Conditional GAN based Mask Guided Portrait Editing

Presented by: Prabhakar Kumar Panday
Supervisors : Mr. Florian Strohm, VIS
Mr. George Basem Fouad Eskandar, ISS

Project Objective & Related Work

- Portrait editing with Conditional GAN based framework
- Using facial mask as reference guide for image generation.
- Generate high quality and diverse images with good controllability.

Related Work:

- Image style transfer using CNN[1]
- Neural color transfer between images [2]
- Paired-CycleGAN for makeup transfer [3] : Training makeup transfer and make removal function are trained in pairs.
- Facial component editing. [4,5,6]
- Face swapping. [7,8]
- Mask-Guided Portrait Editing with Conditional GANs [13]

Figure 1: Component Editing: Adding bangs

Figure 2: Style Transfer: Growing Beard

Background

Conditional GANs: [10]

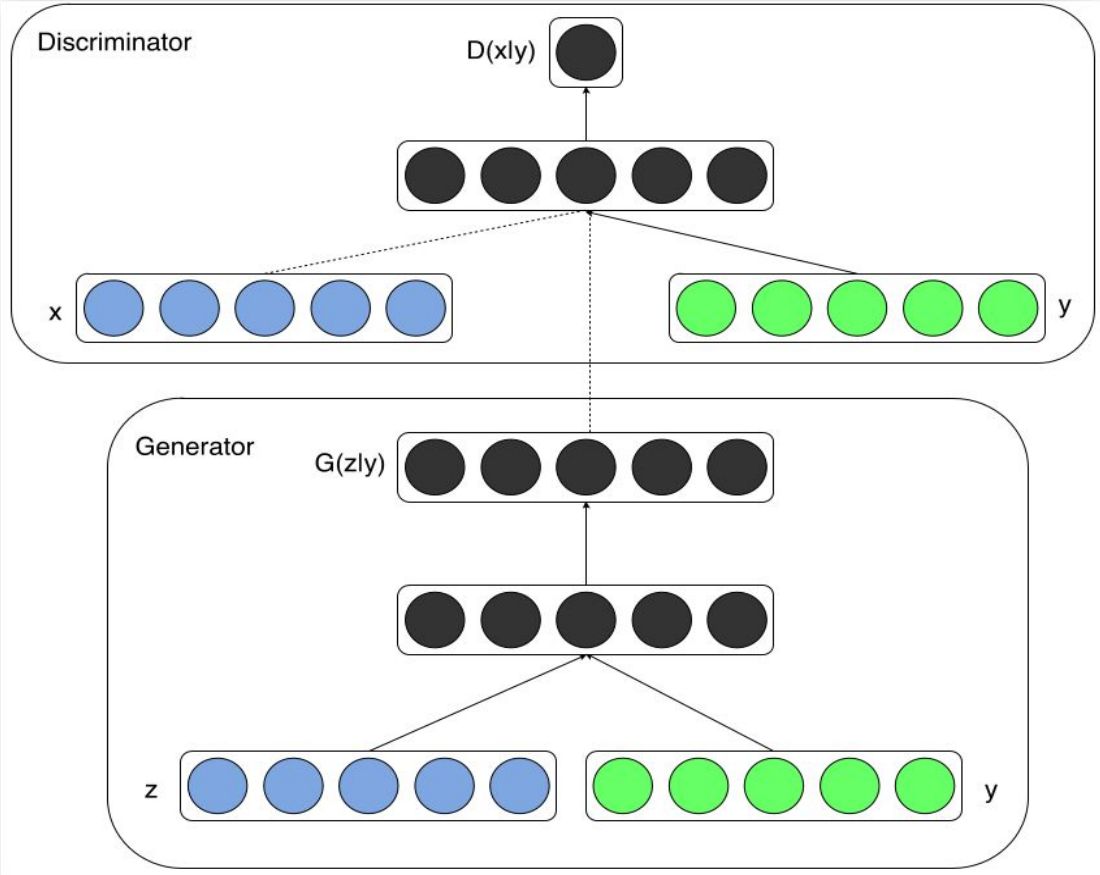

Figure 3: Structure of Conditional GAN

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)} \log [1 - D(G(z|y))]$$

Equation 1: Equation for the Conditional GAN

Background

Autoencoder

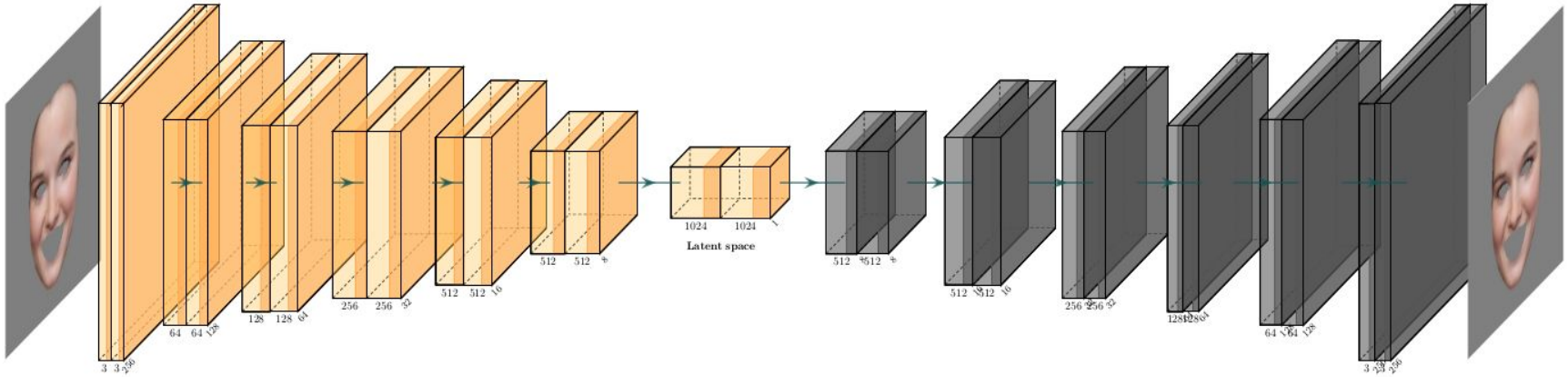

Figure 4: Architecture of one of the Autoencoder used in the project
Input tensor shape is equal to the out tensor's shape (256,256,3)

U-Net for face parsing network [11]

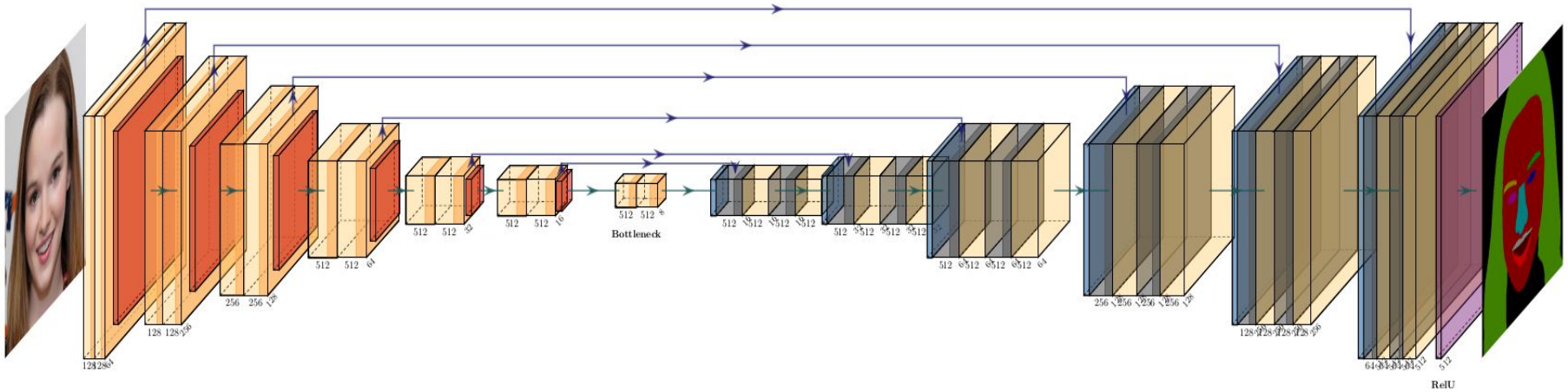

Figure 4: U-Net architecture used in the parsing network to evaluate the quality of the generated images

