BEHIND THE MASK

2.1 Introduction

While the use of Generative Adversarial Networks (GANs) has been a breakthrough in the computer vision industry, there exist multiple styles of GANs that are well-tailored to solve specific problems. Behind the mask, though sounding trivial, points to a critical use case. The situation represents the unsupervised image to image translation by discovering distinctive features from the first set and generating images belonging to the other set by learning distinctions between these two. This technique is more feasible for problems where paired images are not available. Using algorithms like Pix2pix is not viable since paired images are expensive and difficult to obtain. To tackle this problem, CycleGAN, DualGAN, and DiscoGAN provide an insight into which the models can learn the mapping from one image domain to another one with unpaired image data. But even in this case, since the problem is reconstructing human faces by removing their facial masks, which requires non-linear transformations, this is tricky. Moreover, the previously mentioned techniques also alter the background and make changes to unwanted objects as they try to create fake images through generators and discriminators. The goal is to implement an approach that not only detects discriminating factors between two sets of pictures but also generates images without altering the rest of the details and only targets specific areas of the image to change.

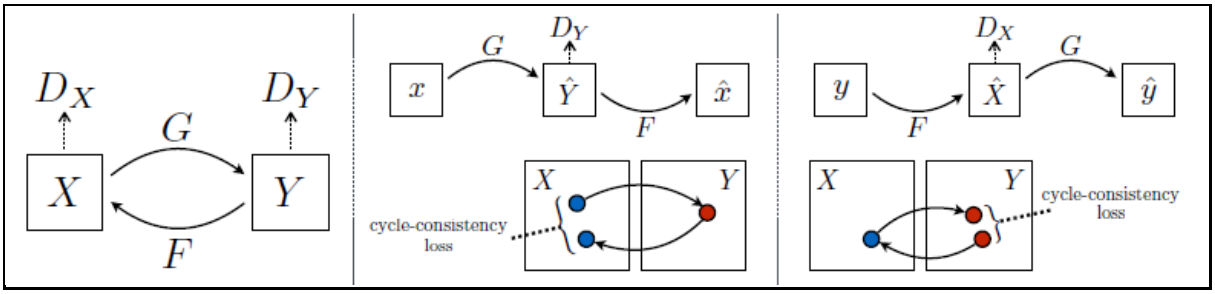
One other technique that can be employed to address this could be to use Contrast GAN, which selects a part of an image, transforms that based on differentiating factors, and then pastes it back to the original image. However, this created an issue since the face masks used in our case had to be of the exact dimensions and identical, which was not the case. To overcome these challenges, we tried to employ an attention-based technique named AGGAN, Attention-Guided Generative Adversarial Networks, for image translation that does not require additional models/parameters to alter a specific part of the image. The AGGAN comprises two generators and two discriminators, like CycleGAN. Two attention-guided generators in AGGAN have built-in attention modules, which can disentangle the discriminative semantic object and the unwanted part by producing an attention mask and a content mask. The underlying image is fused with these masks to create quality fake images. We also consider additional losses to reduce the variance and make the related images pixel consistent. We think of a more sophisticated network by applying two possible subnets to identify the attention and content masks. To avoid omitting any details, the network employs two attention masks, one for the foreground and one for the background, so that the foreground can be better learned, and the background can be preserved. Also, in this case, the generative content mask is introduced to multiple types of facial masks to identify a broad spectrum of them and effectively remove them and create a more decadent generation space. To obtain high-quality unmasked images, we aim and expect to translate masked images to unmasked ones that can be employed on various faces with different skin colors and expressions.

2.2 Problem Analysis

Recently several models based on Generative Adversarial Networks, or GANs, have been used extensively for unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the

original dataset. This property makes them popular in image-to-image translation tasks such as translating photos of summer to winter or day to night and generating photorealistic photos of objects, scenes, and people that even humans cannot tell are fake. We initially implemented Cycle GAN and assessed its performance based on the references provided.

Cycle GAN's primary goal is to learn the mapping functions between two domains X and Y given training samples $\{x_i\}_{i=1}^{N}$ where $x_i \in X$ and $\{y_j\}_{j=1}^{M}$ where $y_j \in Y$. We consider the following model: -

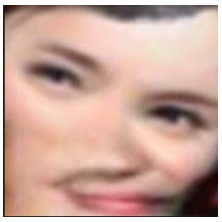


Where, G: X Y and F: Y X are the two mappings with associated adversarial discriminators DY and DX.

