StarGAN

Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation

CVPR 2018

Guide: Prof. Ahlad Kumar

Motivation

Image editing

- Low complexity
- No comprehension of scene or object
- Common operations - filtering

Image to Image translation

- High complexity (challenging modification)
- Learn the mapping between input image and output image
- Common operation - style transfer

Solution for non trivial tasks - **Generative models!**

Role

01

SUMMARIZER (Introduction)

Niharika Dalsania - 201701438

Prior Work

DiscoGAN

Discover relations between different domains and successfully transfer style from one domain to another

Source: https://arxiv.org/pdf/1703.05192v2.pdf

CycleGAN

Translate an image from a source domain X to a target domain Y in the absence of paired examples

Source: https://arxiv.org/pdf/1703.10593.pdf

IcGAN

Identify the latent representation of image using encoder, and modify any attribute to get desired results

Source: https://arxiv.org/pdf/1611.06355.pdf

DIAT

Generate a facial image that owns the reference attribute and keeps similar identity to the input image

Source: https://arxiv.org/pdf/1610.05586.pdf

Major issues...

- Addresses only translation within two domains
- Very inefficient and cumbersome training for multi domain translation
- Low visual quality results - blurred and distorted

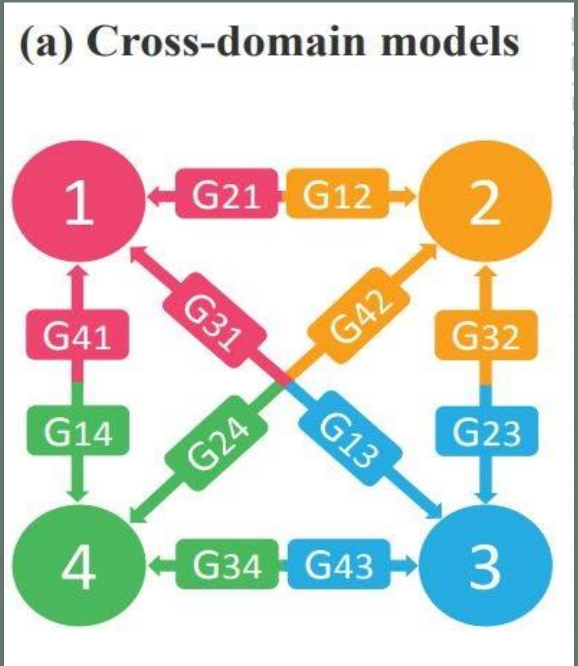
- Some **major amendments needed!** .... not just in the model architecture, but in the underlying training process itself

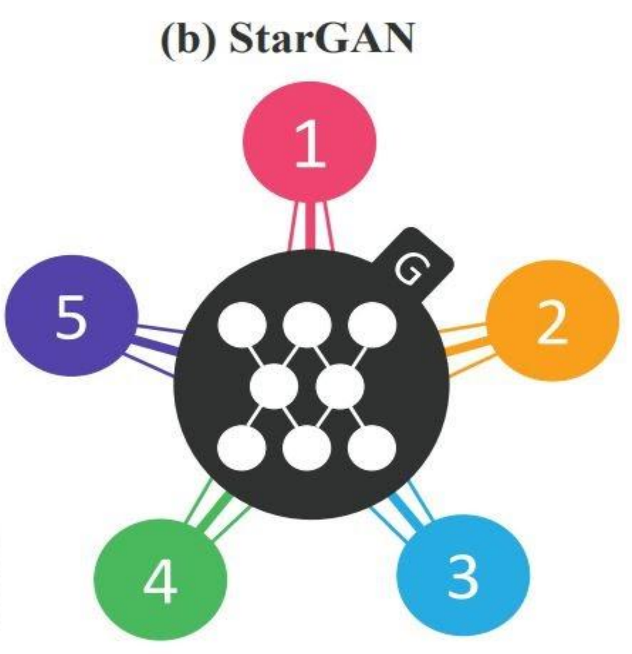
A fresh approach... StarGAN

- **Intuition** behind the new approach
  - Robustness in learning through a multi-task learning framework from very big and different feature datasets
  - Train model to flexibly translate images according to the labels of the target domain