Mask-Guided Portrait Editing Framework

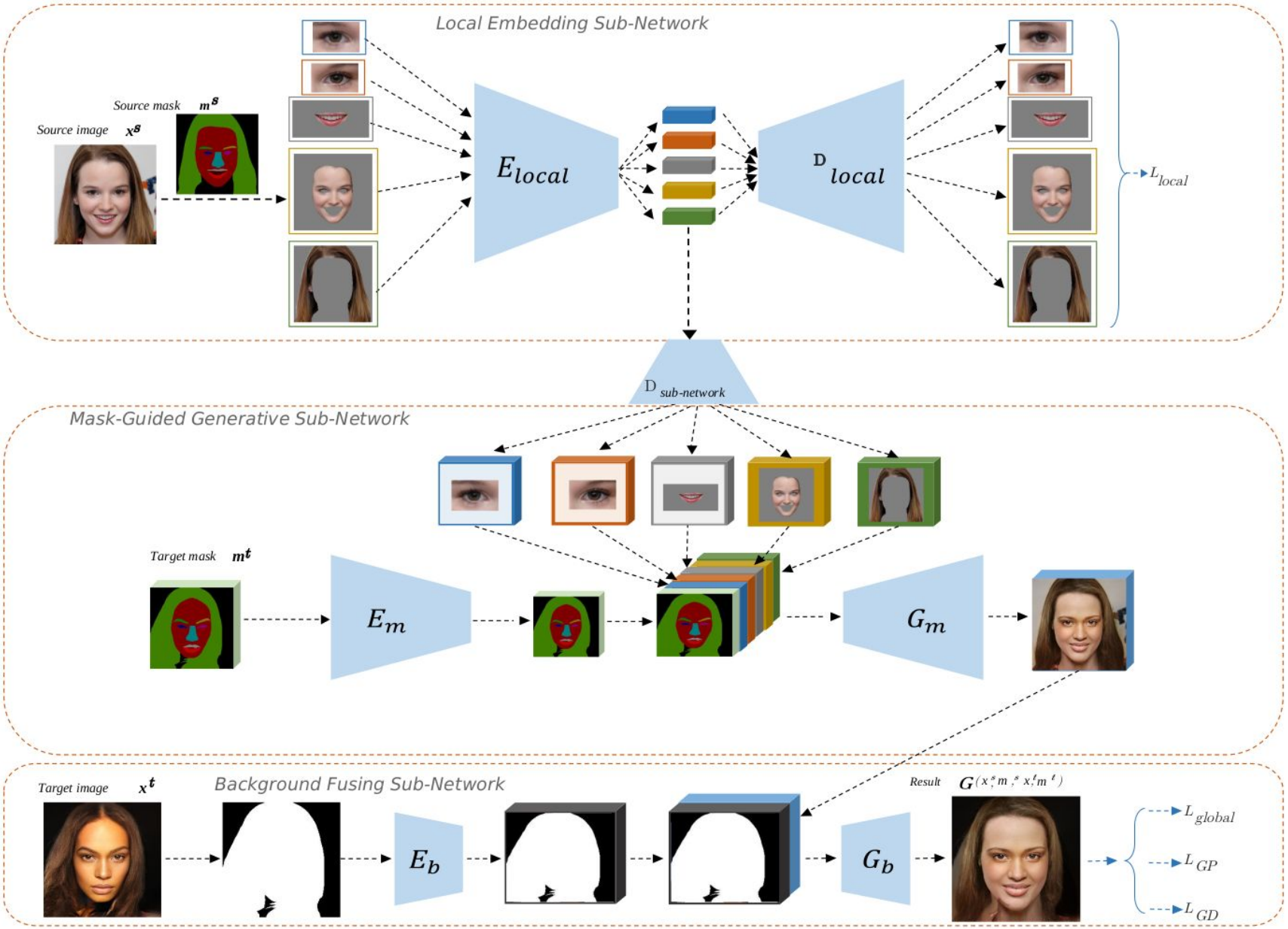

Figure 5: Project Architecture: Consists of three parts viz., local embedding subnetwork, mask guided generative subnetwork, and background f using subnetwork.

