Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Upanshu Srivastava[1]

Abstract

To circumvent the paired image training datasets, the approach of unpaired image-to-image translation is introduced. Since it is challenging to discover paired image datasets and costly too, the two independent image datasets can be mapped by utilizing the Cycle-Consistent Adversarial Networks technique.

1 Introduction

While the original paper demonstrated training and testing multiple unpaired datasets, this paper involves testing and training on limited datasets. Utilizing the three types of datasets, including emojis, Monet paintings and human faces. The system is designed and fed so that the model extracts the features of the emoji training datasets and maps them on the human face dataset. The aim is to get a unique emoji for every face in the human face dataset. Though, the results could be more fulfilling. The other unpaired mapping includes Monet's painting of human faces.

The primary issue the paper addresses is working with unpaired image datasets, which can be cheaper and less demanding, unlike paired image datasets. Moreover, the paired image datasets are biased towards the training data. The translations generated by a model trained on paired data depend highly on the training dataset. If the training dataset is biased towards certain types of images or features, the model may not be able to generalize well to unseen images. (Zhu et al., 2017a)

Paired datasets assume a one-to-one correspondence between the images in the two domains. However, this may only sometimes be the case, as there can be variations in the size, shape, and appearance of objects between the two domains. Handling such variations can be challenging with paired datasets. With the limited scope, paired datasets can only be used for specific image-to-image translation tasks for which the pairs are available. They cannot be easily adapted to other tasks or domains.

The drawbacks mentioned above have led to the development of alternative techniques, such as unpaired image-to-image translation, where there is no one-to-one correspondence between the images in the two domains. (Zhao et al., 2020)

With more extensive availability and better flexibility, the images in the two domains do not need to be related to each other in any particular way. Models trained on unpaired datasets tend to generalize better to unseen data than models trained on paired datasets. This is because models trained on unpaired datasets learn to recognize and translate the underlying patterns and structures of the images rather than simply memorizing the specific pairs in the training dataset. (Gonzalez-Garcia et al., 2018)

Overfitting and generalization is the issue to be looked upon, as discussed, as unpaired image-to-image translation can be susceptible to overfitting, where the model becomes too specialized to the training data and fails to generalize well to new data. This can lead to poor performance on test data or in real-world applications. Though, the problem is very much related to Pix2Pix Model(Choi et al., 2018), irrespective of whether the dataset is paired or unpaired.

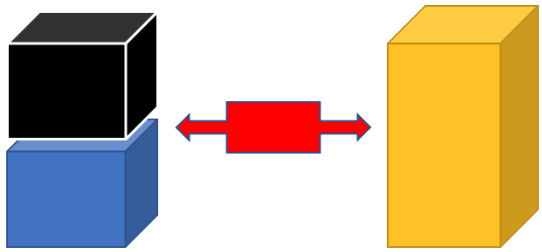

Figure 2: Problematic extraction of Geometry and Characterization using CycleGAN

One of the main drawbacks of geometric changes in unpaired image-to-image translation is that it can lead to unrealistic outputs. When translating images from one domain to another, the generator network may apply geometric modifications to the images to make them look more like the target domain. However, if the generator applies too many or extreme geometric changes, the resulting images may look unnatural and fail to capture the actual characteristics of the target domain. (Yi et al., 2017)