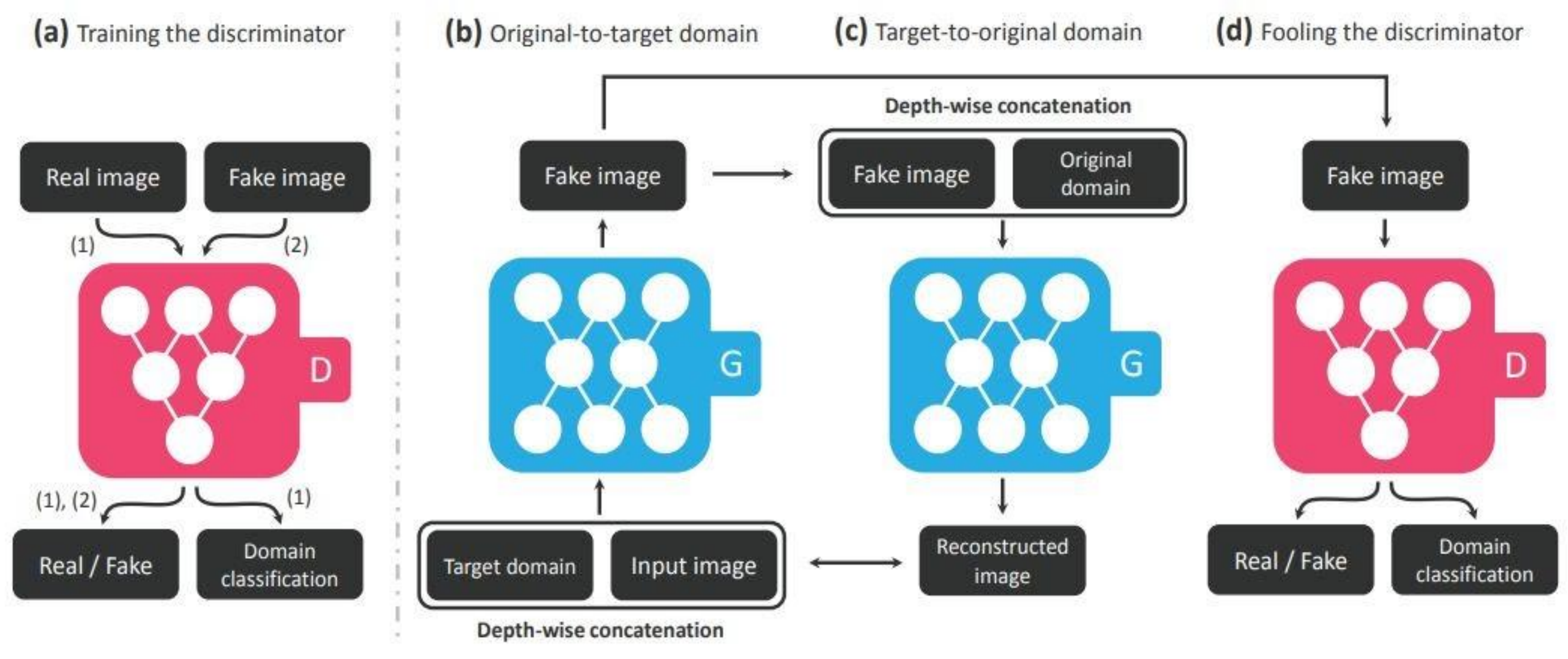
- **Rationale**? Why will this **address the flaws**?
  - Will help improve sharpness of features and reduce efforts of training for each source target pair
  - Not prone to overfitting, as opposed to training a model to perform a fixed translation

The core idea





How will it work?



Looking inside the network - Generator

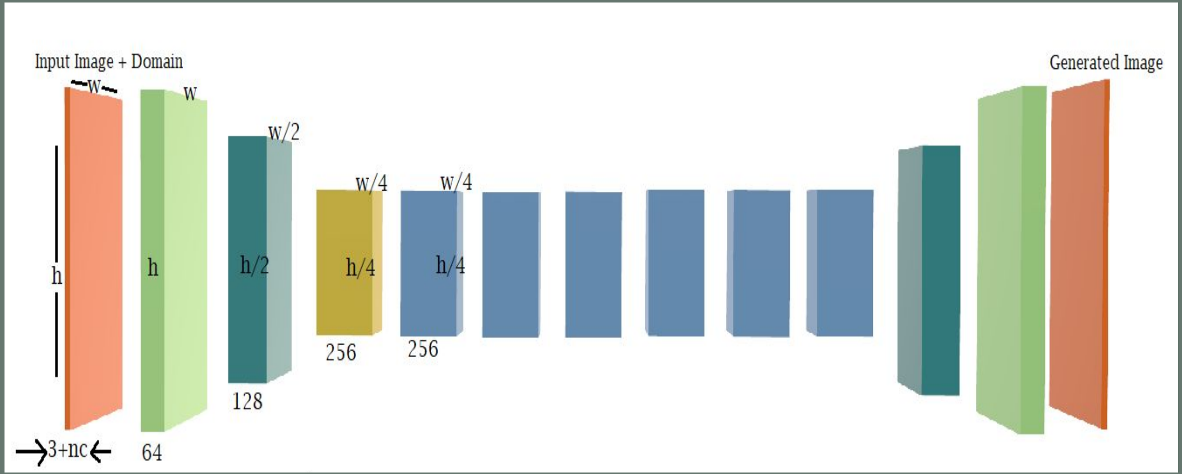
- Downsampling

  2 convolutional layers with the stride size of 2

- 6 residual blocks

- Upsampling

  2 transposed convolutional layers with the stride size of 2



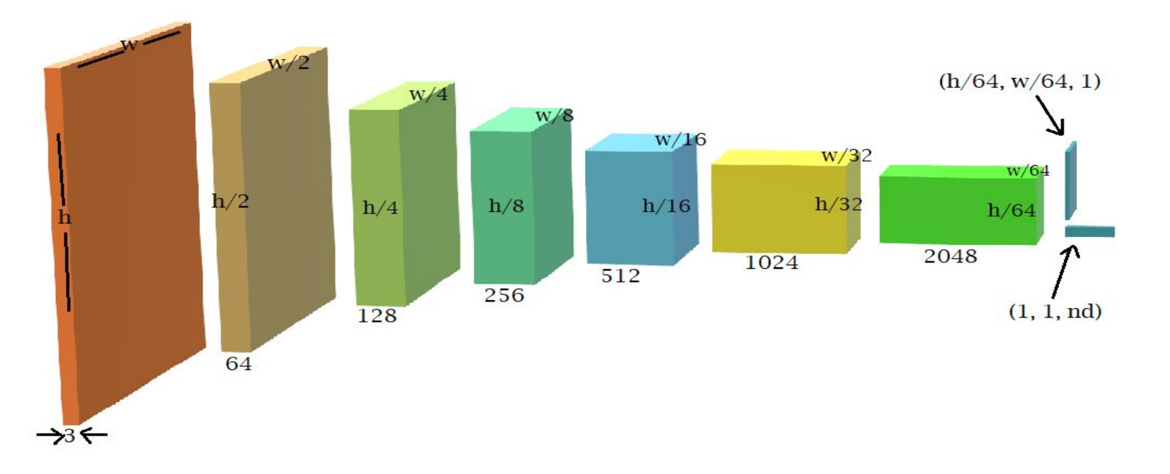
Looking inside the network - Discriminator

- Input Layer
- 5 hidden layers

  convolutional layers with the stride size of 2

- 2 Output layers

  Domain Classification & Real/Fake Identification



Role

02

ADVOCATE

Zeel Patel - 201701443

Losses

Adversarial Loss (GAN Loss)

at Discriminator D	at Generator G
D (x) → should be maximized D (G(z)) → should be minimized	D (G(z)) → should be maximized

$$\mathcal{L}_{adv} = \mathbb{E}_x\left[\log D_{src}(x)\right] + \mathbb{E}_{x,c}[\log\left(1 - D_{src}(G(x,c))\right)],$$

<u>**Dsrc**</u> → Probability distribution of being real or fake
<u>**Dcls**</u> → Probability distribution over domain labels
Fake image → G(x,c)
x → real image
c → target domain

Losses (contd.)