Local Embedding Sub-Network

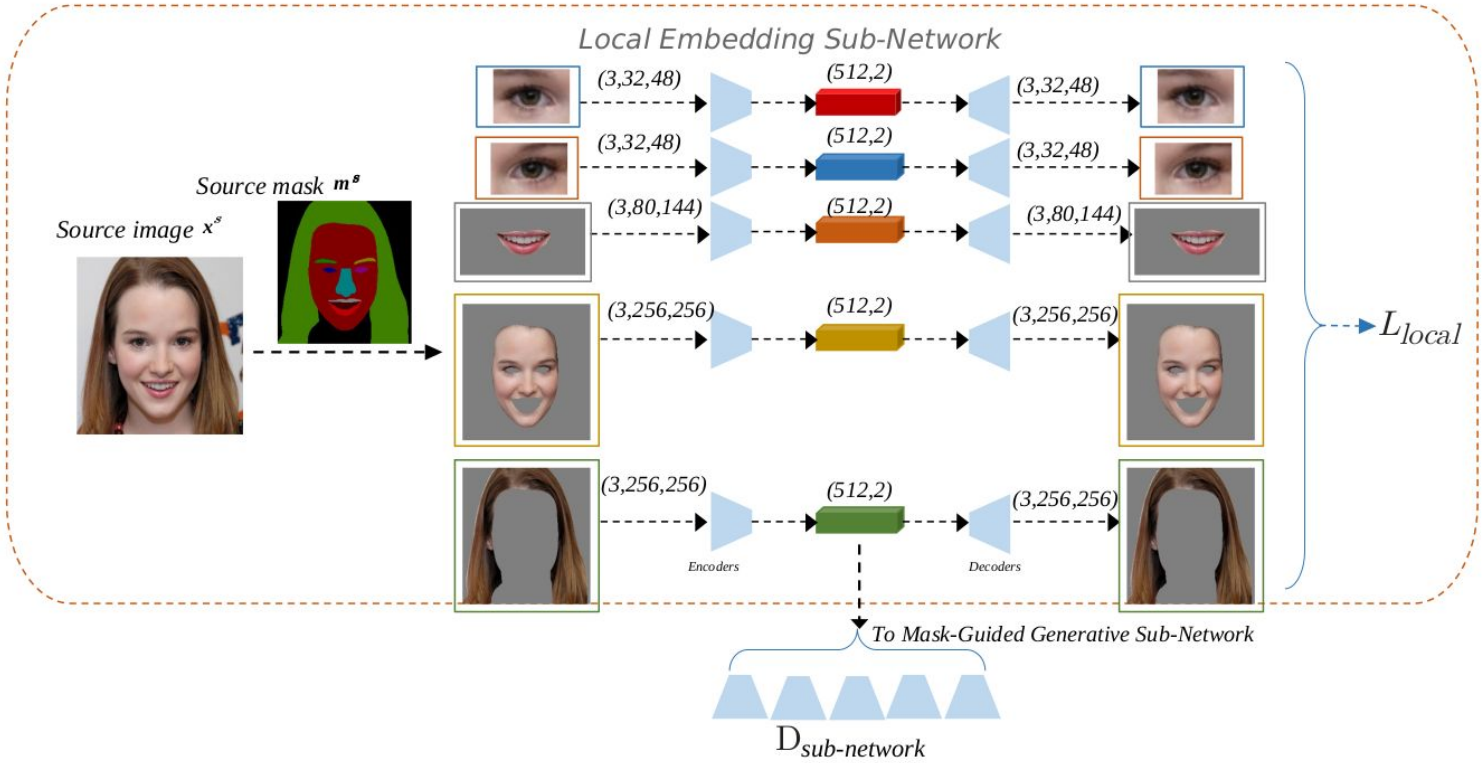

Figure 6: Architecture of the Local Embedding Sub-Network, individual components of the source image are extracted with the help of its mask and used to train an Autoencoder with two heads at its output.

