Unsupervised Image to Image Translation using Attention Guided CycleGANs

Ifrah Siddiqui
MSDS18002

Maham Nasir Khan
MSDS18041

Mirza Elaaf Shuja
MSDS18051

1 Project summary

Image to image translation is a computer vision problem that deals with mapping of one image onto another. Its applications include collection style transfer (image translation), object transfiguration and photo enhancement. A lot of work has been done for supervised as well as unsupervised image translation by using Generative Adversarial Networks (GANs). Unsupervised image-to-image translation learns the joint distribution of images in different domains by using images from the marginal distributions in individual domains. Generally GANs fail at translation of a specific locale of the image therefore current researches in this domain struggle to do image to image translation by taking into account the geometry of translated objects in the images, alongwith the retention of background of the original image. Recent work in GANs uses attention mechanisms to transfer style from one image to another, where attention refers to processing over a specific subset of inputs rather than the entire input set. This project incorporates Attention in CycleGANs to solve the problem of unwanted background image translation. AGGANs perform attentive image style transfer onto a restricted locale of the image but these GANs also fail to deal with the geometry of objects in the images. We propose a different training technique for the attention-GAN that is able to translate not only the style of images but also deal with the geometry of objects in the images to some extent.

2 Introduction

GANs have amassed high popularity in the computer vision world since 2014. Till September 2018, more than 500 variations of GANs have been introduced with each one outperforming the other in various tasks. While a lot of work has been done in the field of image to image translation and GANs, still there are some problems that need to be addressed. In the proposed model we aim at addressing the problem of background translation and handling the basic geometry of objects in the images during translation.

3 Background

3.1 Literature Review

Ian Goodfellow proposed the basic architecture of GANs to improve the generation of fake images [1]. Basic GANs consist of two neural networks pitted against one another and can learn to imitate any distribution of data. The two networks are:

- a generator; that generates content based on some probability distribution
- a discriminator; that discriminates whether the input is real (i.e from the original training data) or fake (i.e generated by the generator)

The ideal goal is to generate such data for which the discriminator gets confused between tagging it as real or fake. During training random noise is fed to the generator which tries to capture the distribution of data and generate fake images from the learned distribution. These generated images act as an input to the discriminator which knows the distribution of real data, and it gives the probability of fakeness of the input image. As we continue with the training process, the generator tries to maximize the Binary Cross Entropy (BCE) loss for the generated fake images to fool the discriminator while the dicriminator tries to minimize this loss to correctly estimate the probability of fakeness of the generated images. This minimax game of the BCE loss function results in the generation of a fake image that renders the discriminator unable to distinguish it from a real one. During testing, the image to be translated is fed to the trained model that translates it by mapping it to the distribution of the target domain.

Radford et. al have used deep convolutional generative adversarial networks (DCGANs) to solve the problem of unsupervised learning in image to image translation [2]. DCGANs basically involve a convolutional layer without the max-pooling or fully connected layers. Convolutional strides and transposed convolutions are used for upsampling and downsampling. Using the power of convolutional neural networks, these GANs can learn a hierarchy of image representations from objects parts to scenes in both the discriminator and the generator.

Isola et. al propsed conditional GANs for supervised image to image translation [3]. They have used a ”U-Net” based architecture for the generator and a convolutional ”PatchGan” classifier for the discriminator to capture local style statistics.

Zhu et. al proposed a novel approach to deal with the problem of unpaired image to image translation [4]. They introduced cycleGANs to learn the representation of images belonging to two different domains. The proposed architecture learns the function G by mapping the images from source domain X to the target domain Y such that G: X → Y using adversarial loss coupled with an inverse mapping F such that F: Y → X using cycle consistency loss to push F(G(X)) = X.

3.2 Previous Work Reproduced for the Project

3.2.1 Vanilla GANs

We have reproduced results of vanilla GANs using MNIST dataset to generate good quality fake data as shown in Figure 1.