Domain Classification Loss

- Task of Generate → generate an image which is classified in the target domain.
- Hence, error in classifying fake → to train generator

$$\mathcal{L}_{cls}^{r} = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)],$$

- Task of discriminator → Detect fake image
- Hence, error in classifying real → to train discriminator

$$\mathcal{L}_{cls}^{f} = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x,c))].$$

Losses (contd.)

Reconstruction Loss

- Just like cycle GAN,(cycle consistency loss)
- L1 Norm

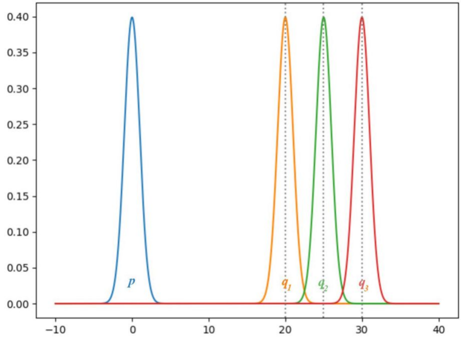
$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}[||x - G(G(x,c),c')||_1],$$

Final Objective

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}^r_{cls},$$

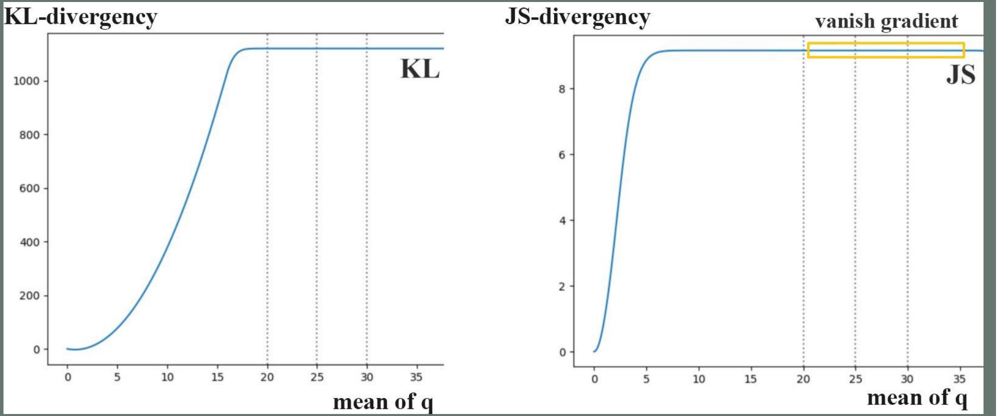
$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}^f_{cls} + \lambda_{rec}\,\mathcal{L}_{rec},$$

Why Gradient Penalty? (Long Story)



- p → original data distribution , q→ generated
- Discriminator is trained first.
- Minimizing the GAN objective function with an **optimal discriminator** is equivalent to **minimizing the JS-divergence.**

- If the generated image has distribution q far away from the ground truth p, the generator barely learns anything because of vanishing Gradient.



Source: https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

Wasserstein Distance

- Alternative cost function to address this gradient vanishing problem is reverse JS divergence and adding noise.
- But it has some limits as well.
- **Wasserstein distance:** minimum cost of transporting mass in converting the data distribution q to the data distribution p. **(We look at horizontal distance)**

- Discriminator → Critic
- Hence → No sigmoid layer at last
- Weights → Clipped
- Critic → G,D functionality of original GAN is not there

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y)\sim\gamma}\left[\, \|x - y\| \,\right]$$

$$\nabla_w \left[\frac{1}{m} \sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m} \sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$$

Source: https://jonathan-hui.medium.com/gan-wasserstein-gan-wgan-gp-6a1a2aa1b490

Now What?

- Weight Clipping is a terrible solution.
- Slow convergence after weight clipping (when clipping window is too large), and vanishing gradients (when clipping window is too small).
- Solved by Gradient Penalty.
- Points interpolated between the real and generated data should have a gradient norm of 1 for D.

$$\begin{aligned}\mathcal{L}_{adv} = & \mathbb{E}_x[D_{src}(x)] - \mathbb{E}_{x,c}[D_{src}(G(x,c))] \\ & - \lambda_{gp}\, \mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D_{src}(\hat{x})||_2 - 1)^2],\end{aligned}$$

Role

03

ADVOCATE

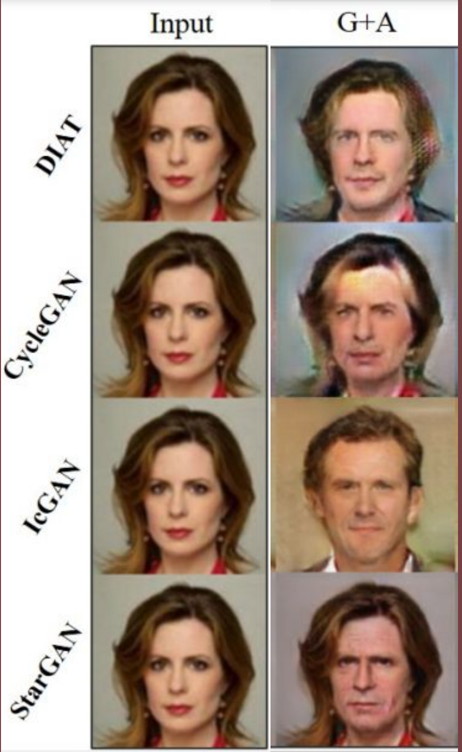
Darshan Patel - 201701436

Multi Domain Translation

```
# Core Algorithm
Shuffle Data
Divide into batches
for every epoch
  for every iteration
    fetch respective batch
    choose a random target label
    train G & D on batch for converting to target label
```

