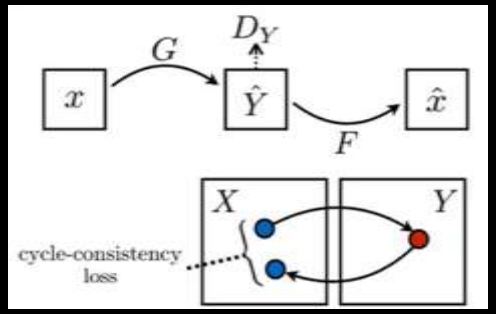
Paper Reading

GANimation: Anatomically-aware Facial Animation from a Single Image (ECCV 2018)

艺术学院　朱峰　1801221169

Related Work: Image-to-Image Translation

As in this framework, several works have also tackled the problem of using unpaired training data. This approach is more related to those works exploiting cycle consistency to preserve key attributes between the input and the mapped image, such as CycleGAN, DiscoGAN and StarGAN.



CycleGAN

StarGAN

Introduction

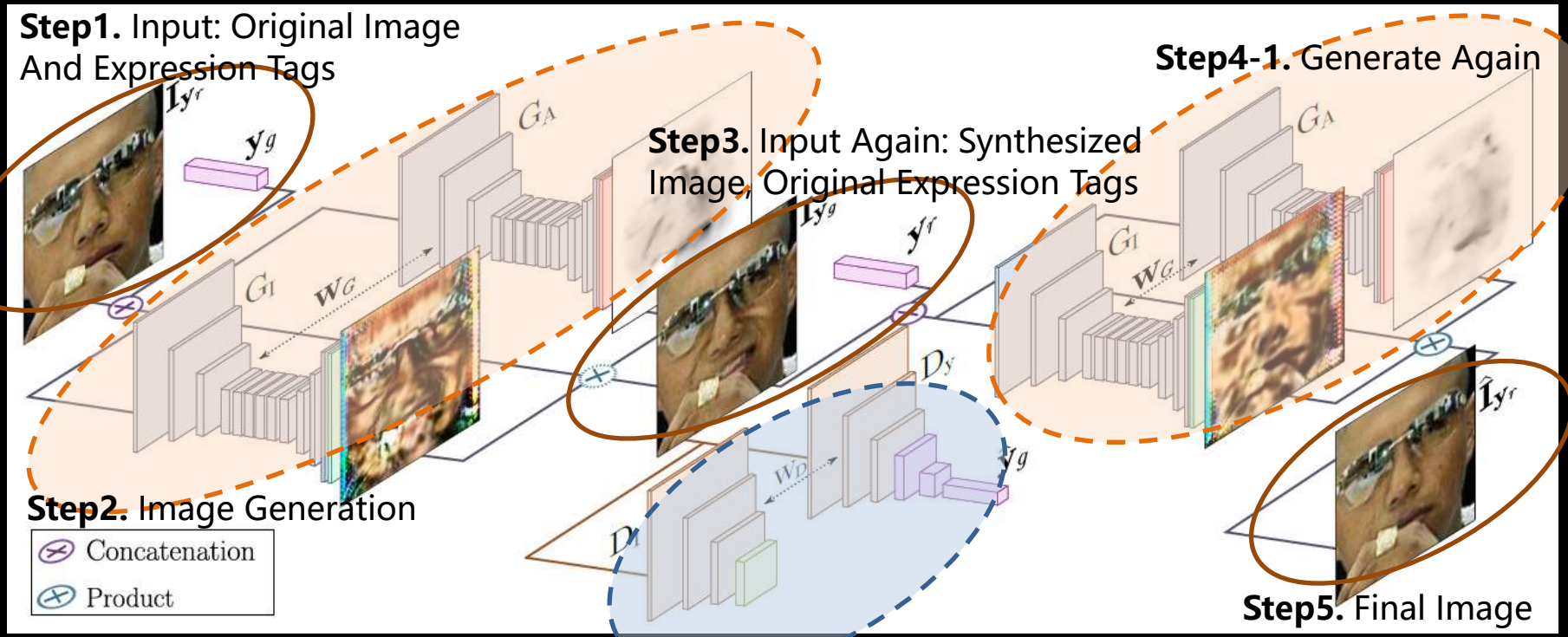
- **StarGAN** can only generate a discrete number of expressions, determined by the content of the dataset.

