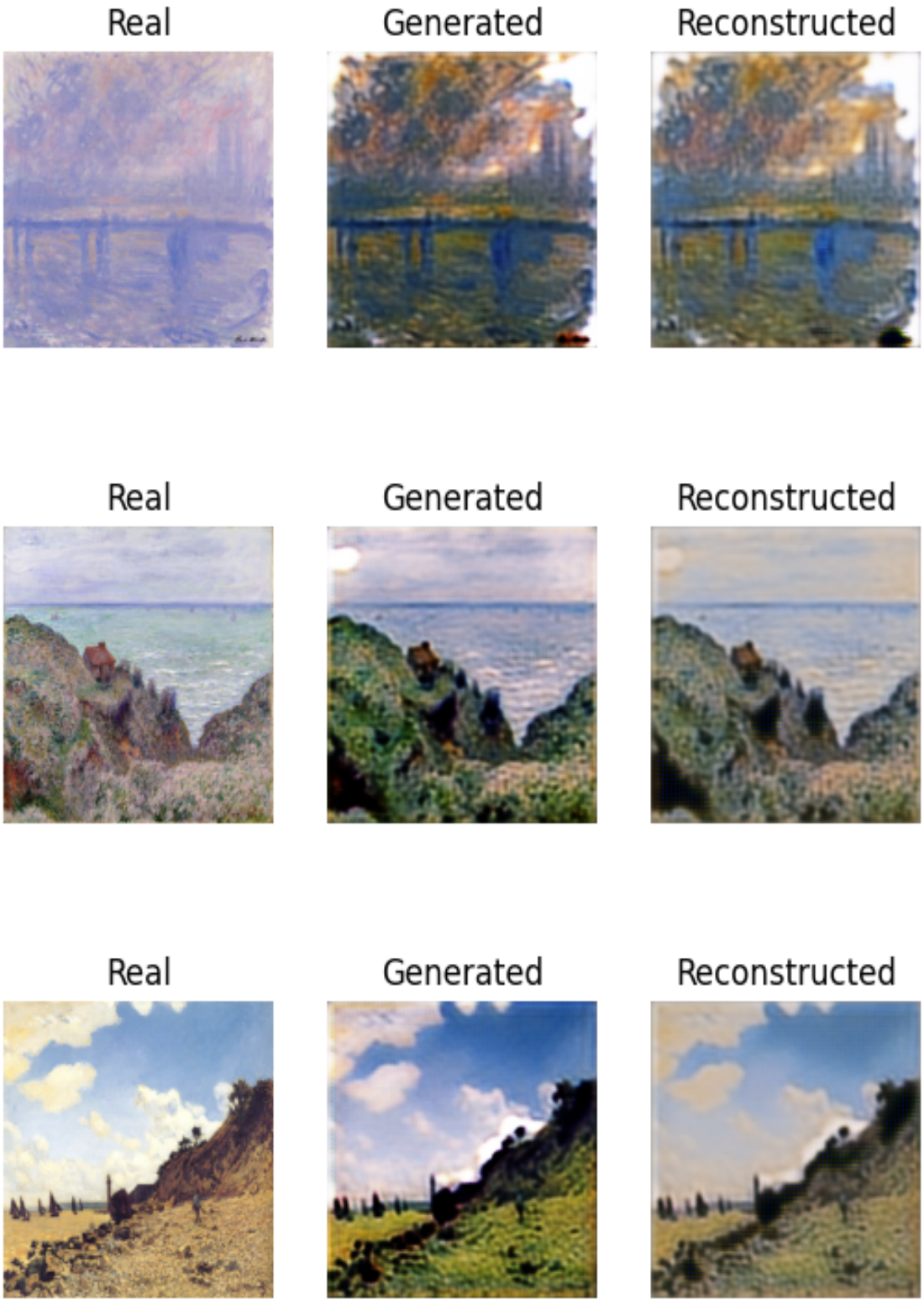

Figure 1: Style transfer Monet's paintings (Real) to a photographic style (Generated)

Figure 2: Style transfer from digital images (Real) to a Monet's painting style (Generated)