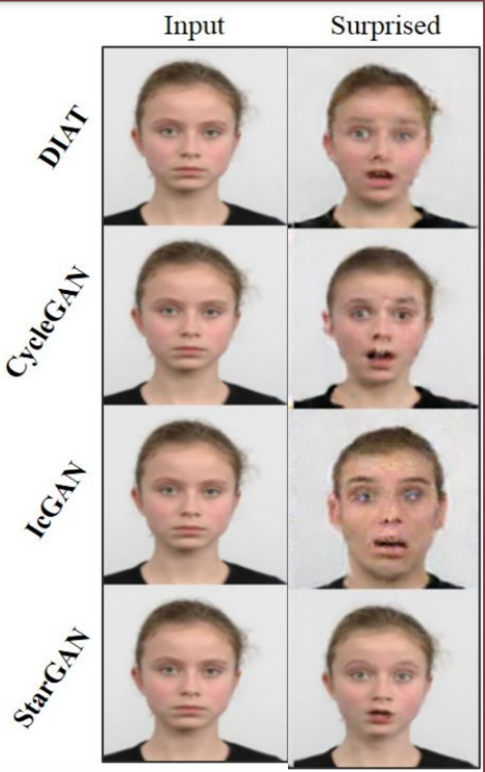
- Multi Domain Image to Image translation becomes possible
- Robust and Effective Implementation

Qualitative Analysis - CelebA



- The **Regularization effect** of StarGAN through a multi-task learning framework.
- Compared to IcGAN, starGAN shows an advantage in **preserving the facial identity** feature of any input

Qualitative Analysis - RaFD



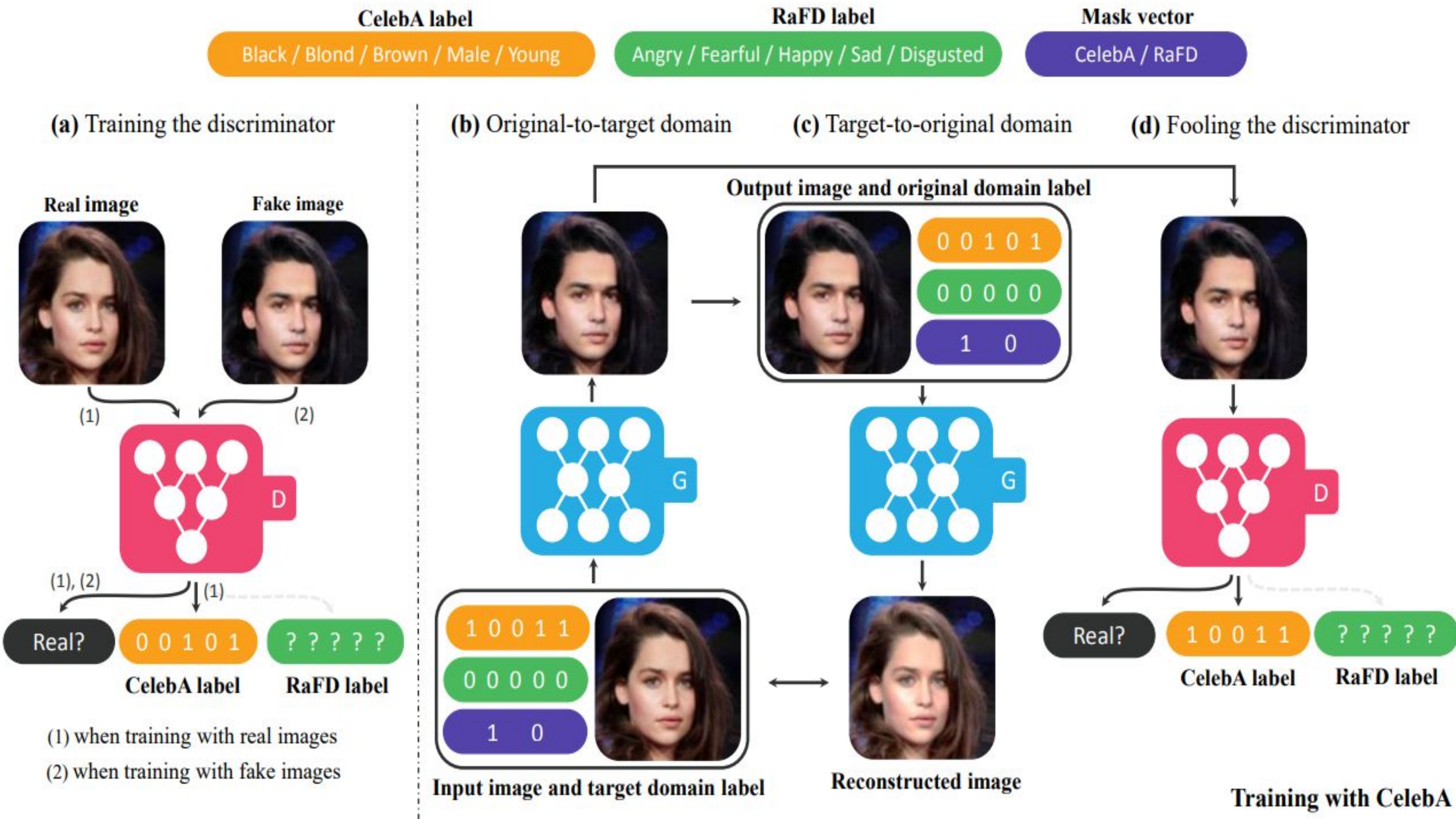
- DIAT and CycleGAN mostly **preserve the identity** , but blurry results.
- starGAN shows sharp results while preserving the identity, because of Implicit **data augmentation** effect from a multi-task learning setting
- $X\_i = 500$, $n = 8$, $X = 4000$