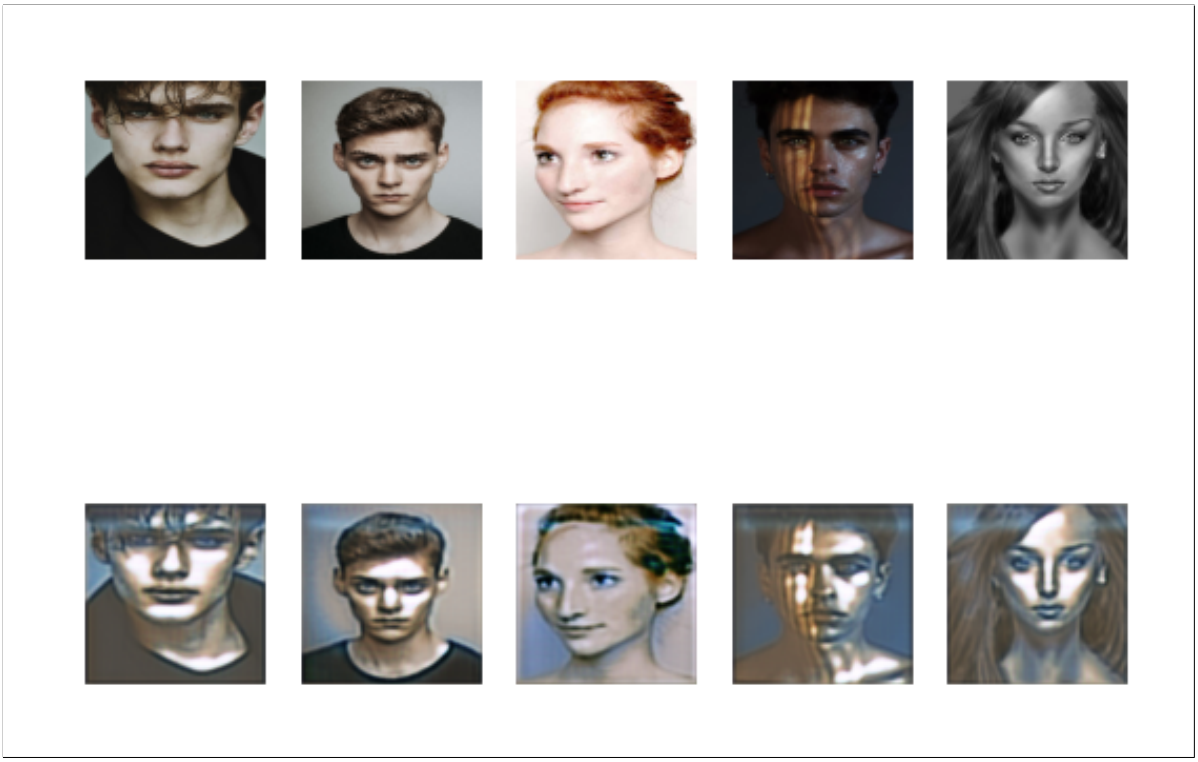

Figure 1: Image Translation from Human Face Dataset to Monet Paints

For example, if the generator translates images of a Cheetah to images of a Leopard, it may try to stretch or compress the cheetah images to make them look more like leopards. However, if it stretches the cheetah's face too much, the resulting image may look bizarre and unrecognizable as a leopard. Similarly, if it compresses the cheetah's body too much, the resulting leopard image may look distorted and unrealistic. Therefore, it is important to balance the amount of geometric changes applied by the generator to ensure that the resulting images look natural and preserve the important characteristics of the target domain. Also, the two datasets of cat species are quite identical. The model could find it hard to extract the features of both leopards and cheetahs. It might be a better case for paired training datasets but not the same wishful result for unpaired image datasets. Reflecting a vedge between the efficiency results of two datasets.

One way to overcome the challenge of unrealistic outputs due to excessive geometric changes in unpaired image-to-image translation is to introduce additional constraints or regularization terms in the loss function.

For example, using a pre-trained neural network, one can use perceptual losses, which measure the difference between high-level features extracted from the generated and target images. By incorporating perceptual losses, the generator is encouraged to preserve the high-level semantic content of the images rather than just focusing on low-level geometric changes. (Zhu et al., 2017b)

Another way to overcome this challenge is to use cycle-consistent adversarial networks (CycleGAN), which learn to translate images between two domains using generator networks and discriminator networks. In CycleGAN, the generator networks are trained to minimize the adversarial loss and the cycle-consistency loss, which enforces that the reconstructed images from the translated images should be similar to the original input images. By doing so, CycleGAN helps preserve the original images' critical characteristics while still allowing for some geometric modifications to occur during the translation process.

Overall, the key is to strike a balance between preserving the essential characteristics of the target domain and allowing for some degree of flexibility and creativity in the translation process.

Moreover, a fused image of two objects can be challenging too. Consider a dataset full of images of cuboids and a dataset full of images containing two cubes with equal dimensions of a cuboid. The colours of the two cubes are different such that our target domain is to acquire an image of a cuboid in accordance with to input image of two cubes. The generator and discriminator network of the model will be challenged to put two different colours in the exact geometry of the input domain. Reflecting on the limitations of maintaining the geometry and distribution of characteristics from the input domain to the target domain.

2 Related Work

Many related works demonstrate different methods for unpaired image-to-image translation. Let's have a look at their efficiency and techniques.

Asymmetric Generative Adversarial Networks (AsymGAN)(Li et al., 2019): Asymmetric Generative Adversarial Networks (AsymGAN) is another approach for unpaired image-to-image translation. It is similar to CycleGAN in that it uses a pair of generators and discriminators to learn the mapping between two image domains. However, AsymGAN uses a different architecture, with one generator that is responsible for generating images in both domains, while the other generator is only used for training purposes.

The main idea behind AsymGAN is to train the generator to produce images in one domain that can fool the discriminator in the other domain while keeping the generated images consistent with the input domain. This is achieved by using two adversarial losses and a reconstruction loss. The first adversarial loss is used to train the generator to generate realistic images in the target domain, while the second adversarial loss is used to ensure that the generated images can fool the discriminator in the source domain. The reconstruction loss is used to ensure that the generated images are consistent with the input images.

Compared to CycleGAN, AsymGAN is more computationally efficient, as it only requires one generator instead of two. It also produces higher-quality images, as it is able to capture more complex and subtle variations between the two domains. However, it can be more difficult to train, as it requires careful tuning of the hyperparameters to balance the different losses.