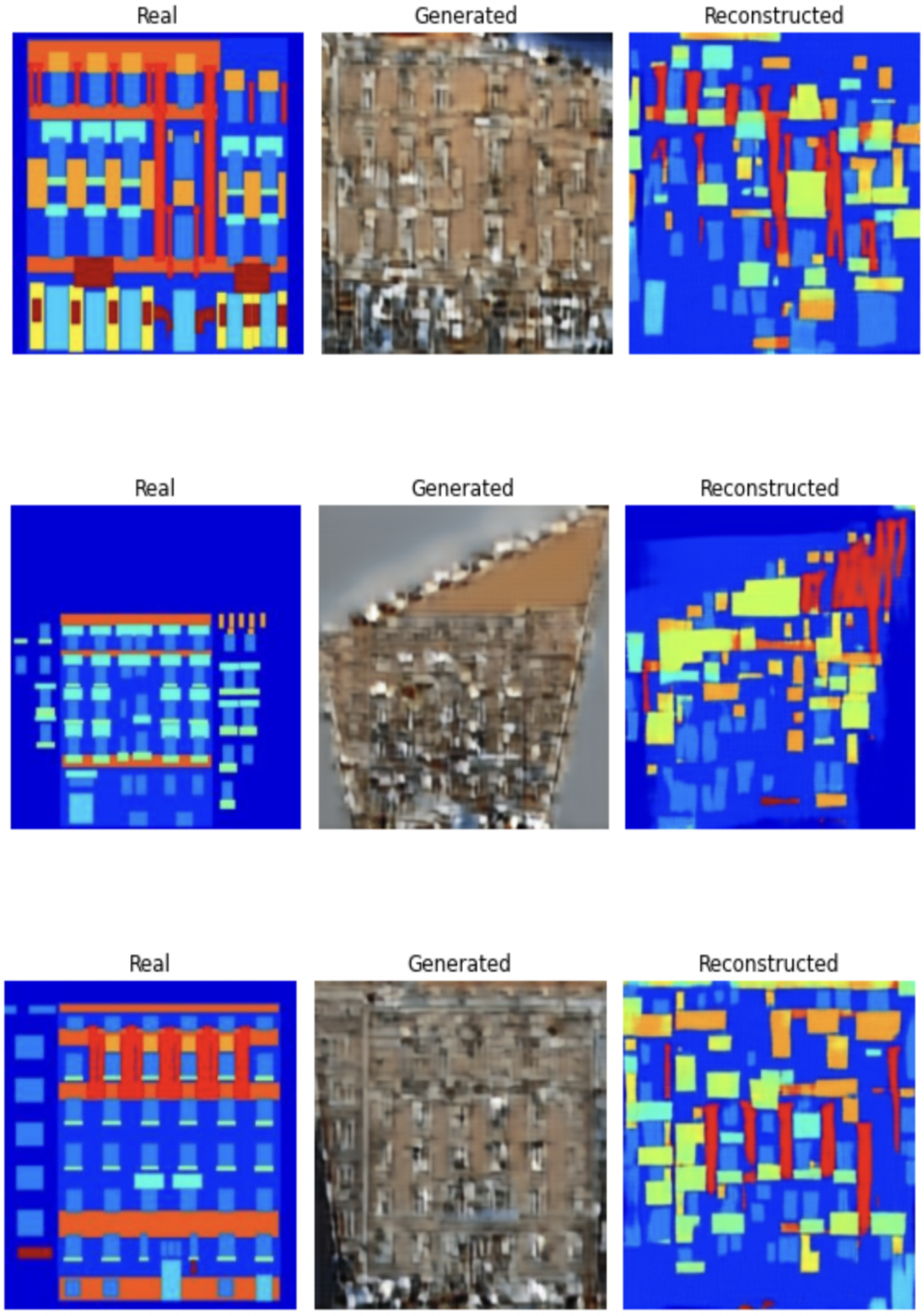

Figure 3: Transfer from label images to a Photographs (Semantic Segmentation)