Multiple Dataset Training

- Able to simultaneously incorporate multiple datasets
- Proposing the "Mask Vector" denoted as "m"

$$G(x, c) \rightarrow y$$

$$\tilde{c} = [c_1, \ldots, c_n, m]$$


CelebA label
Black / Blond / Brown / Male / Young
RaFD label
Angry / Fearful / Happy / Sad / Disgusted
Mask vector
CelebA / RaFD
(a) Training the discriminator
(b) Original-to-target domain
(c) Target-to-original domain
(d) Fooling the discriminator
Real image
Fake image
(1)
(2)
D
(1), (2)
(1)
Real?
0 0 1 0 1
? ? ? ? ?
CelebA label
RaFD label
(1) when training with real images
(2) when training with fake images
Output image and original domain label
0 0 1 0 1
0 0 0 0 0
1 0
G
G
1 0 0 1 1
0 0 0 0 0
1 0
Input image and target domain label
Reconstructed image
D
Real?
1 0 0 1 1
? ? ? ? ?
CelebA label
RaFD label
Training with CelebA

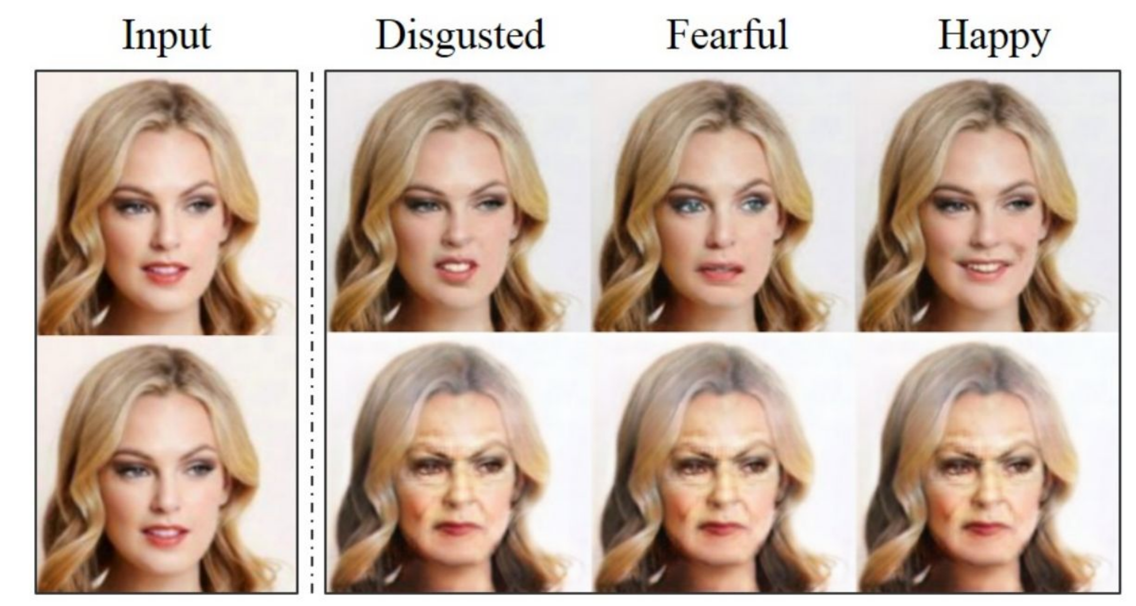
Importance of Joint Training

- Improvement in shared low-level tasks such as facial keypoint detection and segmentation

Importance of mask vector

Proper mask vector

Wrong mask vector



Role

04

DEVIL'S ADVOCATE

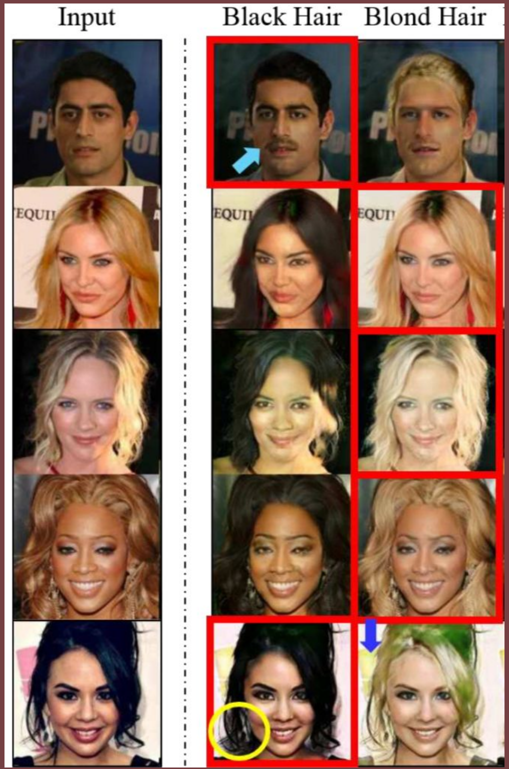
Ruchit Shah - 201701435

A look at the Results...