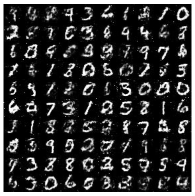

Figure 1: *Fake images of MNIST dataset generated from vanilla GANs*

3.2.2 DCGANs

DCGANs makes use of Convolutional layers for the task of unsupervised learning. The basic architecture of DCGANs is shown in Figure 2. We reproduced the results of DCGANs using the simpson cartoon dataset.

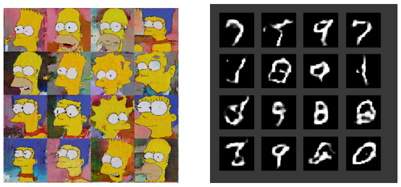

Figure 2: *Fake images of simpson cartoon character (left) and mnist data (right) generated by DCGAN.*

3.2.3 Conditional GANs

Conditional GANs are the variants of GANS used for supervised learning. Their basic architecture is shown in figure 4. They generate fake data given the ground truth of the data. We have reproduced results of conditional GANs using facades dataset as shown in Figure 3.

3.2.4 CycleGANs

We trained cycleGANs on celebA dataset to compare the results of our model with cycleGANs. The results from cycleGANs are shown in appendix A.5.

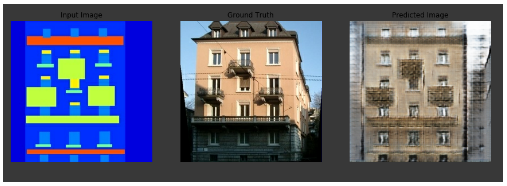

Figure 3: *Left: Input mask ; Centre: Ground Truth ; Right: Output image generated by Conditional GAN*

4 Project Description

4.1 Paper implemented

Youssef et. al proposed attention module in cycleGANs to solve the problem of translation of backgrounds [5]. The proposed architecture gives attention to the specific regions of the image without requiring manual supervision.It creates attention maps and based on those maps, creates more realistic mappings from source domain to the target domain. The basic architecture of attention guided GANs (AGGAN) is shown in Figure 4.

4.1.1 Basic Approach

Independently sampled data instances of the source and target domains are fed to the model. The model first detects the target location to be translated using attention maps and then apply translation on those specific parts.

4.1.2 Attention Module

Input image is fed to the generator which maps it to the target domain, then the same input is fed to the attention module that generates the foreground and background objects separately via an element-wise product on each RGB sample. Translation of image is done using the foreground map of the attention module and then the background is added in the masked output to retain the background of the original image.

4.1.3 Loss Function

Since this work uses cycleGANs coupled with attention module so it uses the following two losses during training:

- **Adversarial loss:** It uses adversarial loss for training of GANs. Let θg and θd represents the parameters of generator and discriminator respectively and z represents the distribution of the generated fake images. The adversarial loss is given in Equation 1.

$$L_{adversarial} = min_{\theta_g} max_{\theta_d} [\mathbb{E}_{x \sim p_{data}} log D_{\theta_d}(x) + \mathbb{E}_{z \sim p_z} log(1 - D_{\theta_d}(G_{\theta_g}(z)))] \tag{1}$$

- **Cycle consistency loss:** The second loss being used is cycle consistency loss. Let S be the source image and $S^{'}$ be the fake source image generated translating S to the target domain and then inverse mapping that translated image back to the source domain. The cycle consistency loss is given in Equation 2.

$$L_{cyclic}(S, S^{'}) = ||S - S^{'}||_1 \tag{2}$$

The total loss of the network is given as:

$$Loss_{total} = Loss_{cyclic} + Loss_{adversarial} \tag{3}$$

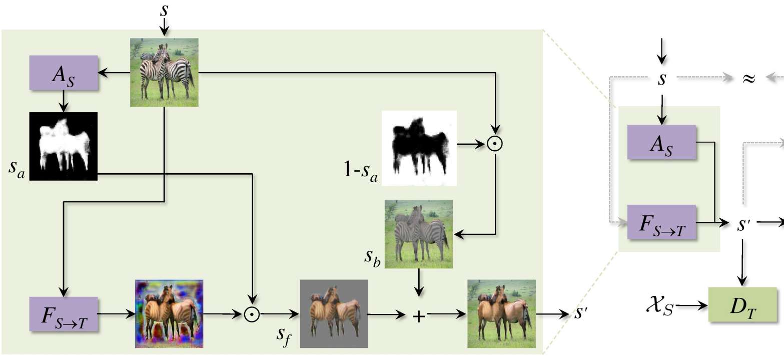

Figure 4: *Data-flow diagram from the source domain S to the target domain.*

4.1.4 Results Reproduced from AGGAN

We have reproduced results of AGGAN on horse2zebra dataset using the pre-trained models. The results are shown in the figure .