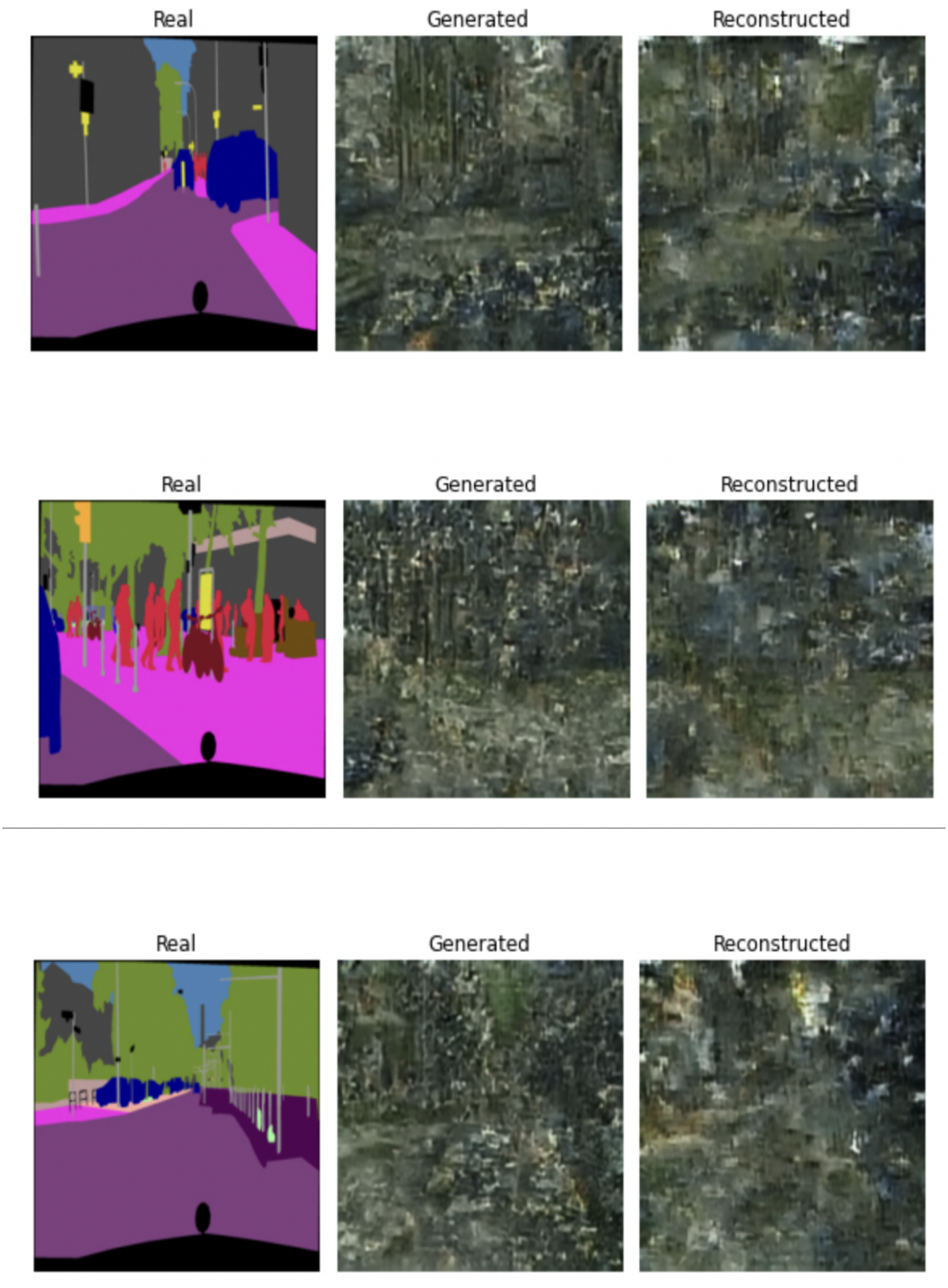

Figure 4: Transfer of Label Images of Cityscape Dataset to photographs followed by reconstruction (Semantic Segmentation)