Local Image Reconstruction Loss:

$$L_{local} = \frac{1}{2} \parallel x_i^s - D_{local}^i(\mathbb{E}_{local}^i(x_i^s)) \parallel_2^2$$

Equation 2: Pixel wise MSE loss for each of the facial features. x^s_i *represents "left eye", "right eye", "mouth", "skin, and "hair".*

Mask Guided Generative Sub-Network

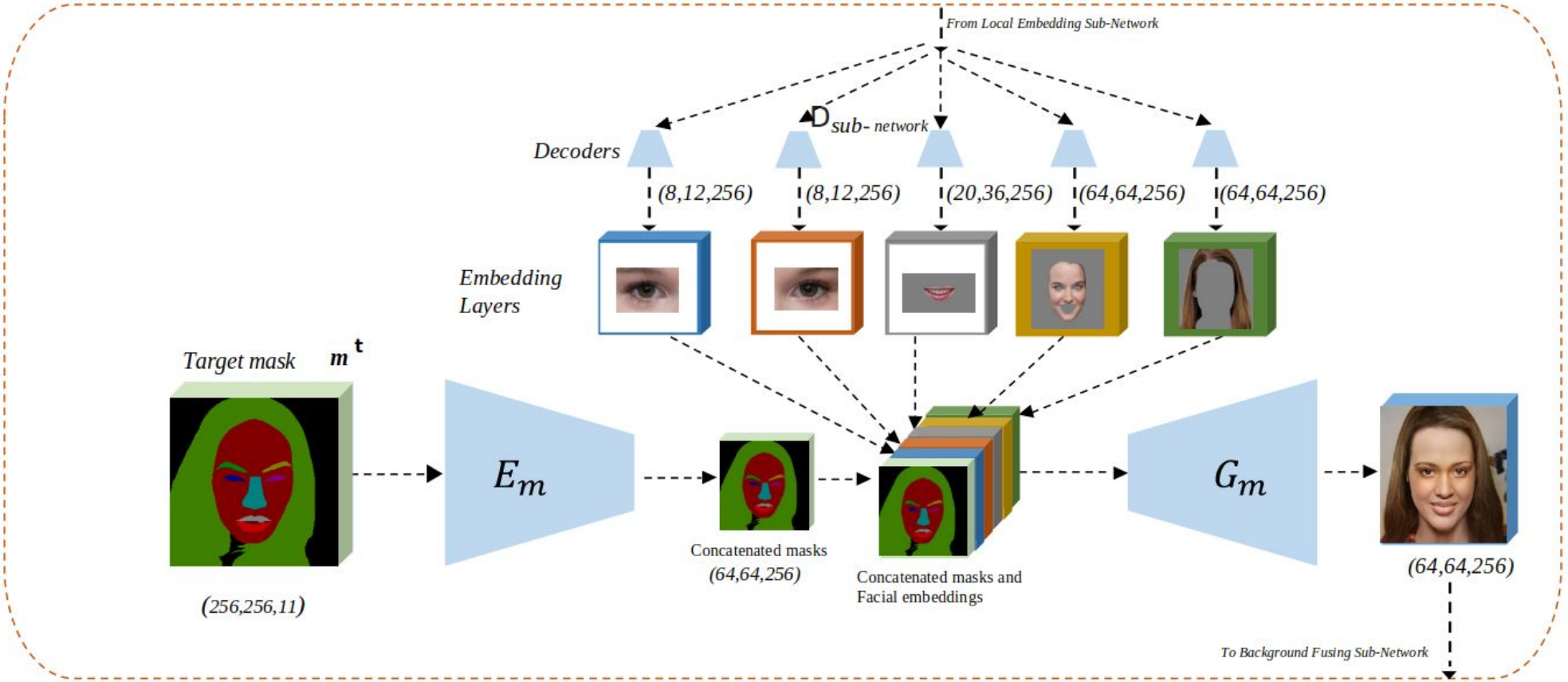

Figure 7: The structure of the Mask-Guided Generative Sub-Network, the Generator G_m learns to generate foreground images using the feature embeddings from the local embedding network and the target mask from the encoder E_m.