- This paper introduces a novel GAN conditioning scheme based on **Action Units (AU) annotations,** which describes in a continuous manifold the anatomical facial movements defining a human expression.

- We leverage on the recent **EmotioNet dataset**, which consists of one million images of facial expressions (we use 200,000 of them) of emotion in the wild annotated with discrete AUs activations.

- Additionally, we propose a fully unsupervised strategy to train the model, that only requires images annotated with their activated AUs, and exploit **Attention Mechanisms** that make our network robust to changing backgrounds and lighting conditions.

Framework

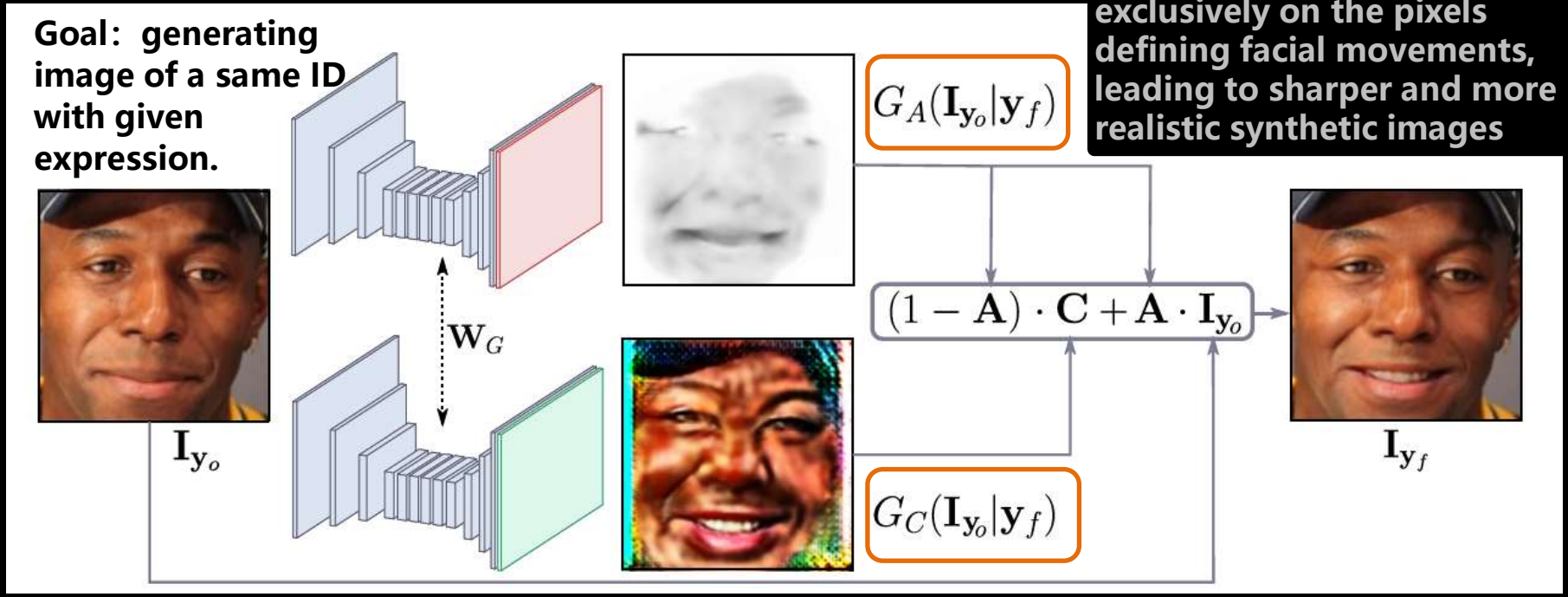


Two Blocks:
Generator: GA, GI
Discriminator: DI, Dy

Step4-2. Discriminator: evaluate the quality of the generated image and its expression

Attention-based Generator

Attention mask A:
the generator can focus exclusively on the pixels defining facial movements, leading to sharper and more realistic synthetic images