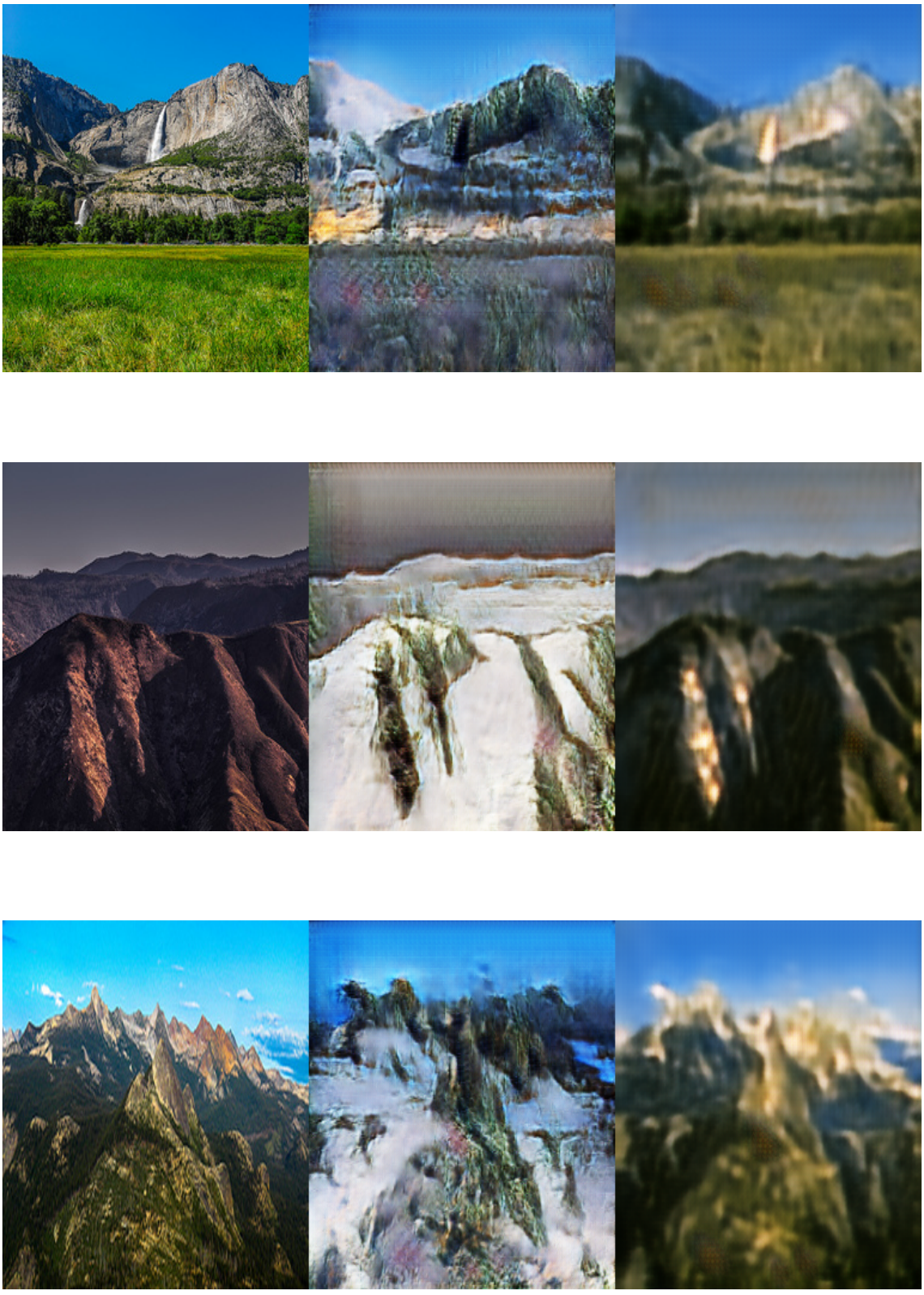

Figure 5: Season transfer from Summer images of Yosemite to Winter

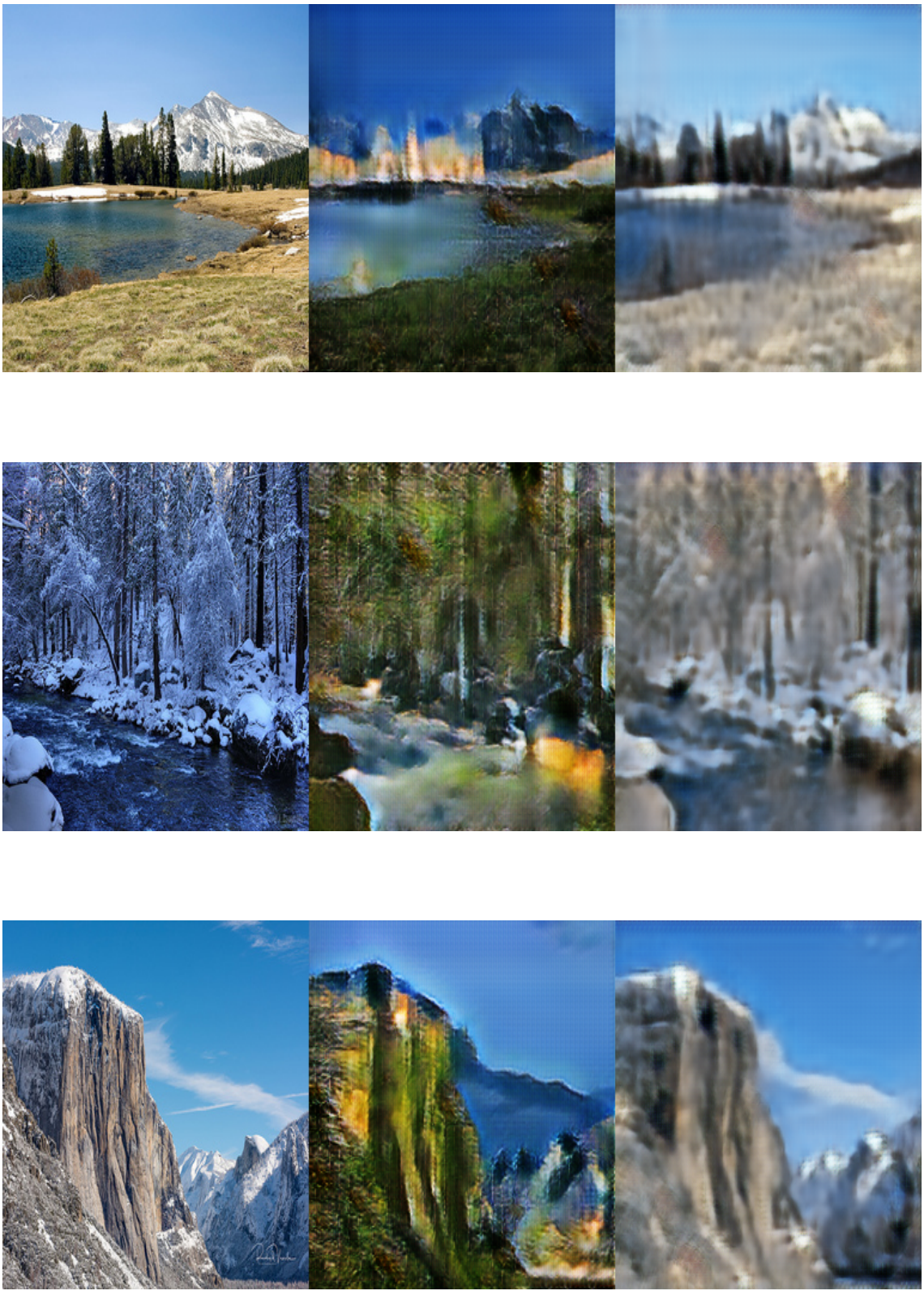

Figure 6: Season transfer from Winter images of Yosemite to Summer

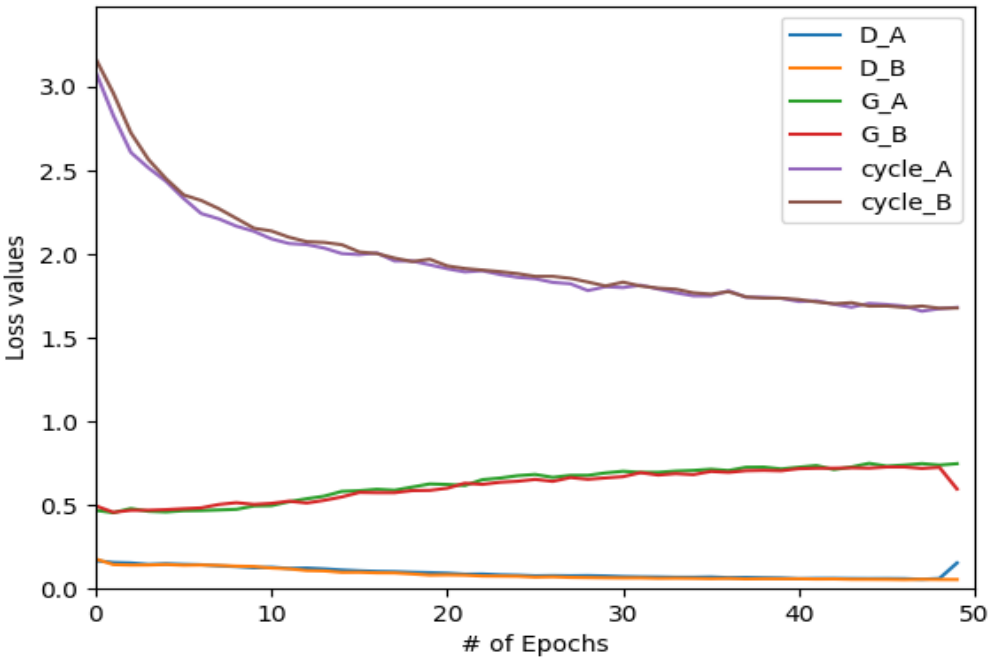

Figure 7: Plot for Loss vs Number of Epochs

The Figure 1 shows transfer from Monet's Painting to Photo to Reconstructed back to Monet's Painting style. And the Figure 2 shows transfer from photo to Monet's paintings to Reconstructed back to input image. The results were obtained after training with 300 images over 50 epochs to transfer input images into artistic styles similar to Monet's paintings.

The Figure 3 shows transfer from Label image to Photo to Reconstructed label image for the Facades Dataset. These results are later used to evaluate model performance. The Figure 4 shows results for the conversion of label images of the CityScapes dataset to photographs followed by reconstruction of the label image from the generated photograph. These results are later used for evaluation of model performance.