Cycle GANs introduce two cycle consistency losses that work with the central principle of translating an image from one domain to another and back again to achieve the initial input image. For this purpose, we consider forward cycle-consistency loss: x G(x) F(G(x)) x and backward cycle-consistency loss: y F(y) G(F(y)) y. Overall we considered the following combination of Adversarial loss, cycle consistency loss and identity loss for optimization as well,

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

For training from an infrastructural standpoint, we utilized Tensors performing operations on GPUs. All the 1898 images of masked people and 1918 images of people without a mask were used in training the model. Performing transformations such as flipping, cropping, and scaling of these images using the albumentations was also necessary. We trained our model in batches of 1 for a total of 150 epochs with a learning rate of 1e-4, identity= 5 and cycle= 12. Adam optimizer was used with a learning rate of 1e-4 which linearly decayed after the first 100 epochs, 1= 0.5 and 2= 0.999.

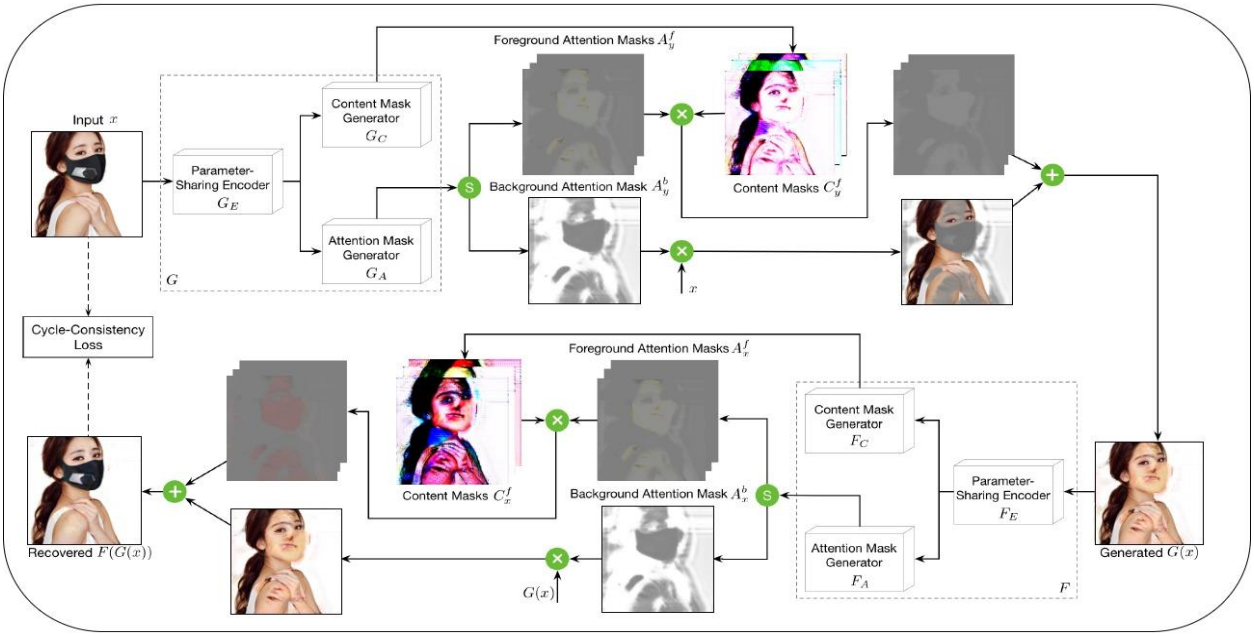


However, we observed from the above output images that the model wasn't able to retain face structure. Often features such as the nose or mouth were inserted onto the mask unevenly. Another drawback that we observed in Cycle GANs is that it is pretty challenging to perform the minimization of the Generator and maximization of the Discriminator. Our model fetched inconsistent results due to this, and as such unpaired image-to-image translation remains a

challenging problem. The resultant model changes unwanted parts in the translation and can be easily affected by background changes. We thereby decided to explore another approach in the form of Attention GANs.