- Results generated from the code
  - Images stored directly in drive folder

- Results from the pre-trained model
  - Images stored directly in local machine

Drawbacks of StarGAN



- StarGAN tends to make unnecessary changes during cross-domain translation.
  - Alters the face colour
  - Unnecessarily changes the background

- StarGAN fails to competently handle same-domain translation
  - Adds a moustache to the face
  - Adds extra hair

Image source : Learning Fixed Points in Generative Adversarial Networks:From Image-to-Image Translation to Disease Detection and Localization , *Md Mahfuzur Rahman Siddiquee et al.*

Rectifications

StarGAN Loss Equations

$$\mathcal{L}_{adv} = \mathbb{E}_x\left[\log D_{src}(x)\right] + \mathbb{E}_{x,c}[\log(1 - D_{src}(G(x,c)))],$$

$$\mathcal{L}_{cls}^r = \mathbb{E}_{x,c'}[-\log D_{cls}(c'|x)],$$

$$\mathcal{L}_{cls}^f = \mathbb{E}_{x,c}[-\log D_{cls}(c|G(x,c))].$$

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}[||x - G(G(x,c),c')||_1],$$

$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}_{cls}^r,$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{cls}\,\mathcal{L}_{cls}^f + \lambda_{rec}\,\mathcal{L}_{rec},$$

Rectified Loss Equations

$$\mathcal{L}_{adv} = \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c}\left[\log\left(1 - D_{r/f}(G(x,c))\right)\right] + \mathbb{E}_x[\log D_{r/f}(x)]$$

$$\mathcal{L}_{domain}^r = \mathbb{E}_{x,c_x}[-\log D_{domain}(c_x|x)]$$

$$\mathcal{L}_{domain}^f = \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c}[-\log D_{domain}(c|G(x,c))]$$

$$\mathcal{L}_{cyc} = \sum_{c\in\{c_x,c_y\}} \mathbb{E}_{x,c_x,c}[||G(G(x,c),c_x) - x||_1]$$

$$\mathcal{L}_{id} = \mathbb{E}_{x,c}\,[||G(x,c) - x||_1] \text{ if } c = c_x;\ 0 \text{ otherwise}$$

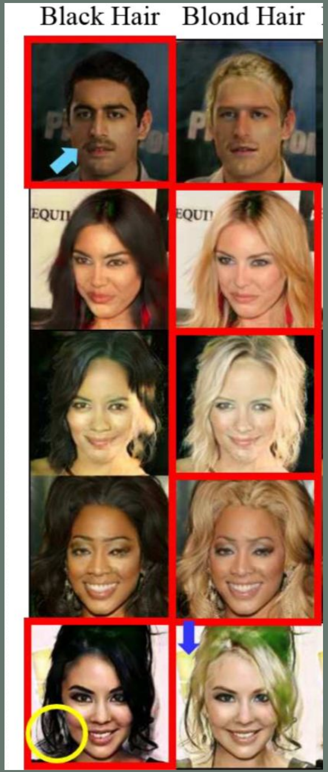
$$\mathcal{L}_D = -\mathcal{L}_{adv} + \lambda_{domain}\mathcal{L}_{domain}^r$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{domain}\mathcal{L}_{domain}^f + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id}$$

Source : Learning Fixed Points in Generative Adversarial Networks:From Image-to-Image Translation to Disease Detection and Localization , *Md Mahfuzur Rahman Siddiquee et al.*

Improvement in Results

StarGAN



Rectified version

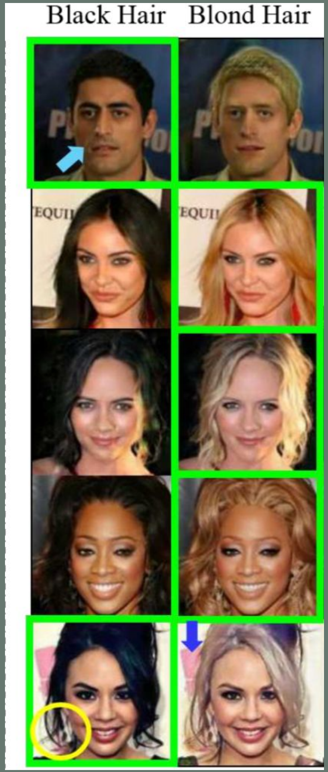


Image source : Learning Fixed Points in Generative Adversarial Networks:From Image-to-Image Translation to Disease Detection and Localization , *Md Mahfuzur*

Yet another Drawback...

- Methods used in StarGAN assume binary-valued attributes and thus cannot yield satisfactory results for fine-grained control.

- These methods require specifying the entire set of target attributes, even if most of the attributes would not be changed.

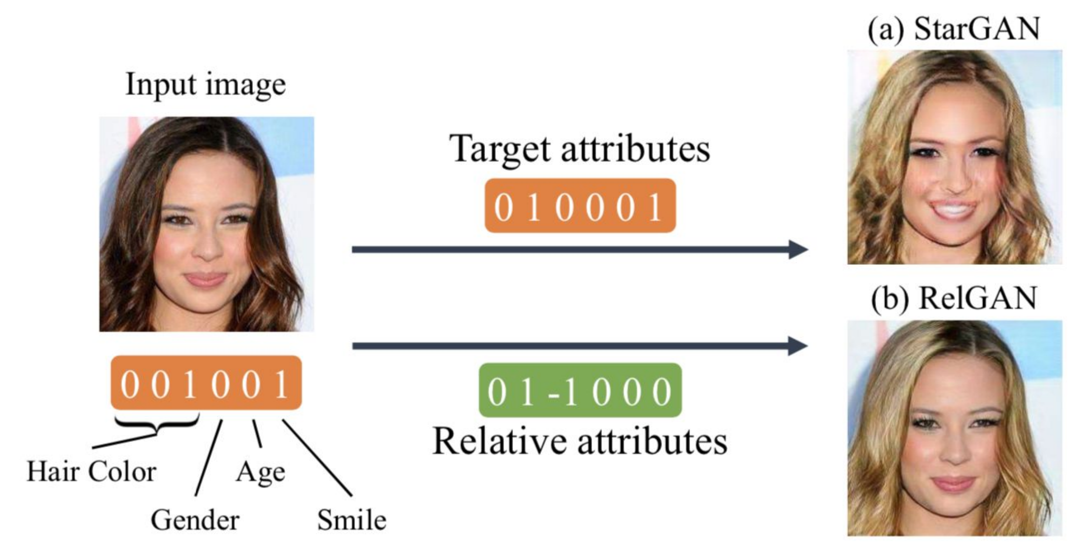


Image Source : RelGAN: Multi-Domain Image-to-Image Translation via Relative Attributes