The results shown in Figure 5 and 6 were obtained as a result of training the model on Summer to Winter Yosmite dataset consisting of 1540 Summer Photos and 1200 Winter Photos with each split into train and test subsets. The model was trained for 50 epochs. The Figure 5 shows transfer from Summer to Winter to Reconstructed back to Summer. Whereas Figure 6 shows transfer from Winter to Summer to Reconstructed back to Winter.

The Figure 7 shows plot for different losses of the discriminator, generator, and cycle consistency over the course of training. As the epochs increase, the discriminator loss decreases as it becomes more skilled at distinguishing real and generated images and the cycle consistency loss decreases as the translation process becomes more consistent and reversible.

5 Model Evaluation

Evaluation of our generative model is difficult since the image translation is carried out on unpaired images with no pairwise supervision. Thus, for each translated image our model generates, we have no ground truth for reference. Since traditional image quality metrics such PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index) require a ground truth to quantify image quality, usage of these metrics is ruled out. Thus, we look for alternative methods to evaluate our model. We used two methods, based on the literature we studied. These are as follows-

5.1 FCN score

The FCN metric evaluates how interpretable the generated photos are according to an off-the-shelf semantic segmentation algorithm (the fully-convolutional network, FCN, from [3]). The FCN predicts a label map for a generated photo. This label map can then be compared against the input ground truth labels using standard semantic segmentation metrics such as per-pixel accuracy, per-class accuracy, etc.