Attention-Guided Generative Adversarial Networks (AttentionGAN) bear the vital advantage of having generators that can capture the foreground of the target domain and preserve the background of the source domain effectively. The AGAN generator learns both foreground and background attention. It leverages the foreground attention to select from the generated output for the foreground regions while using the background attention to maintain the background information from the input image. In this way, AGAN can focus on the most discriminative foreground and ignore the unwanted background. With this framework, we aim to stabilize the GANs' training and improve the quality of generated images by jointly approximating attention and content masks with several losses and optimization methods.

Behind the Mask Architecture



In the proposed architecture, the two generators G and F are composed of two subnets each for generating attention masks and content masks as shown above. For instance, is comprised of a parameter-sharing encoder GE, an attention mask generator GA, and a content mask generator GC. GEaims at extracting both low-level and high-level deep feature representations. GC targets to produce multiple intermediate content masks. GA tries to generate multiple attention masks. By the way, both attention mask generation and content mask generation have their own network parameters and will not interfere.

We had a dataset consisting of 1896 images with mask & 1918 images without mask. We split given images in 80:20 ratio, with 80% used for training and 20% for testing. After initially having trained masked and unmasked images using Cycle GAN and not fetching satisfactory results, we went ahead with unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks model. We observe that AttentionGAN achieves significantly better results than CycleGAN. It produces clearer transformed images. The proposed generators are equipped with a built-in attention module, which can disentangle the discriminative semantic objects from the unwanted parts via producing an attention mask and a content mask. Then we

fuse the attention and the content masks to obtain the final generation. Moreover, we design two novel attention-guided discriminators which aim to consider only the attended foreground regions. The proposed attention-guided generators and discriminators are trained in an end-to-end fashion. Then we fused the attention and the content masks to obtain the final generation. This framework stabilizes the GANs' training and thus improves the quality of generated images through jointly approximating attention and content masks with several losses and optimization methods. We designed two novel attention-guided generation schemes for the proposed framework, to better perceive and generate the most discriminative foreground parts and simultaneously preserve the unfocused objects and background. We performed transformations on the images by resizing, cropping, and scaling to 256X256, left-right flip and random crop for data augmentation. We trained our model in batches of 4, for 200 epochs with a learning rate of 1e-4, identity= 5 and cycle= 10. Adam optimizer was used with a learning rate of 1e-4 which linearly decayed after the first 100 epochs, 1= 0.5 and 2= 0.999. We iteratively increased the number of epochs by observing intermediary train results progressively.

The proposed attention-guided generation scheme I can achieve promising results on the facial expression translation as seen in the architecture diagram, where the change between the source domain and the target domain is relatively minor. However, it performs unsatisfactorily on more challenging scenarios in which a more complex semantic translation is required. To tackle this issue, we further proposed a more advanced attention-guided generation scheme II. The improvement upon scheme I is mainly three-fold: First, in scheme I the attention and the content masks are generated with the same network. To have a more powerful generation of them, we employ two separate sub-networks in scheme II. Second, in scheme I we only generate the foreground attention mask to focus on the most discriminative semantic content. However, to better learn the foreground and preserve the background simultaneously, we produce both foreground and background attention masks in scheme II. Third, as the foreground generation is more complex, instead of learning a single content mask in the scheme I, we learn a set of several intermediate content masks, and correspondingly we also learn the same number of foreground attention masks. The generation of multiple intermediate content masks is beneficial for the network to learn a richer generation space. The intermediate content masks are then fused with the foreground attention masks to produce the final content masks. Scheme I take a three-channel RGB image as input and outputs a one-channel attention mask and a three channel content mask. Scheme II takes a three-channel RGB image as input and outputs n attention masks and n-1 content masks, thus we fuse all of these masks and the input image to produce the final results. We considered the following losses for the model - GAN, Attention-guided GAN, Cycle-consistency, Attention, and Pixel loss.

$$\begin{aligned}\mathcal{L} =& \lambda_{cycle} * \mathcal{L}_{cycle} + \lambda_{pixel} * \mathcal{L}_{pixel} \\ &+ \lambda_{gan} * (\mathcal{L}_{GAN} + \mathcal{L}_{AGAN}) + \lambda_{tv} * \mathcal{L}_{tv},\end{aligned}$$

The cycle-consistency loss can be used to enforce forward and backward consistency. The cycle-consistency loss can be regarded as "pseudo" pairs of training data even though we do not have the corresponding data in the target domain which corresponds to the input data from the source domain.