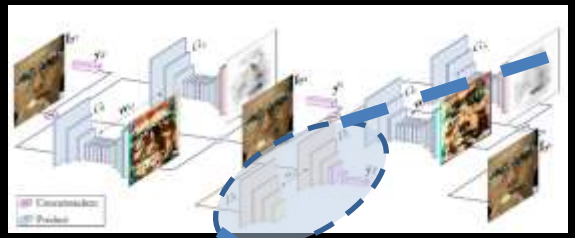


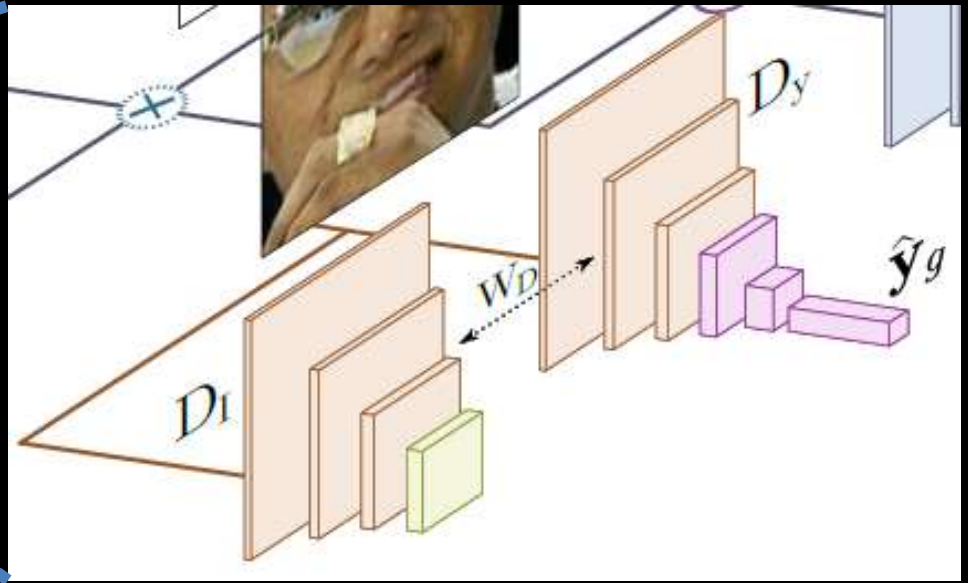
Two Blocks:
Generator: GA, GI
Discriminator: DI, Dy

Color mask C: the generator does not need to render static elements

Discriminator

Goal: Evaluate the quality of the generated image (Real Photo? Given ID?) and its expression (Given tag?)





Two Blocks:
Generator: GA, GI
Discriminator: DI, Dy

Loss Function

Goal of Generator: Generating image of a same ID with given expression.

Goal of Discriminator: Evaluate the quality of the generated image (Real Photo? Given ID?) and its expression (Given tag?)

Gradient penalty

1 Image Adversarial Loss

$$\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}[D_{\mathrm{I}}(G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f))] - \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}[D_{\mathrm{I}}(\mathbf{I}_{\mathbf{y}_o})] + \lambda_{\mathrm{gp}}\mathbb{E}_{\tilde{I}\sim\mathbb{P}_{\tilde{I}}}\left[(\|\nabla_{\tilde{I}}D_{\mathrm{I}}(\tilde{I})\|_2 - 1)^2\right]$$

2 Attention Loss

$$\lambda_{\mathrm{TV}}\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\sum_{i,j}^{H,W}\left[(\mathbf{A}_{i+1,j} - \mathbf{A}_{i,j})^2 + (\mathbf{A}_{i,j+1} - \mathbf{A}_{i,j})^2\right]\right] + \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}[\|\mathbf{A}\|_2] \quad (2)$$

3 Conditional Expression Loss

$$\mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|D_{\mathrm{y}}(G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f)) - \mathbf{y}_f\|_2^2\right] + \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|D_{\mathrm{y}}(\mathbf{I}_{\mathbf{y}_o}) - \mathbf{y}_o\|_2^2\right]. \quad (3)$$

4 Identity Loss

$$\mathcal{L}_{\mathrm{idt}}(G, \mathbf{I}_{\mathbf{y}_o}, \mathbf{y}_o, \mathbf{y}_f) = \mathbb{E}_{\mathbf{I}_{\mathbf{y}_o}\sim\mathbb{P}_o}\left[\|G(G(\mathbf{I}_{\mathbf{y}_o}|\mathbf{y}_f)|\mathbf{y}_o) - \mathbf{I}_{\mathbf{y}_o}\|_1\right]. \quad (4)$$