ITTR (Implicit Template-based Transformation Network)(Zheng et al., 2022): ITTR (Implicit Template-based Transformation Network) is another approach for unpaired image-to-image translation. In ITTR, the mapping between two image domains is learned implicitly using a template-based approach. The idea behind ITTR is to use a template image from the target domain as a reference for generating the corresponding image in the source domain. The template is used to define a transformation function that can map any image from the source domain to the target domain.

The main advantage of ITTR is that it does not require a large amount of paired data for training, as the mapping is learned implicitly from the template. This makes it particularly useful for applications where paired data is scarce or unavailable. However, ITTR also has some limitations. For example, the quality of the generated images depends heavily on the quality of the template, and the performance of ITTR can deteriorate when the images in the two domains have significant differences in scale or rotation.

Compared to CycleGAN and AsymGAN, ITTR is a relatively new approach and has not been extensively evaluated on a wide range of datasets and applications. Therefore, its effectiveness and limitations need to be further explored in future research.

Attention GAN(Tang et al., 2021): AttentionGAN is a modified version of the GAN architecture that incorporates an attention mechanism to selectively focus on certain parts of the input image during the translation process. This allows the model to generate more realistic and visually coherent images by better capturing the underlying structure of the input image.

In Attention GAN, the generator consists of an encoder-decoder architecture with skip connections, similar to other GAN variants. However, in addition to the standard convolutional layers, Attention GAN includes attention blocks that learn to selectively amplify or suppress certain features in the input image. These attention blocks are integrated into the generator architecture at multiple scales, allowing the generator to attend to different levels of detail in the input image.

Compared to CycleGAN, AttentionGAN has been shown to produce higher-quality results in certain image-to-image translation tasks. For example, in one study that compared the two methods on the task of generating realistic facial images from sketches, AttentionGAN was found to outperform CycleGAN in terms of visual quality and fidelity to the input sketches.

However, AttentionGAN can be computationally expensive and requires more training time than CycleGAN. Additionally, it may not be as effective in some image-to-image translation tasks that do not require the selective focus provided by the attention mechanism. As with any machine learning method, the choice of which approach to use ultimately depends on the specific requirements and constraints of the task at hand.

Quality Aware:(Chen et al., 2019) Quality-aware unpaired image-to-image translation refers to a class of methods that aim to improve the visual quality of images generated by unpaired image-to-image translation models. These methods typically use a combination of perceptual and adversarial losses to encourage the generated images to be both visually appealing and consistent with the input images.

One approach to quality-aware unpaired image-to-image translation is to use a multi-scale perceptual loss, which measures the difference between the generated and target images at multiple levels of abstraction. This loss function is designed to encourage the generated images to capture both the fine-grained details and the overall structure of the input images.

Overall, quality-aware unpaired image-to-image translation methods can help to address some of the limitations of traditional unpaired image-to-image translation methods, such as the tendency to generate visually inconsistent or distorted images. However, these methods can also be more computationally expensive and may require more training data than traditional methods.

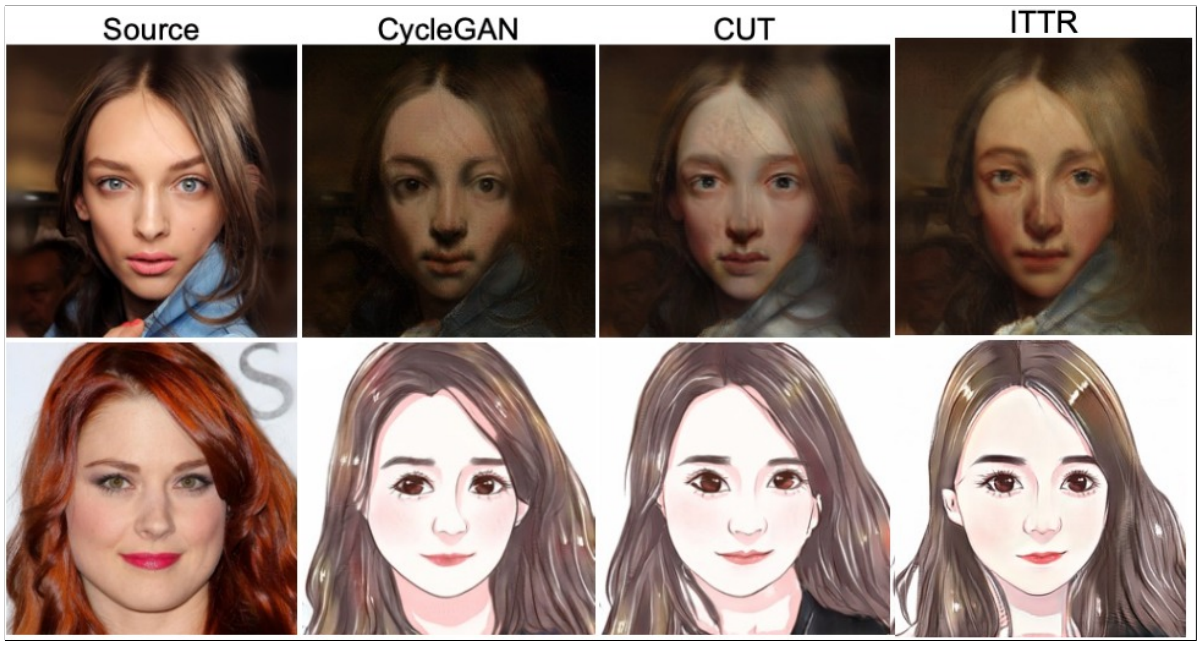

Figure 3: Comparison of different Unpaired Image-to-Image translation Techniques

The main difference between QUILT and CycleGAN is that QUILT introduces an additional quality-aware loss that encourages the generator to produce high-quality images. This loss is defined based on the perceptual similarity between the generated image and its corresponding ground truth image, which is measured using a pre-trained image classification network.