Quantitative Support

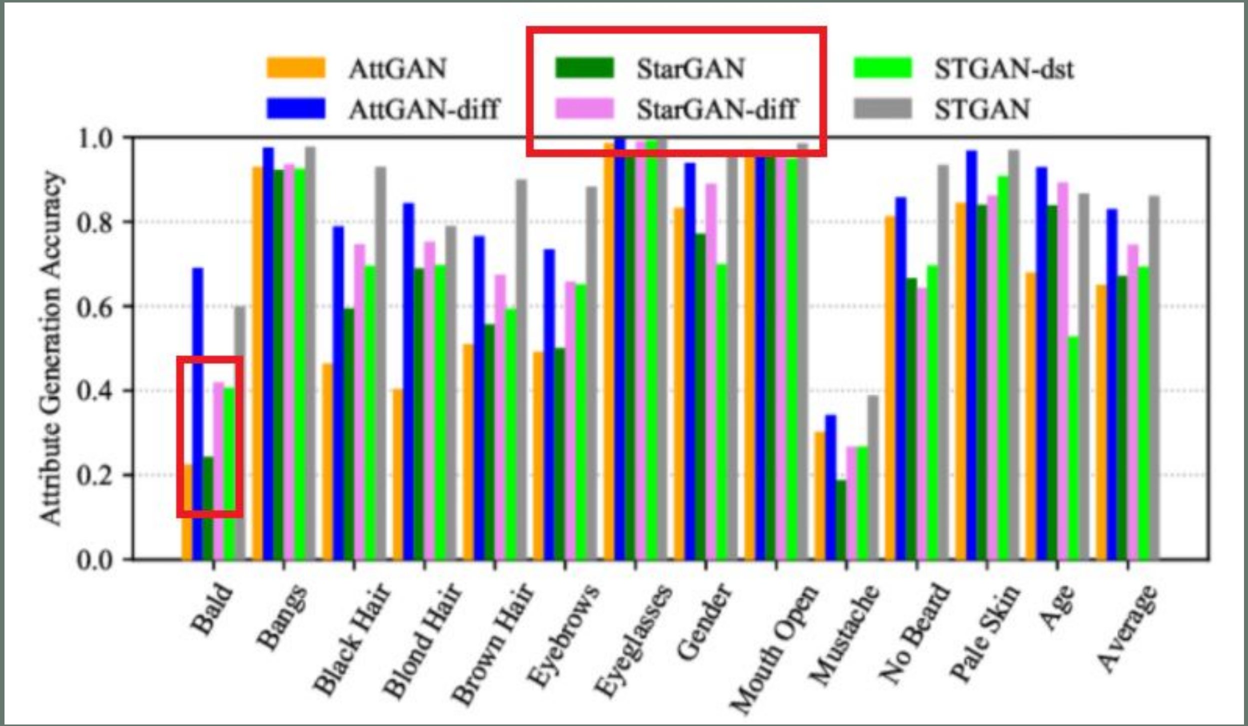


Image Source : STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing

Role

01

SUMMARIZER (Conclusion)