The cityscapes dataset [4] is a popular dataset used for semantic segmentation of images and consists of several photographs with annotated labels associated with them. Although we don't use pairwise supervision while training the model, the availability of the reference label images for this dataset enables us to calculate an FCN score to evaluate the performance of our model. In addition to training the CycleGAN model to translate Monet's paintings into photos and vice-versa, we trained the model with the cityscapes dataset to translate the label images to photographs. We then get the label maps from these generated images using FCN and calculate the metrics as shown below for the semantic segmentation task-

Loss	**CycleGAN**	**pix2pix**
Per-pixel acc.	0.48	0.71
Per-class acc.	0.16	0.25
Class IOU	0.11	0.18

Table 4: FCN model score summary

Thus, we see that the CycleGAN performs similar to pix2pix in terms of Per-class accuracy and class IOU (intersection of union) despite being trained without pairwise supervision.

5.2 Inception Score

The Inception Score, or IS for short, is an objective metric for evaluating the quality of generated images, specifically synthetic images output by generative adversarial network models. To evaluate the performance of generative models in image synthesis, the Inception Score is commonly used. This score is named after the Inception v3 model, which was introduced in a 2015 research paper by Christian Szegedy and colleagues, titled "Rethinking the Inception Architecture for Computer Vision."[5]

The generated images are fed into a pre-trained Inception v3 model, which produces a probability distribution over a set of classes for each image. The score is then calculated by taking the exponential of the average KL divergence between the marginal distribution of class labels and the conditional distribution of class labels given the generated images.

The intuition behind the Inception Score is that diversity is captured by the entropy of the marginal distribution of class labels, while quality is measured by the expected value of the conditional distribution of class labels given the generated images. By combining these two measures, the Inception Score provides a comprehensive evaluation of the overall quality of the generated images.

Our aim is to calculate inception scores for model outputs from the CycleGAN and compare them with the corresponding outputs from one of our baseline models like Pix2Pix. To this end, we train the CycleGAN and Pix2Pix model with the facades dataset [6] to compare performance. Facades dataset consists of Building Facades and corresponding segmentations split into train and test subsets. We train both the models to perform the translation task of label images to photographs. We then use the trained model to generate the resulting fake photographs from certain test label images. Once the fake photographs are obtained from each model, we calculate the inception score on them using the standard pytorch implementation of inception score. The score is calculated on random splits of the images such that both a mean and standard deviation of the score are returned. The results are tabulated as follows-

Model	Score (Mean, Variance)
CycleGAN	(1.1008, 0.1161)
Pix2Pix	(1.3475, 0.2886)

Table 5: Inception Score Summary

As seen above, despite the fact that Pix2Pix has the advantage of pairwise supervised learning, the CycleGAN gives similar performance for the same task. In fact, although the mean is slightly lower for the CycleGAN, the reduced variance indicates that the CycleGAN has more consistent performance across all types of images.

6 Discussion & Related Work

In this section, we will discuss some of the popular approaches proposed in the field of image-to-image translation and contrast with our approach CycleGAN [7].

One of the most well-known methods for paired image-to-image translation is Pix2Pix [1], which was proposed by Isola et al. in their paper "Image-to-Image Translation with Conditional Adversarial Networks" at CVPR 2017. Pix2Pix is a conditional generative adversarial network (cGAN) that learns a mapping between an input image and an output image using a paired dataset. The generator network takes the input image and generates the corresponding output image, while the discriminator network tries to distinguish between the generated output image and the real output image. The cGAN framework ensures that the generated output image is conditioned on the input image, leading to high-quality and visually pleasing output images. While Pix2Pix is suitable for tasks where paired datasets are available and generates highly realistic images, CycleGAN is suitable for tasks where datasets are unpaired.

Another related work is UNIT [8] (Unsupervised Image-to-Image Translation Networks), proposed by Liu et al. Similar to CycleGAN, UNIT addresses the problem of unpaired image-to-image translation. However, UNIT relies on a shared-latent space model. The shared-latent space model disentangles the input and output factors of variation in the data, allowing for the manipulation of the input images to create different outputs which can be useful when specific input features need to be manipulated. CycleGAN, on the other hand, preserves important visual features of the input which ensures that the generated images are consistent with the input.