In addition to the quality-aware loss, QUILT also introduces a multi-scale discriminator that operates on multiple scales of the image. This helps the discriminator to capture more detailed information about the image, which in turn helps to improve the quality of the generated images.

Overall, QUILT aims to address the limitation of CycleGAN, which sometimes produces low-quality images due to the lack of a direct supervision signal. By introducing a quality-aware loss and a multi-scale discriminator, QUILT is able to produce higher-quality images while still maintaining the advantages of CycleGAN, such as the ability to perform unpaired image-to-image translation.

CutGAN is a recent approach for unpaired image-to-image translation, which is based on the idea of patch-level processing of images. The name CutGAN comes from the fact that it uses a cutting and stitching mechanism to process images at the patch level.

In CutGAN, an input image is first divided into overlapping patches, which are then processed by two separate generative models: one for translating the patches from the source domain to the target domain and another for translating the patches back from the target domain to the source domain. These models have trained adversarially, with the objective of minimizing the difference between the generated images and the real images in the target and source domains.

The main advantage of CutGAN over other approaches is its ability to handle complex transformations between image domains. By processing images at the patch level, CutGAN can capture local variations and details that may be missed by other methods. Additionally, CutGAN has been shown to be more efficient than other patch-based methods due to its use of a multi-scale generator network that can process patches at different scales.

However, one potential drawback of CutGAN is the need to preprocess images into patches, which can add additional computational overhead. Additionally, the quality of the generated images may be affected by the patch size and overlap used during training.

CutGAN(Han et al., 2021): CutGAN is a method for unpaired image-to-image translation that aims to overcome some of the limitations of CycleGAN. The key difference between CutGAN and CycleGAN is the use of object and semantic segmentation maps as additional input to the generator and discriminator networks. These segmentation maps are used to guide the image translation process and ensure that the translated images retain the same object and semantic structures as the input images.

Compared to CycleGAN, CutGAN is able to produce more accurate and semantically meaningful translations, particularly in cases where the input and output images have complex semantic structures. This is because CutGAN is able to explicitly model and preserve these structures

during the translation process rather than relying solely on adversarial loss and cycle consistency loss.

However, one potential drawback of CutGAN is the increased computational cost and complexity of generating and using the segmentation maps. Additionally, CutGAN may require more training data and longer training times to achieve optimal performance compared to CycleGAN.

The key difference between CutGAN and CycleGAN is using the object and semantic segmentation maps as additional input to the generator and discriminator networks. These segmentation maps are used to guide the image translation process and ensure that the translated images retain the same object and semantic structures as the input images.

Compared to CycleGAN, CutGAN is able to produce more accurate and semantically meaningful translations, particularly in cases where the input and output images have complex semantic structures. This is because CutGAN is able to explicitly model and preserve these structures during the translation process rather than relying solely on adversarial loss and cycle consistency loss.

3 Method

Since the demonstration involves the conversion of human faces dataset to emoji datasets, the target was to generate a unique amalgam of human faces and emojis. Though even after running several epochs and in relation to iterations, heavy loss to the above combination was encountered.

Moreover, the emoji dataset included 10,000 images, whereas the human faces dataset items 7200. Passing on the datasets to two different generators and making it a part of the cycle doesn't reduce the loss substantially, as seen in the figure. Though, I tried experimenting with the model with different datasets reflected in the figures. There were barely any changes observed when the datasets were swapped.

Like observing the loss of Generators A and B, including Discriminator A and B, consider after 100 iterations, "dA[0.120,0.182] dB[0.290,0.059] g[7.490,7.162]" refers to based on this blog, after 100 iterations the discriminator loss for domain A (dA) is 0.120 for the current iteration and 0.182 on average over the last few iterations, while the discriminator loss for domain B (dB) is 0.290 for the current iteration and 0.059 on average over the last few iterations. The generator loss (g) is 7.490 for the current iteration and 7.162 on average over the last few iterations. One can observe that the current iteration loss is less than the average loss over the 99 iterations. The model looks promising, but generators must look more promising as the result isn't wishful.

The model utilizes the 70x70 PatchGAN with a total of four blocks of two generators and two discriminators each. Using the 70x70 PatchGAN introduces instance normalization, training the discriminator directly on fake and real images.