- The generated foreground images have the structure from the target mask m^t and the style from the local embeddings coming from the Decoder of the Local Embedding Sub-Network.

- The target mask m^t is a tensor with eleven channels, and the shape of the tensor is (256, 256, 11). Each channel of the tensor encodes one feature of the image in the form of the labels

Background Fusing Sub-Network

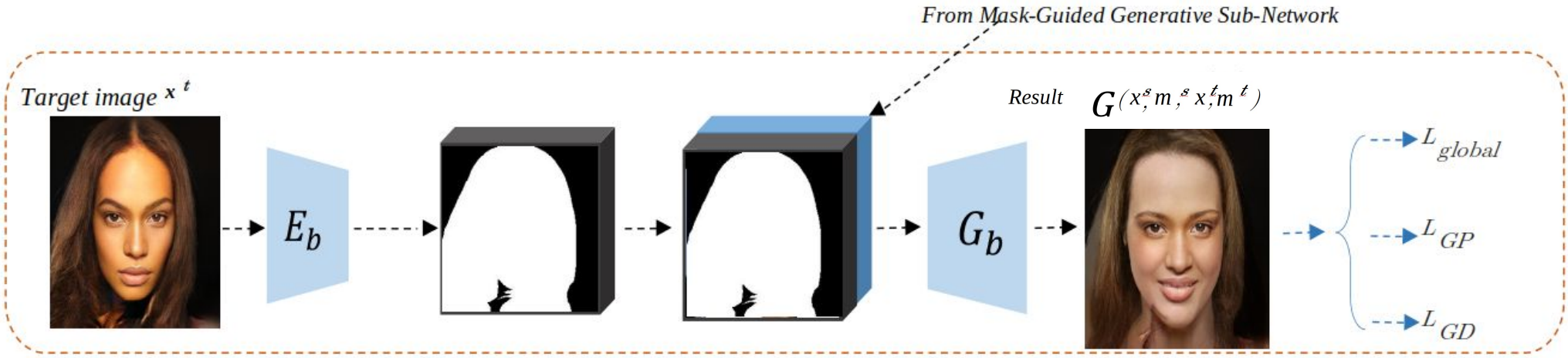

Figure 8: The output of the Generator G_b is a three channel image tensor, it is used to compute losses: Global Reconstruction Loss (L_{global}), Adversarial Loss (L_{GD}), and Face Parsing Loss (L_{GP}).

Global Reconstruction Loss:

$$L_{global} = \frac{1}{2} \| G(x^s, m^s, x^t, m^t) - x^s) \|_2^2$$

Equation 3: Pixel wise MSE loss, During training, when the input source image x^s is the same as the input target image x^t, the resultant image generated by the generator G_b which is a function of all four inputs $G(x^s, m^s, x^t, m^t)$ (section 3), must be the same as x^s .

Loss Functions

Adversarial Loss:

- The generator (G_b) uses a **Multiscale discriminator** [11,(Pix2Pix)], two identical discriminators are used.
- The first discriminator is fed with the full scale images.
- For the other discriminator, both the generated and real images are scaled down by a factor of two using an average pooling layer.

$$L_{sigmoid} = -\mathbb{E}_{x^t \sim p_r}[\log (D_i(G(x^s, m^s, x^t, m^t), m^t))]$$

Equation 4: Loss function of the framework

Pairwise Feature Matching Loss (L_{FM}) [12]:

- The features of the discriminator network D are matched pairwise for real and generated images in order to overcome the unstable gradient problem of the GAN [2].

$$L_{FM} = \frac{1}{2} \parallel f_{D_i}(G(x^s, m^s, x^t, m^t), m^t) - f_{D_i}(x^t, m^t) \parallel_2^2$$

Equation 5: L_{FM} stands for pairwise feature matching loss, and it is defined as the Euclidean distance between feature representations for generated images and the target image.