4.2 Novelty introduced

We have used attention guided cycle consistent GANs for gender translation in images using celebA dataset. Figure 5 shows some of the images from our dataset.

Figure 5: *The left three images are of females and the right three are of males*

4.2.1 Parameter tuning

The experiments were carried out using 100 epochs and batch size of 1. After analyzing the attention maps generated during training we found out that the maps generation was good till epoch 40 and after that the maps began to distort, so we let the attention module learn till epoch 40, till then the discriminator was trained using the full image, and after epoch 40 it was trained using the foregrounds only. A threshold is used in the network for feeding the foreground to the discriminator that restricts the discriminator to be distracted by very small values. According to our application we set the value of that threshold to 0.05.

Table 1: Frechet Inception Distance between original and fake generated images

	Cycle-GAN	**Attention-GAN**
Male to Female	107.475	102.947
Female to Male	91.912	90.684

4.2.2 Results

Our model is able to translate the gender from males to females and vice versa as shown in Figure 6. The FID scores of the cycle GANs and our model is reported in table I. FID score measures the distance between the original images and the generated fake images.

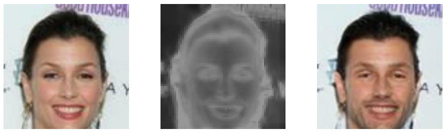

Figure 6: *Input image(left), generated attention map (middle), translated image (right)*

4.2.3 Conclusion

By incorporating the attention module in the cycleGAN and carefully tuning its parameters according to our application, our model is able to produce better results than cycleGAN for image translation.Our model also caters the basic shape of faces in the images. Hence, with a slight further improvement, our model can produce better results than [5] by taking into account the geometry of objects in the images.

References

[1] Goodfellow, Ian, et al. ”Generative adversarial nets.” Advances in neural information processing systems. 2014.

[2] Radford, Alec, Luke Metz, and Soumith Chintala. ”Unsupervised representation learning with deep convolutional generative adversarial networks.” arXiv preprint arXiv:1511.06434 (2015).

[3] Isola, Phillip, et al. ”Image-to-image translation with conditional adversarial networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[4] Zhu, Jun-Yan, et al. ”Unpaired image-to-image translation using cycle-consistent adversarial networks.” Proceedings of the IEEE international conference on computer vision. 2017.

[5] Mejjati, Youssef Alami, et al. ”Unsupervised Attention-guided Image-to-Image Translation.” Advances in Neural Information Processing Systems. 2018.

A Appendix

A.1 Basic Architecture of vanilla GANs

The basic structure of vanilla GANs is shown in the Figure 7.

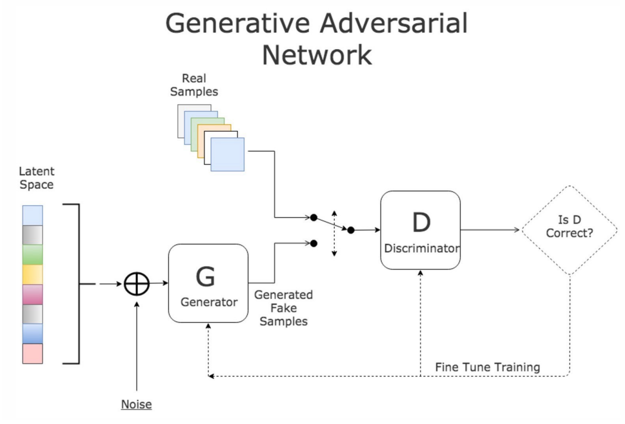

Figure 7: *Basic architectures of vanilla GAN*

A.2 Architecture of DCGANs

The architecture of convolutional GANs is shown in Figure 8.