The generator architecture includes two generators called generators one and two. The first generator inputs paint to generate the photos while the second generator generates paint from the photos as input. Each generator has a corresponding discriminator model. The output of generator one is fed to discriminator one to identify being fake or real. Also, the output of generator one is fed to discriminator one to identify being fake or real.

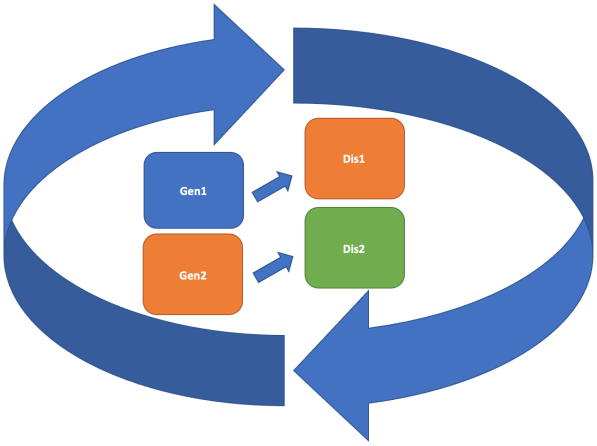

Figure 4: CycleGAN utilizing two Generator and two Discriminators

CycleGAN is a type of generative adversarial network (GAN) that enables unpaired image-to-image translation. The main idea behind CycleGAN is to learn a mapping between two domains without requiring a one-to-one correspondence between images from each domain. For example, we might want to convert images of one item into images of another item, without having to match each item image to a specific item image.

To accomplish this, CycleGAN uses two generators (one for each domain) and two discriminators (also one for each domain). The generators learn to map images from one domain to another, while the discriminators learn to distinguish between real images from each domain and generated images.

The CycleGAN training process involves two main objectives: adversarial loss and cycle consistency loss. The adversarial loss encourages the generators to produce images that are similar to the target domain, while the cycle consistency loss ensures that the generators produce consistent mappings.

To achieve the cycle consistency loss, CycleGAN introduces a cycle consistency constraint, which requires that mapping an image from one domain to the other and then back again should result in the original image. This constraint helps to prevent the generators from learning non-reversible mappings, which could lead to blurry or distorted

Table 1: Classification average losses for Discriminator on various data sets after 1000 iterations (10 Epochs)

DATA SET	DIS1	DIS2
EMOJI	0.276	0.166
HUMAN FACES	0.229	0.176
MONET	0.198	0.150

images.

During training, the generators and discriminators are updated in an alternating fashion to minimize the adversarial and cycle consistency losses. Once trained, the generators can be used to translate images between the two domains, even if there is no one-to-one correspondence between them.

Overall, CycleGAN has several advantages over other image-to-image translation methods, including its ability to handle unpaired data and its ability to learn both forward and backward mappings, making it a powerful tool for various image manipulation tasks.

4 Experiments

The generator model from the cycle consistency is trained to produce original images. The six figures listed below reflect the translation from one image form to another. Considering Figure 5, the forward propagation of the generator (Gen1) reflects the conversion from Monet Paintings to Photos. Though, there is a substantial loss.

The Backward Propagation reflects the generator (Gen2) reflects the conversion from photos to Monet Paintings. Though, the same could be said for the output generated by Generator 2. All the six figures on the following page reflect the demonstration of the experiments.

As the functionality and loss functions are co-dependent on each of the propagation that is forward and backwards. Let's consider talking about the four losses:

1. Adversarial Loss (L2 MSE): Minimizing the loss predicted by Discriminator 1 and 2 for generated images marked as real

2. Identity Loss (L1 MAE): To minimize the loss, it outputs the source image as is without translation

3. Cycle Loss Forward Propagation (L1 MAE): The loss generated during the translation to generated image from the input image in the generator process.

4. Cycle Loss Backward Forward Propagation (L1 MAE): The loss works in the backward direction as it works in the above case.

Also, during the experiments, the dataset was distributed into training and testing datasets. Eighty per cent of the data was allocated for training, while the remaining 20 per cent was allocated for testing the model. Personally, I face hardware limitations since I wasn't set up with the server. So, all the computation was done through a personal laptop. After running the computer for approx five days or 120 hours, the results weren't as satisfying. Working with huge and complex datasets can make the process tedious. Suffered from training instability, which can result in mode collapse or poor image quality.

5 Conclusion

The paper includes four primary sections Introduction, Related Work, Methods and Experiments. The introductions sections talk about why the Unpaired method is advantageous and what is the motivation behind the development of such a model. Also, concentrating on how it's cost-effective. But the process drawbacks have also been discussed, and how it fails for fused objects translation. Lacking the geometry and symmetry of the objects for conversions.

Related work discusses the major developments in the field of Unpaired image-to-image translation, such as AttentionGAN, ITTR, AsymGAN and Quality Aware. I focussed on these methods' efficiency and their comparison with the CycleGAN. The paper also mentions the latest development on the Unpaired Image translation, that is, ITTR (Implicit Template-based Transformation Network).