Full Loss

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\mathrm{I}}(G, D_{\mathrm{I}}, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_g) + \lambda_{\mathrm{y}}\mathcal{L}_{\mathrm{y}}(G, D_{\mathrm{y}}, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_r, \mathbf{y}_g) \quad (5)\\ &+ \lambda_{\mathrm{A}}\left(\mathcal{L}_{\mathrm{A}}(G, \mathbf{I}_{\mathbf{y}_g}, \mathbf{y}_r) + \mathcal{L}_{\mathrm{A}}(G, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_g)\right) + \lambda_{\mathrm{idt}}\mathcal{L}_{\mathrm{idt}}(G, \mathbf{I}_{\mathbf{y}_r}, \mathbf{y}_r, \mathbf{y}_g).\end{aligned}$$

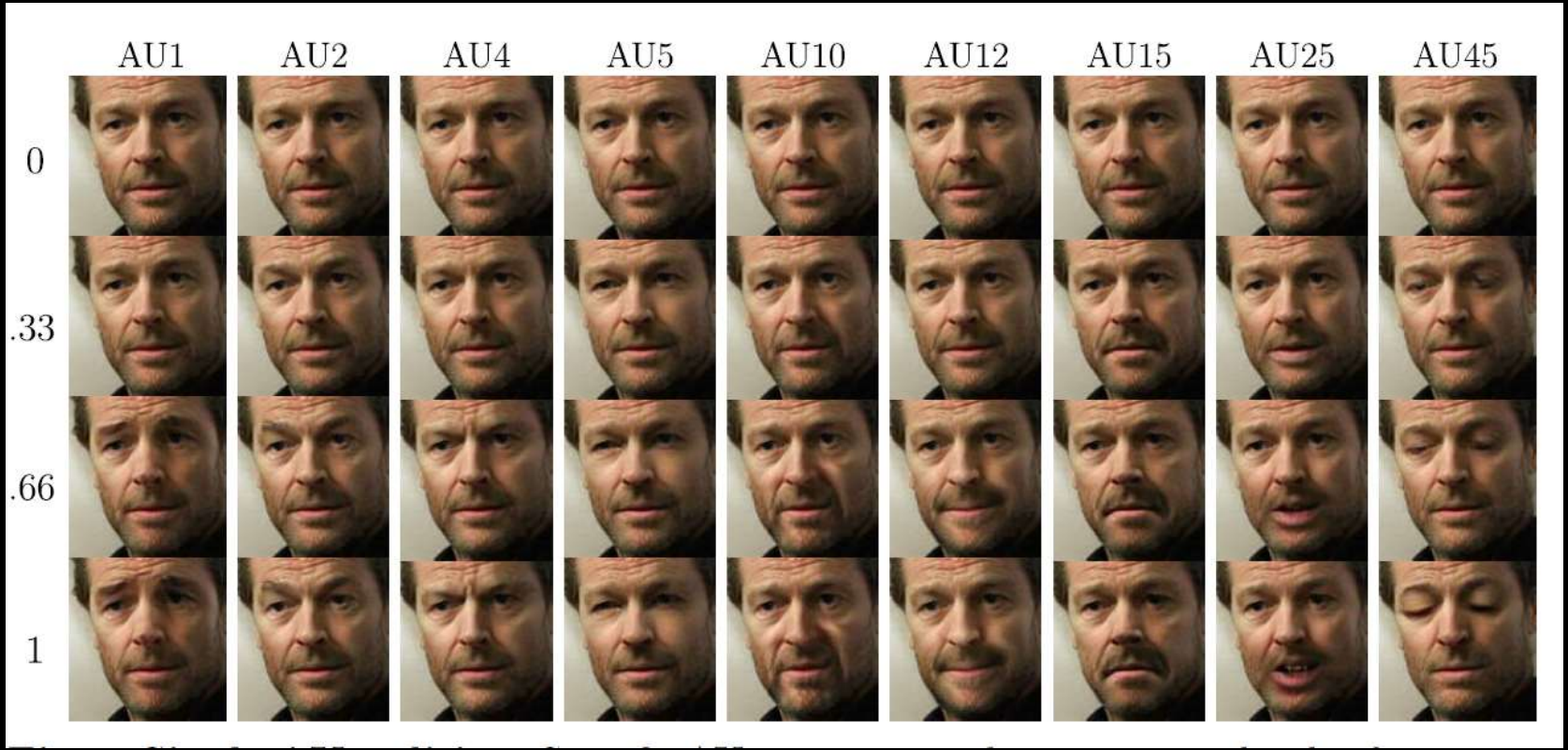
Implementation Details

Generator: built from **CycleGAN;** slightly modified by substituting the last convolutional layer with two parallel convolutional layers, one to regress the **color mask C** and the other to dene the **attention mask A.**

Discriminator: adopted the **PatchGan** architecture, with the **gradient penalty** computed with respect to the entire batch.