MUNIT [9] (Multimodal Unsupervised Image-to-Image Translation), proposed by Huang et al., extends the CycleGAN framework to enable multimodal image-to-image translation. MUNIT is an unsupervised approach that can handle image-to-image translation between different modalities, such as changing the style of an image while keeping its content the same. It uses a combination of reconstruction loss, adversarial loss, perceptual loss and diversity loss to train the generator and discriminator. The adversarial loss encourages the generator to produce images that are indistinguishable from the target domain while the reconstruction loss ensures that the generated image is similar to the input image. Perceptual loss helps to preserve the content of the image while changing its style and diversity loss encourages the generator to produce diverse outputs. MUNIT is different from CycleGAN in the sense it generates multiple diverse outputs for a single input image, making it a multimodal approach.

Finally, we would like to discuss about DRIT [10] (Disentangled Representation for Image-to-Image Translation) proposed by Lee et al. which is another approach that uses a disentangled representation approach for image-to-image translation with unpaired data. This approach aims to separate the latent space into two parts: a content space that encodes shared information between different domains, and a domain-specific attribute space that captures the differences in variations for the same content. To achieve this, a content discriminator is used to help separate the two spaces.

7 Challenges

7.1 High Computational Demands

One of the major obstacles we encountered during our project was the computational intensity of training the CycleGAN model. The process necessitated high-end computational resources and lengthy training periods, posing significant computational difficulties.

7.2 Mode Collapse and Oscillations

While training the model, we encountered a significant obstacle in preventing issues such as mode collapse and oscillations. To overcome these challenges, we focused on enhancing the diversity and quality of the dataset, while also ensuring that the model captured the complex patterns in the data.

7.3 Model Evaluation

As specified above, evaluation of the model is difficult since pairwise supervised images are not available in the case of most datasets used for training. Thus, we have used FCN and Inception Score to get an idea about the model performance. However, there are some caveats with respect to these scores. FCN score is only valid for the semantic segmentation problem, and hence other problems of style transfer, etc. can't be evaluated by this method. Inception Score also has its own limitations. Inception Network is trained on the Imagenet dataset consists of 1000 classes only. The Inception score will be low if the GAN is trained on a class outside of these 1000 classes, which is why we observe

a low absolute inception score in our model evaluation, despite it giving us an idea about the relative performance of Pix2Pix and CycleGAN on the same problem. Thus, ideally in the future, better metrics must be used to evaluate CycleGAN performance.

8 Future Scope

While we conducted a comprehensive literature review of recent models, including UNIT (Unsupervised Image-to-Image Translation Networks), MUNIT (Multimodal Unsupervised Image-to-Image Translation), and DRIT (Disentangled Representation for Image-to-Image Translation), there is still more to be explored in this area. To further explore the effectiveness of these models in comparison to CycleGAN, it would be valuable to implement them and conduct a comparative study of their results. This will allow us to evaluate the strengths and limitations of these models and identify areas for improvement in unpaired image-to-image translation.

Acknowledgments

We would like to express our sincere gratitude to our Professor Jean Ponce and Teaching Assistants Ayush Jain and Zuhaib Akhtar for their invaluable guidance, support and mentorship during our research project. We would also like to thank them for providing us with access to High-Performance Computing clusters that were vital to our experiments.

References

[1] Zhu J.Y. Zhou T. Efros A.A. Isola, P. Image-to-image translation with conditional adversarial networks. CVPR, 2017.

[2] M. Mirza B. Xu D. Warde-Farley S. Ozair A. Courville I. Goodfellow, J. Pouget-Abadie and Y. Bengio. Generative adversarial nets. NIPS, 2014.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. IEEE, Jun 2015.

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. IEEE, Jun 2016.

[6] Radim Tyleček and Jiri Šivic. The facades dataset: a large-scale benchmark for image boundary detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 746–753, 2013.

[7] Park T.-Isola P. Efros A.A. Zhu, J.Y. Unpaired image-to-image translation using cycle-consistent adversarial networks. ICCV, 2017.

[8] Breuel T.-Kautz J. Liu, M.Y. Unsupervised image-to-image translation networks. NIPS, 2017.

[9] S. Belongie X. Huang, M.-Y. Liu and J. Kautz. Multimodal unsupervised image-to-image translation. ECCV, 2018.

[10] J.-B. Huang M. K. Singh H.-Y. Lee, H.-Y. Tseng and M.-H. Yang. Diverse image-to-image translation via disentangled representations. ECCV, 2018.