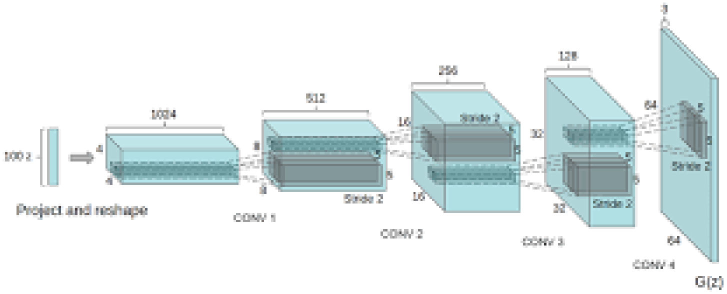

Figure 8: *Basic architectures of DCGAN*

A.3 Flowchart of Conditional GANs

The architecture of convolutional GANs is shown in Figure 9

.

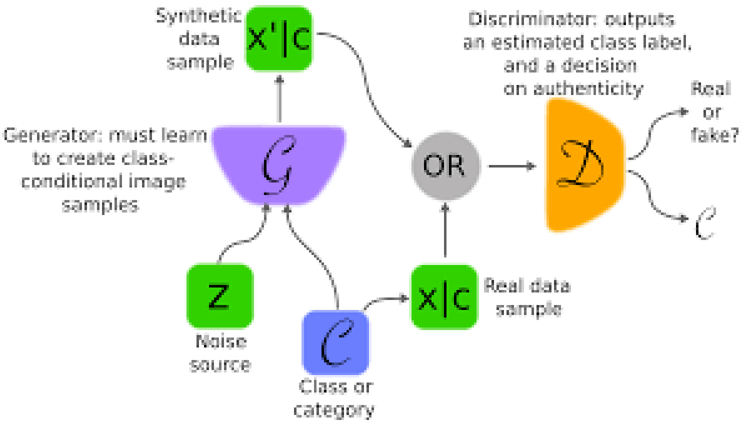
Synthetic data sample
X'|C
Discriminator: outputs an estimated class label, and a decision on authenticity
Real or fake?
Generator: must learn to create class-conditional image samples
G
OR
D
C
Z
Noise source
C
Class or category
X|C
Real data sample

Figure 9: *Flow diagram of conditional GANs*

A.4 Architecture of CycleGANs

Figure 10 depicts the basic architecture and flow diagram of image translation using cycleGANs.

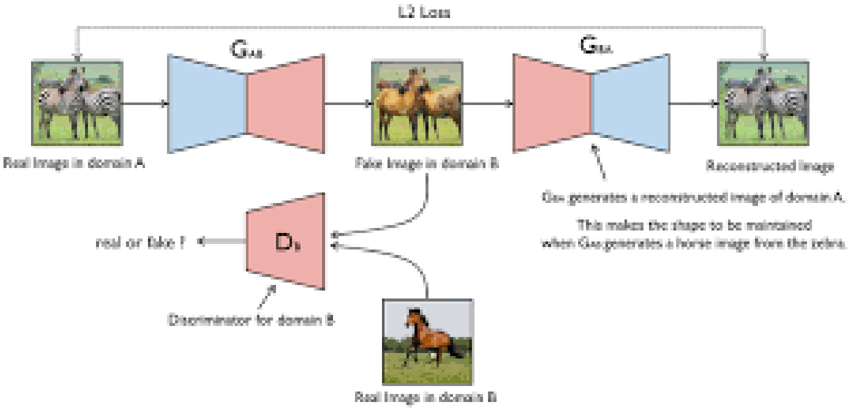
L2 Loss
G_AB
G_BA
Real Image in domain A
Fake Image in domain B
Reconstructed Image
G_BA generates a reconstructed image of domain A.
This makes the shape to be maintained when G_AB generates a horse image from the zebra.
real or fake ?
D_B
Discriminator for domain B
Real Image in domain B

Figure 10: *Flow diagram of cycleGANs*

A.5 Additional Results from Our Experiments

Figure 11 and Figure 12 show the comparison of images produced by our proposed model to that of cycleGANs.