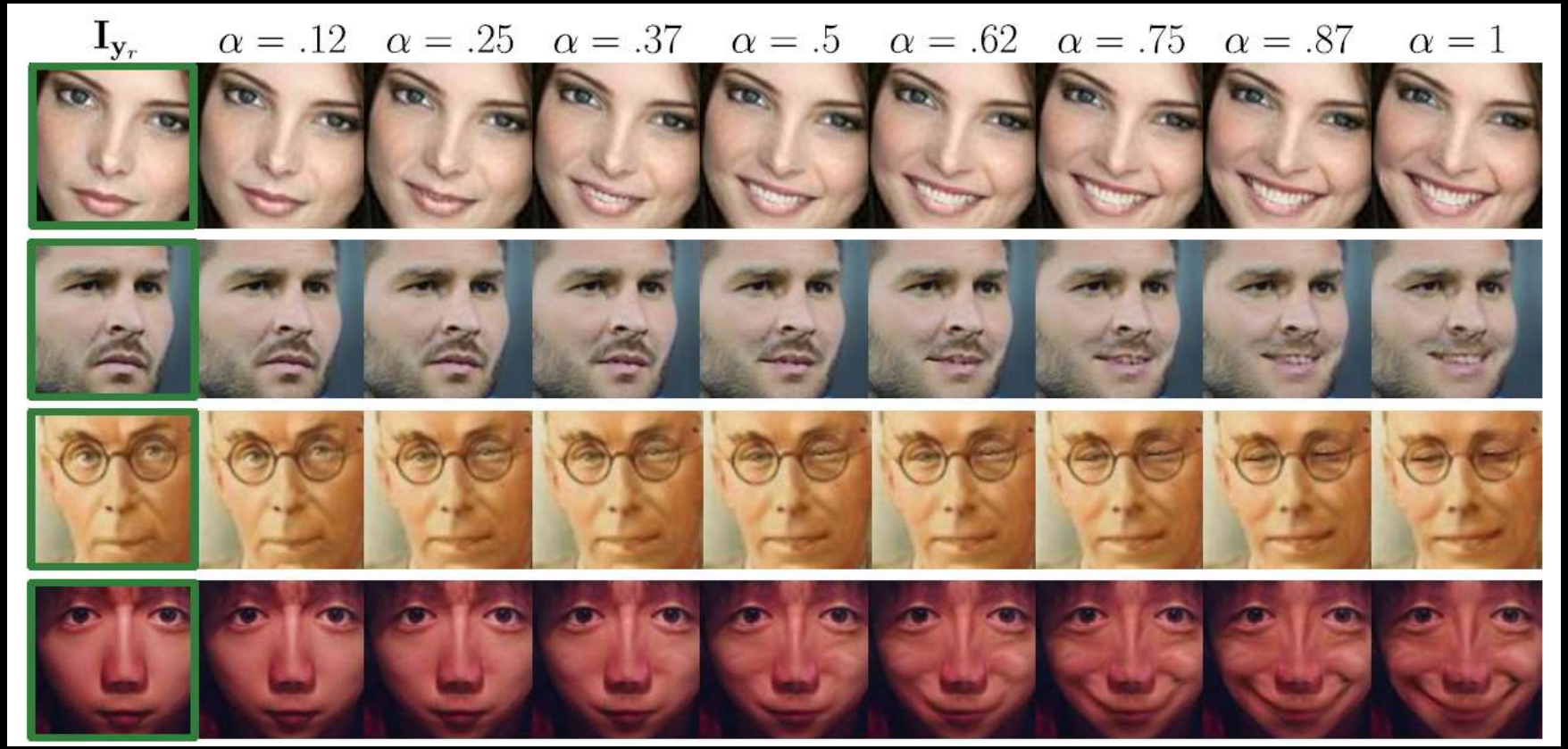
Other Details: The model is trained on the **EmotioNet dataset**. We use a subset of **200,000 samples (over 1 million)** to reduce training time. We use Adam with learning rate of 0.0001, beta1 0.5, beta2 0.999 and batch size 25. We train for 30 epochs and linearly decay the rate to zero over the last 10 epochs. Every 5 optimization steps of the critic network we perform a single optimization step of the generator. The model takes two days to train with a single GeForce GTX 1080 Ti GPU.