The core part of the paper is Methods and Experiments; the paper talks about the generator, discriminator architecture and losses related to it. Moreover, the details of CycleGAN and how the reimplementation was taken into account.

In conclusion, CycleGAN has shown great potential for unpaired image-to-image translation, allowing for the transformation of images between different domains without the need for paired training data. However, there are some limitations and challenges associated with this approach, including issues with geometric changes and the dependence on the dataset distribution.

Despite these limitations, there are still many potential future directions for work in this area. One possible avenue of research is to explore different loss functions and training strategies that can help to address some of the challenges associated with CycleGAN. Another potential area of focus is to investigate the use of other deep learning techniques, such as attention mechanisms and reinforcement learning, to improve the performance and efficiency of unpaired image-to-image translation. Additionally, there is a need for the development of more comprehensive evaluation metrics to assess the quality of the generated images accurately.

Overall, CycleGAN has opened up new possibilities for image-to-image translation, and future research in this area has the potential to advance the field and its applications

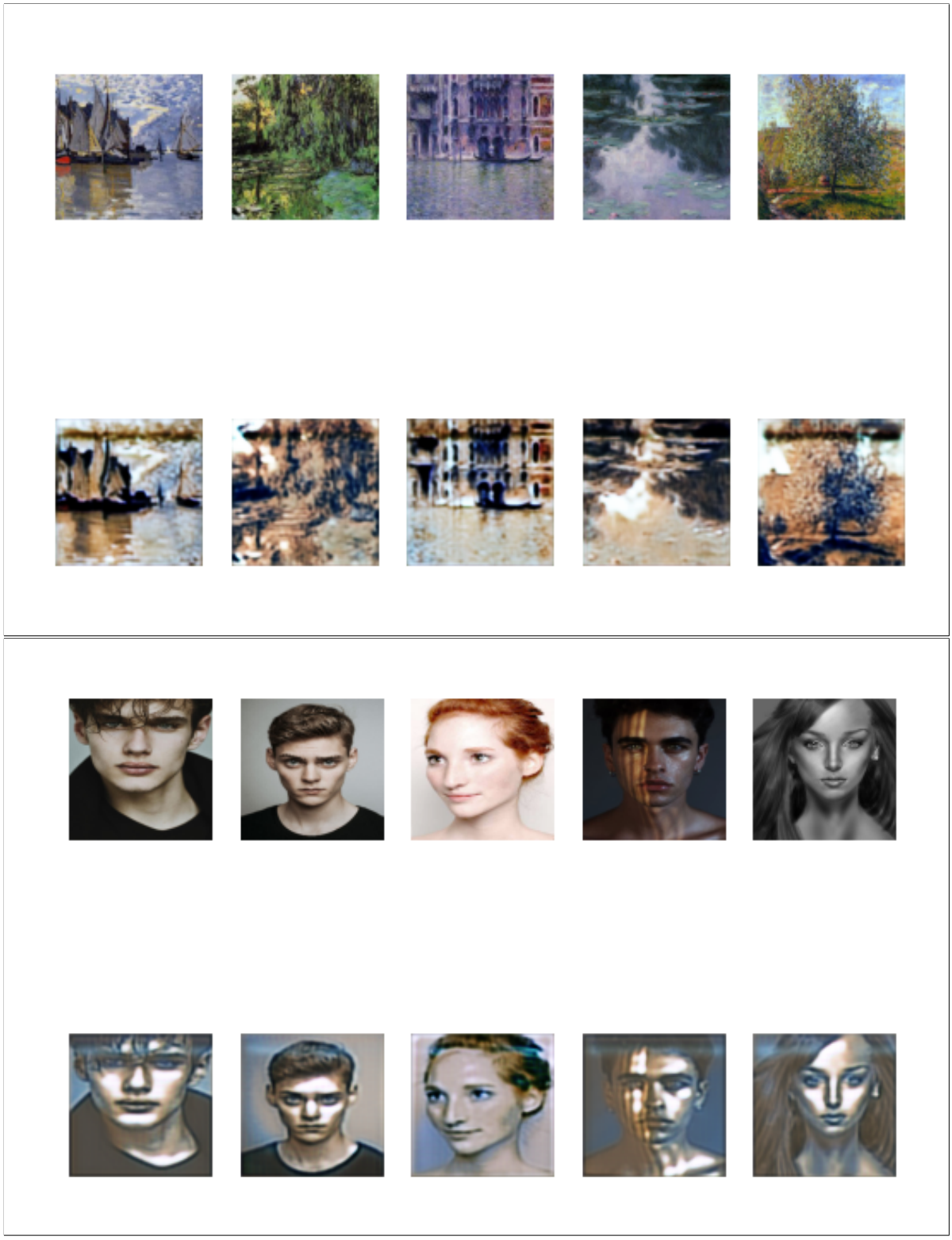

Figure 5: Results reflected using Different Datasets: Paint to Photos/Photos to Paint