When training our AGGAN we do not have ground-truth annotation for the attention masks. They are learned from the resulting gradients of the attention-guided discriminators and the rest of the losses. However, the attention masks can easily saturate to 1 which makes the attention-guided generator has no effect as indicated in the GAN animation. To prevent this situation, we perform a Total Variation Regularization over attention masks M_y and M_x.

To reduce changes and constrain generators, we adopt pixel loss between the input images and the generated images. We adopt L1 distance as loss measurement in pixel loss.

2.3 Conclusion

We propose using AGGAN to translate images from the face mask domain to the unmasked domain by focusing on the image area that contains the facial mask. The generators in AttentionGAN have the built-in attention mechanism, which can preserve the background of the input images and discover the most discriminative content between the source and target domains by producing attention masks and content masks. We can then utilize the input image, the attention mask, and the content mask to generate unmasked images. The attention-guided discriminator also focuses only on the attention area. Some of the results obtained by using such a technique are presented below:

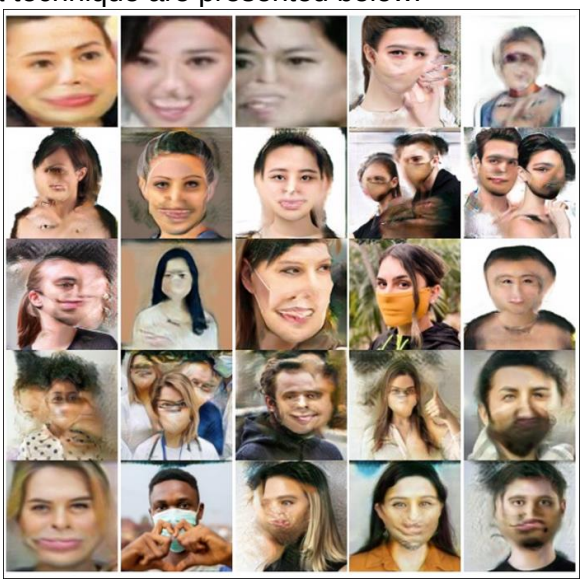


Fig: Masked Images used as test images and inputted into the network

Fig: Fake Images generated through the AGGAN network

By looking at the fake images we can say that the network does a considerably good job when it comes to removing face masks from the input images but does require some further enhancements. For example, when the color of the face mask matches the color of the face, the discriminator is not able to differentiate among those, and hence the generated image doesn't undergo many changes. Also, the network is not able to remove the masks in case of images where the hands interfere with the mask and hence no changes are done to those images. Additionally, for certain images, we see that the structure of the face is not proper especially when there are multiple people or if people are wearing caps in the image. Apart from these specific instances, the AGGAN does a relatively better job even when this problem requires a non-linear transformation of the images. The attention-guided mechanism does require some improvements to make the network more robust to any kind of human face image with a face mask and computationally remove the face mask to generate fake images. Alternatively, we can try several other approaches and hyperparameter tuning to improve the network.

One approach could be ContraGAN which considers relations between multiple image embeddings in the same batch (data-to-data relations) as well as the data-to-class relations by using a conditional contrastive loss. The discriminator of ContraGAN discriminates against the authenticity of given samples and minimizes a contrastive objective to learn the relations between training images. This, as presented by another team produced more realistic results and did a better job in removing the face masks. However, we can improve our model in the following ways:

- Increasing the total number of epochs on which we are training our model with a variety of images. We could increase the epochs to 300 and simultaneously check the improvements.
- We can tune the hyperparameters in such a way that they make the model fit better with the existing dataset. Here, as we are dealing with different losses, changing the regularization parameters for identity, cyclic, and attention losses can give us varied results. Also carefully training the model by sub setting the training pictures to make sets containing identical masks can be explored.
- Additional augmentations can be added and used to train the network such that the network identifies pictures where there are people wearing multiple masks and is able to deal with corner cases.

These improvements can lead to a better image-image translation of unmasked images from the masked ones and can help us fit this complex model into other facial disfigurements to generate high-quality people’s images.