Results – Single AUs with Various Weights

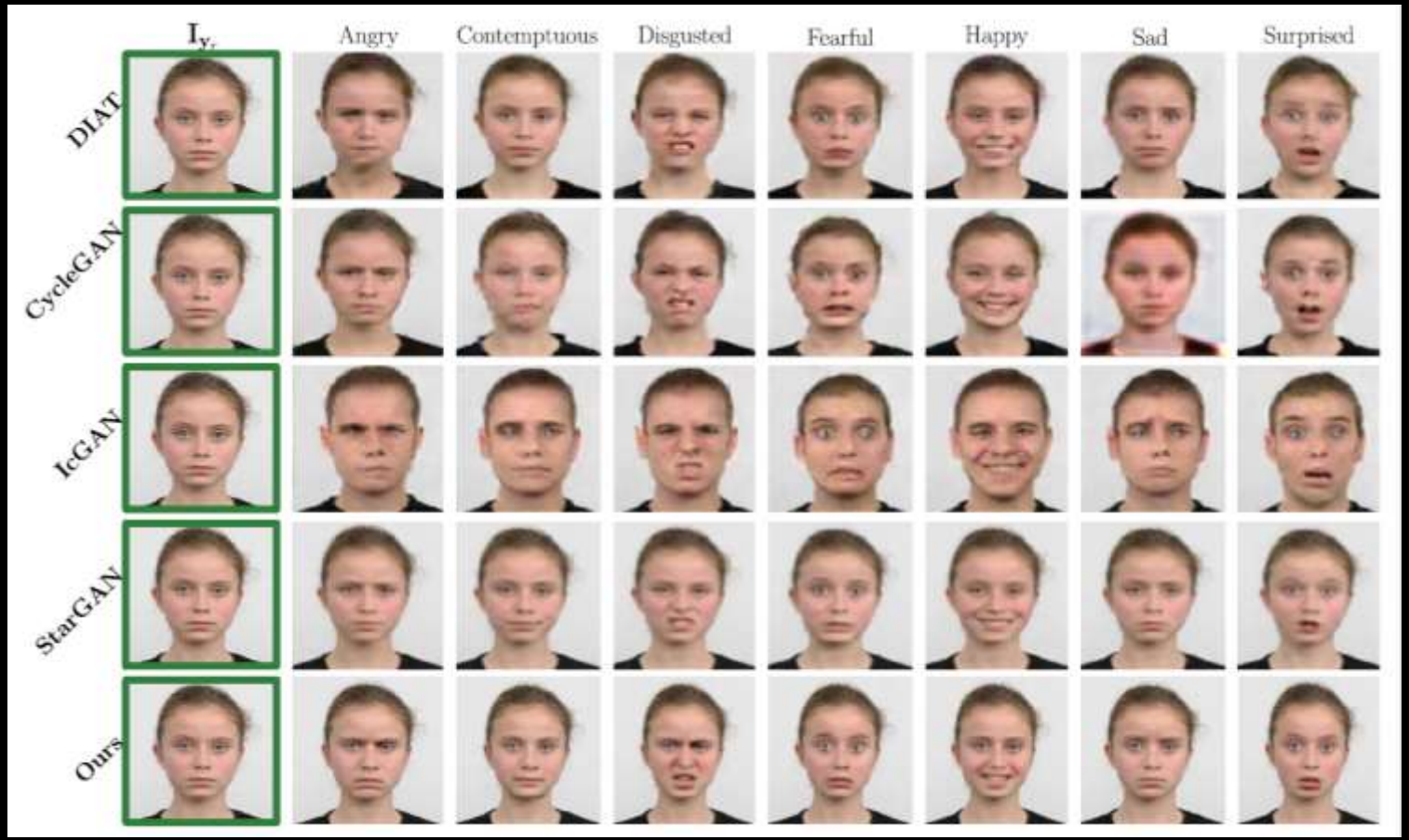


Results – Single AUs' Masks

Results – Multiple AUs