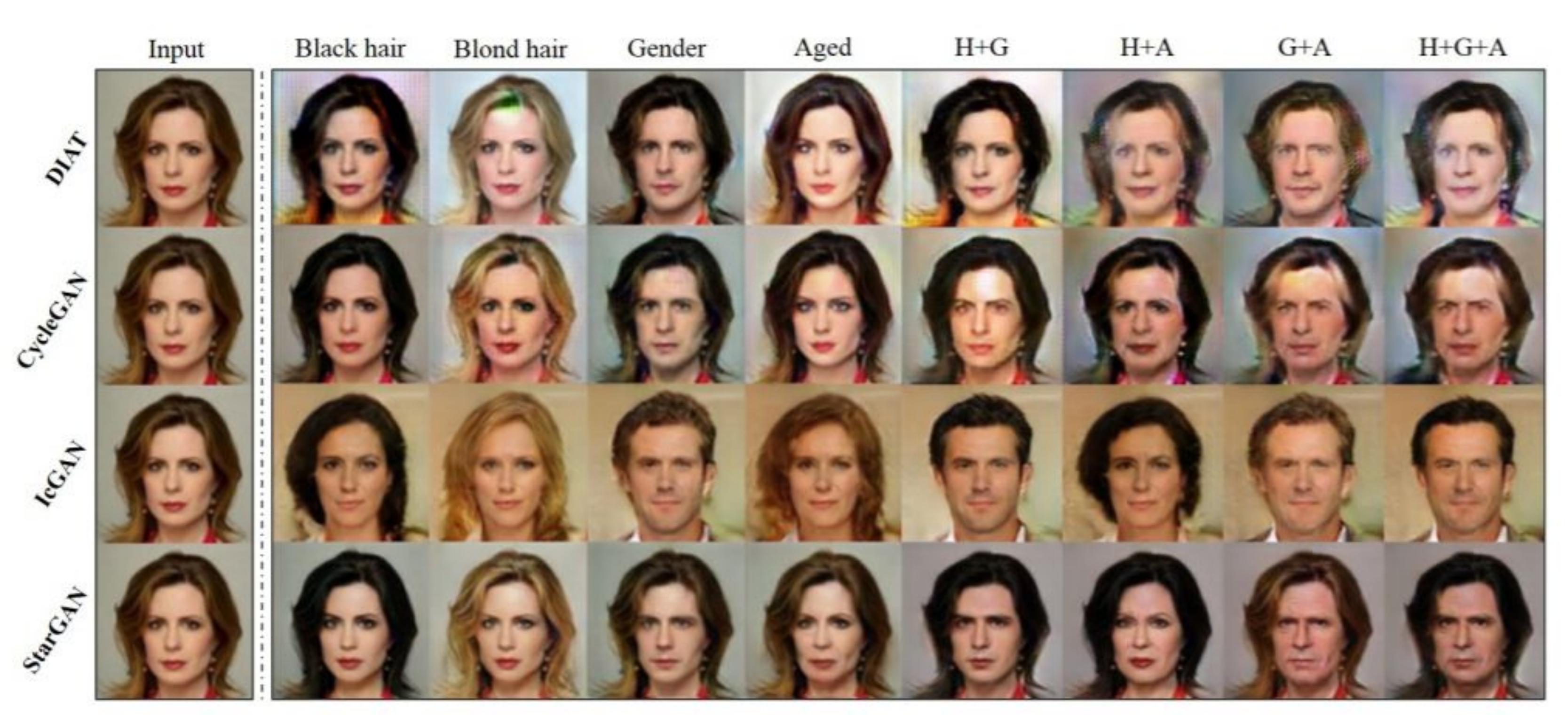
Niharika Dalsania - 201701438

Major Takeaways...

- Existing methods of image to image translation
  - Two domain translation
  - Inefficient training
  - Blurry, distorted results

- A novel approach - StarGAN
  - Multi domain translation
  - Joint training over multiple datasets (mask vector!)
  - Visually superior results compared to previous results
  - Not yet perfect! .... fails for same domain translation

The plus points...



The minus points...

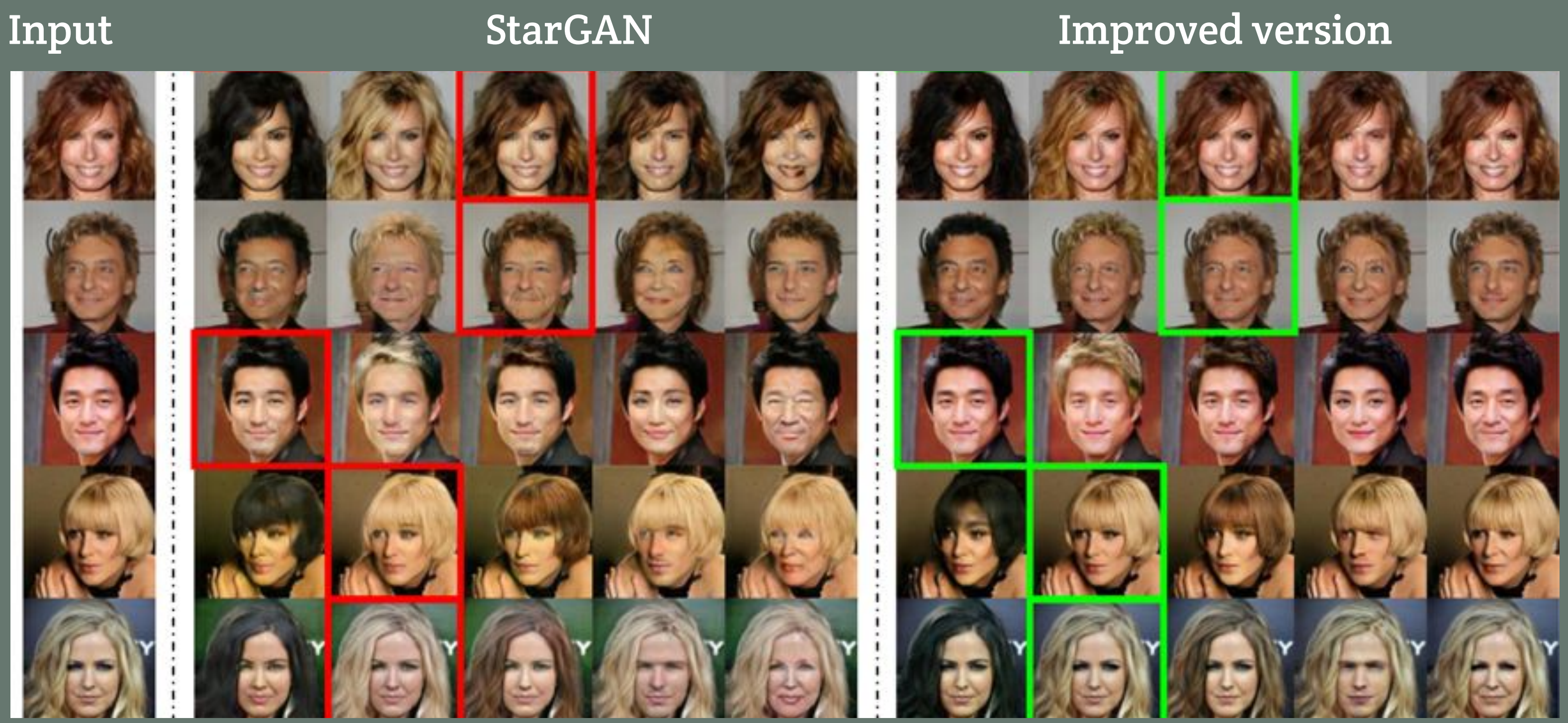


Image Source : Learning Fixed Points in Generative Adversarial Networks:From Image-to-Image Translation to Disease Detection and Localization , *Md Mahfuzur Rahman Siddiquee et al.*

StarGAN in a nutshell…

- An extremely scalable model with high visual quality generated image owing to the generalization capability behind the multi-task learning setting.
- Yet, there are some flaws addressing which researchers can develop superior image translation applications across multiple domains.

Thank you!

Group 7

Ruchit(201701435), Darshan(201701436), Niharika(201701438), Zeel(201701443)