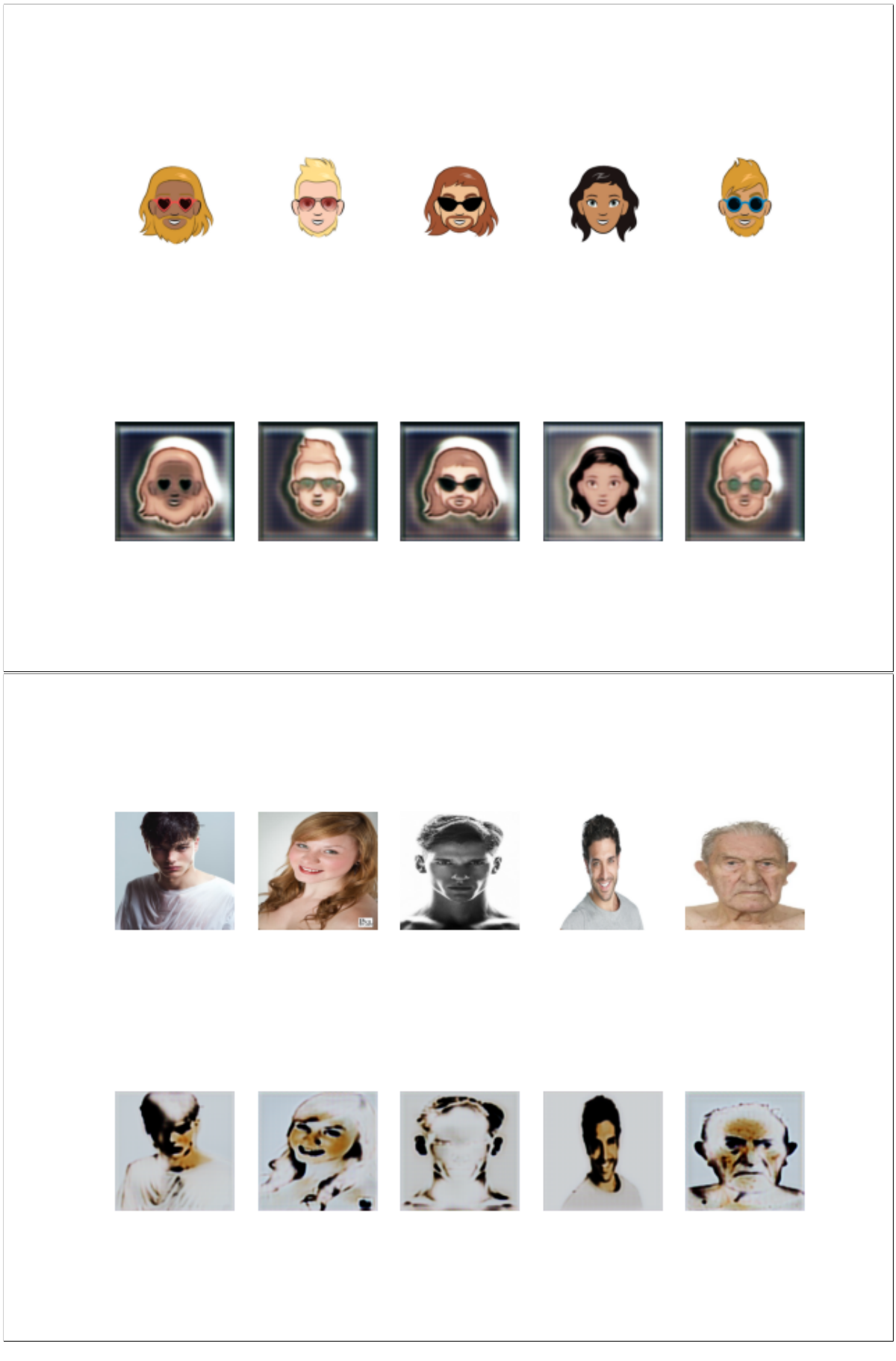

Figure 6: Results reflected using Different Datasets: Emojis to Paints/Photos to Emojis