A.6 Contribution of group members

Since we had to develop the basic understanding of GANs and its variants so all the group members worked together generally so that every one has a clear intuition of the basic architecture and working of GANs. However, Table 2 shows the detailed contribution of each group member for the completion of the project.

Table 2: Task-wise contribution of the group members

Task	**Done by**
Vanilla GANs	Ifrah, Maham, Elaaf
DC GANs	Maham
Conditional GANs	Maham
Cycle GANs	Ifrah, Maham, Elaaf
Reproduce the results of AGGAN	Ifrah
CelebA dataset resizing	Elaaf
Implementation of proposed model	Ifrah, Elaaf
Report writing	Ifrah
Poster Design	Maham

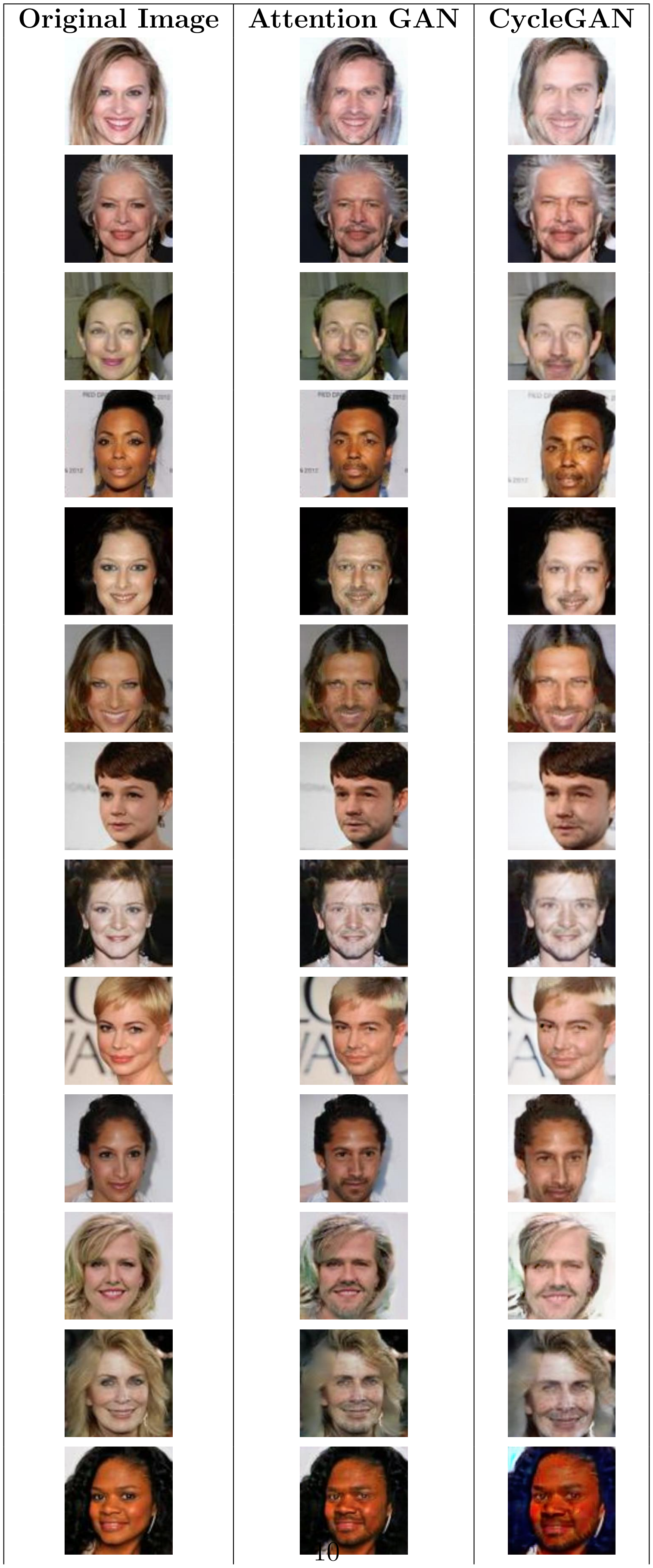

Figure 11: *Comparison of our proposed AGGAN with cycleGANs for female to male translation.*