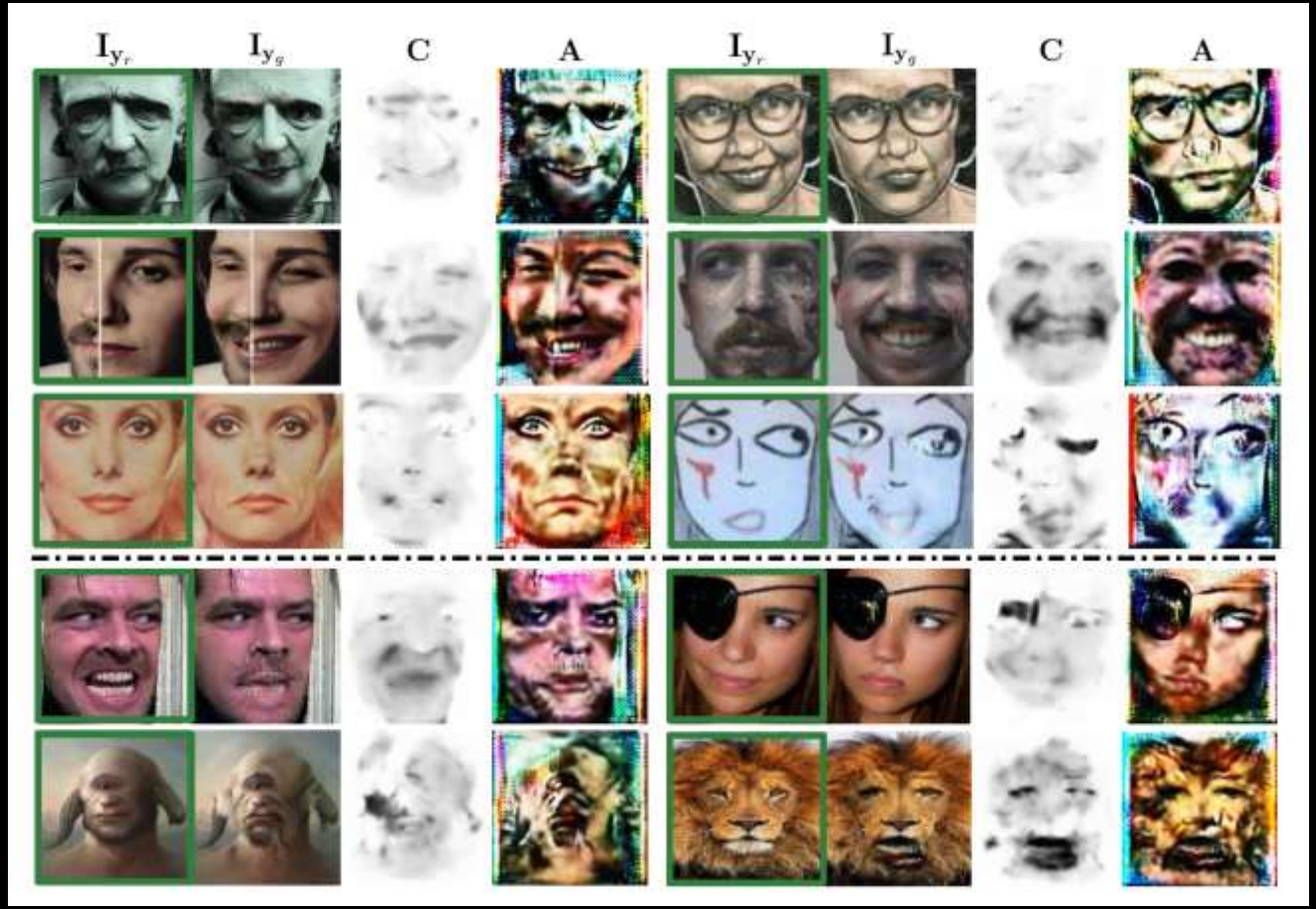


Results – Comparison with SOTA



Results –Evaluation on Images in the Wild

Results – Success and Failure Cases



Results – More

Thanks.