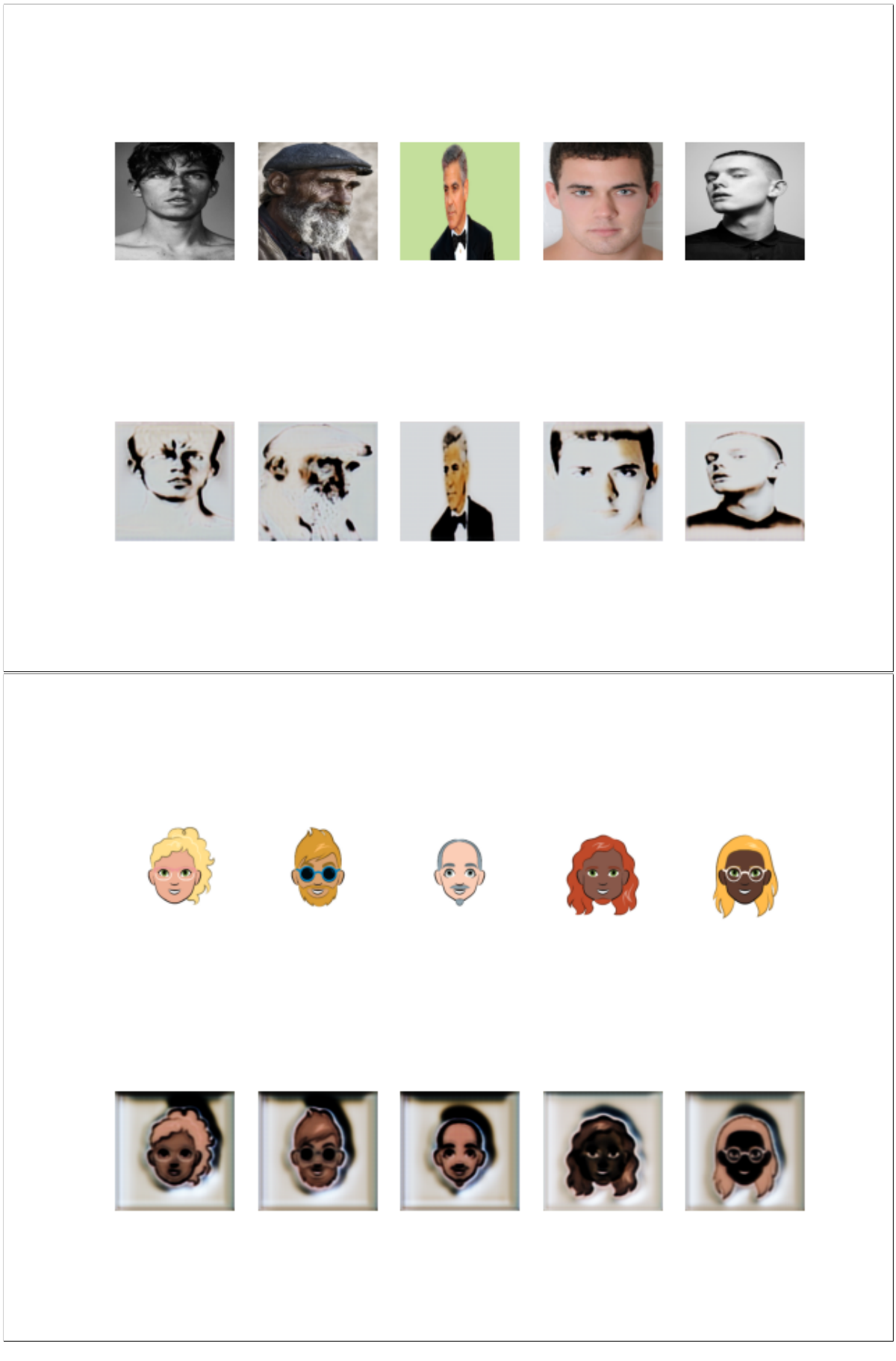

Figure 7: Results reflected using Different Datasets: Photos to Paints/Emojis to Paints

further.

6 Acknowledgement

I want to express my gratitude to all those who have contributed to the success of this work. I acknowledge the support received from the Unpaired Image-to-Image translation study group and Prof. Qi Guo. I am also grateful to the participants who generously gave their time to contribute to this study. I will also like to thank the other course project group, from Ideas to Innovation coursework. The team demonstrated the paired image dataset using Pix2Pix GAN Model, helping me greatly in the unpaired image translation. I would like to thank my teammates, Ruijie Song and Yien Chein Huang, for their support. I extend my sincere thanks to our colleagues who provided valuable feedback and assistance throughout the course of this project. I appreciate their insights and expertise, which helped us to refine and improve our work.

I want to acknowledge the contributions of the reviewers who provided insightful feedback that significantly improved the quality of this paper.

6.1 Code Availability

GitHub

References

Chen, L., Wu, L., Hu, Z., and Wang, M. Quality-aware unpaired image-to-image translation. *IEEE Transactions on Multimedia*, 21(10):2664–2674, 2019.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Gonzalez-Garcia, A., Van De Weijer, J., and Bengio, Y. Image-to-image translation for cross-domain disentanglement. *Advances in neural information processing systems*, 31, 2018.

Han, J., Shoeiby, M., Petersson, L., and Armin, M. A. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 746–755, 2021.

Li, Y., Tang, S., Zhang, R., Zhang, Y., Li, J., and Yan, S. Asymmetric gan for unpaired image-to-image translation. *IEEE Transactions on Image Processing*, 28(12):5881–5896, 2019.

Tang, H., Liu, H., Xu, D., Torr, P. H., and Sebe, N. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 2021.

Yi, Z., Zhang, H., Tan, P., and Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017.

Zhao, Y., Wu, R., and Dong, H. Unpaired image-to-image translation using adversarial consistency loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 800–815. Springer, 2020.

Zheng, W., Li, Q., Zhang, G., Wan, P., and Wang, Z. Ittr: Unpaired image-to-image translation with transformers. *arXiv preprint arXiv:2203.16015*, 2022.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017a.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017b.