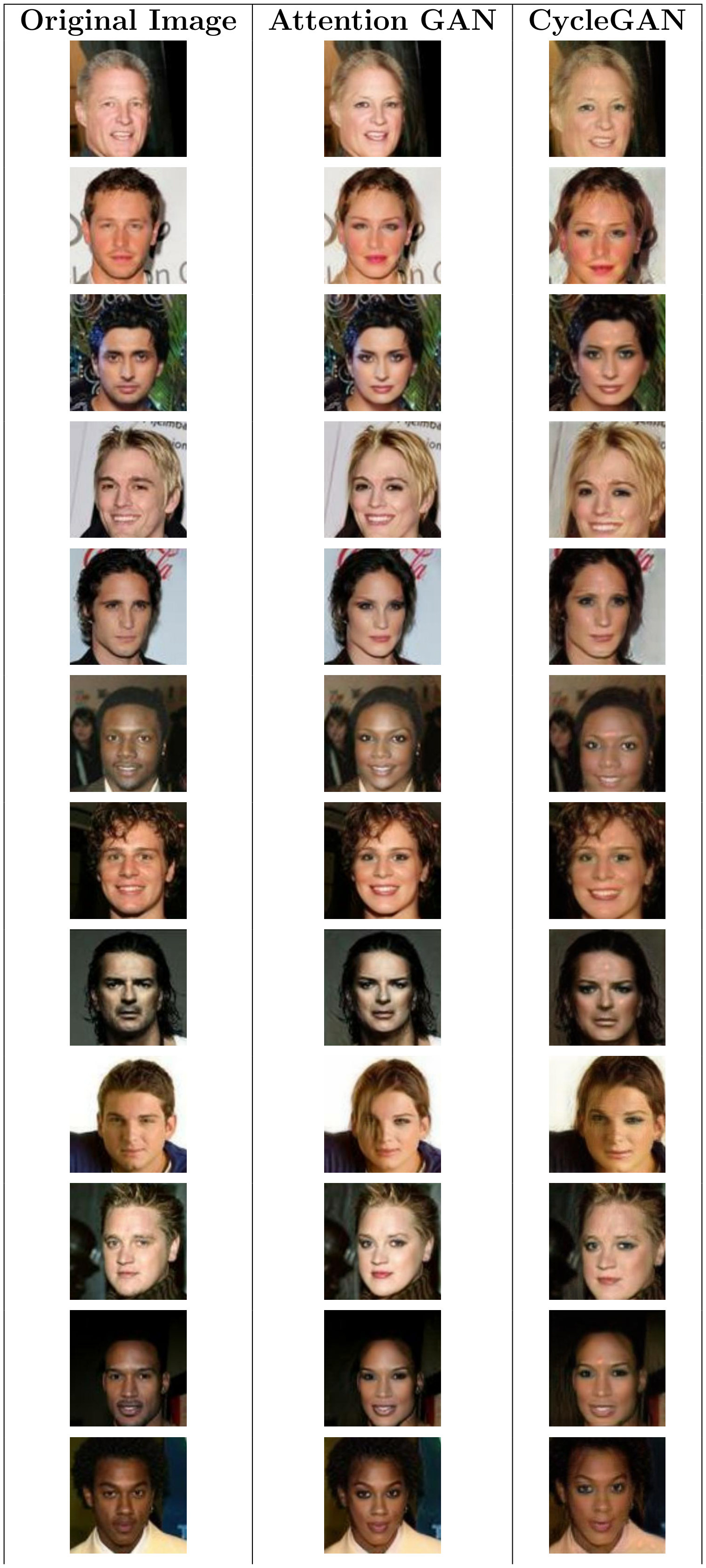

Figure 12: *Comparison of our proposed AGGAN with cycleGANs for male to female translation.*