Loss Functions

Face Parsing Loss:

- The fully trained face parsing network (U-Net) is utilized to stimulate the produced samples of the framework to have the same mask as the target mask

$$L_{GP} = -\mathbb{E}_{x\sim p_r}[\sum_{i,j} \log P(m^t_{i,j}|P_F(G(x^s, m^s, x^t, m^t))_{i,j})]$$

Equation 6: $m^t_{i,j}$ is the input mask label for x^t located at (i, j) and $P_F(G(x^s, m^s, x^t, m^t))_{i,j}$ is the predicted pixels of the face parsing network at the location (i, j).

Overall Loss Functions:

$$L_G = \lambda_{local} \cdot L_{local} + \lambda_{global} \cdot L_{global} + \lambda_{GD} \cdot L_{GD} + \lambda_{GP} \cdot L_{GP}$$

Equation 7: where, λ_{local} , λ_{global} , λ_{GD} , and λ_{GP} are the parameters to balance the importance of different losses.

Training Strategy:

- **Paired data**, source and target images are **Same**.
- **Unpaired data**, source and target images are **Different**.

$$L_G = \lambda_{local} \cdot L_{local} + L_{sigmoid} + \lambda_{GP} \cdot L_{GP}$$

Equation 8: Renewed loss functions for paired data

Results-Original Method

Mask-to-face Synthesis:

- The model is able to generate diverse images from a fixed target mask.
- The generated images have similar facial features, specially the skin with different style from the source image.
- We can also add glasses on the face, as seen on the last row of the Figure: .

Figure 10: The framework synthesizes realistic, diverse and mask equivariant faces from one target mask. In the above figure, the first and scond columns of each row are the target image and target mask, respectively. All the remaining images are the generated images by the framework with random source faces.

Results-Original Method

Face Component Editing:

- By leveraging the fact that the generated image of the framework depends on the mask of the target image for facial components, we can control the facial components of the generated image.

Figure 11: In the first column, the expression of the source image is changed from smile to giggle and in the second column it is changed from grin to stern face by changing the mouth component of the target image mask. The generated image of the last column shows the ability of the framework to remove bangs from the face.

Figure 12: The framework is capable of generating new images by replacing the target facial mask of the source image completely, with good amount of control. In the above figure, first row images are the input and second row images are
the images generated by the framework. The generated image has changed face, eyes, and hairline when compared to the source image.

Results-Original Method

Style Transfer:

- Mask of the target image was fixed, but the same target image was used as the source image such that the local feature embedding of one facial feature was copied from a different source image

Figure 13: The figure demonstrates the ability of the model to transfer the style of an image without copying the facial features. In the first row, the eye color of the source image changed, in the second row, the model is able to generate beard from the input image which do not have any beard. The third row demonstrates the ability of the image to change hair color and in the final row the model is able to change the color of the lips of the source images, here the first and the third images are the source image, and the second and fourth images are the outputs respectively.

Experiments

Changing the Eye mask:

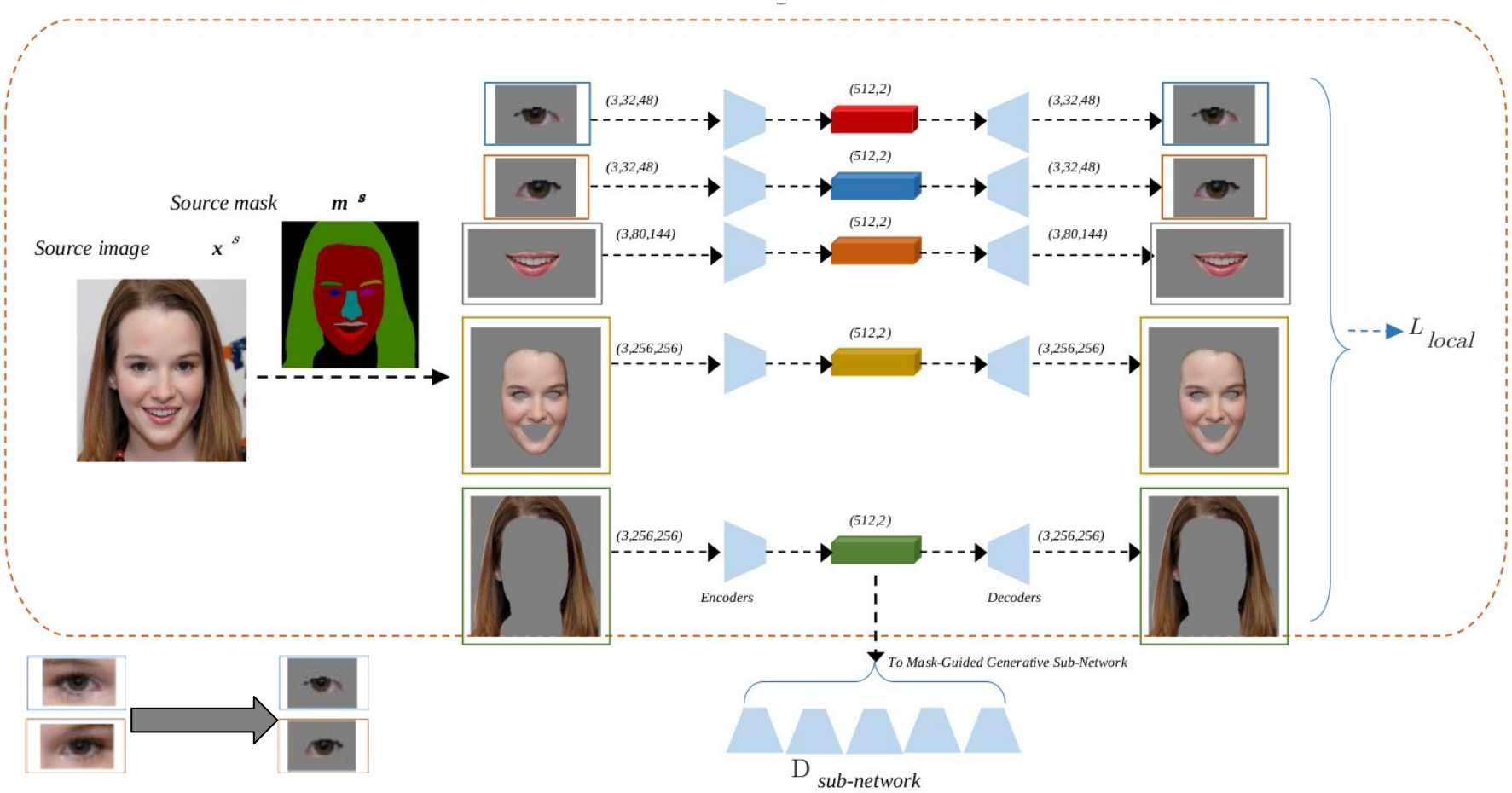

Figure 14: The cropped image for the eyes from the source image x^s is free from the skin encoding, it is fed to the autoencoder which learns the style encoding of the source image.

Experiments-Results

Results of changing the Eye mask:

Figure 15: The images in the first column are the images to be edited, and the second column consists of the images whose style is to be copied onto the first image. The images of the third column is the output image of the framework having the skin style of the second column images with the style of eyes unchanged. The fourth column is the output of the new eye mask changed method, where the skin tone around the eye is maintained constant. The final column is the out generated by the first method but in order to keep the skin tone same even the style of the eye must be copied from the second column images. Since only the eyes of the source image is used by the local encoder during training, the generated images are free of any sudden skin tone change.

Experiments

Multi Target Mask Encoder:

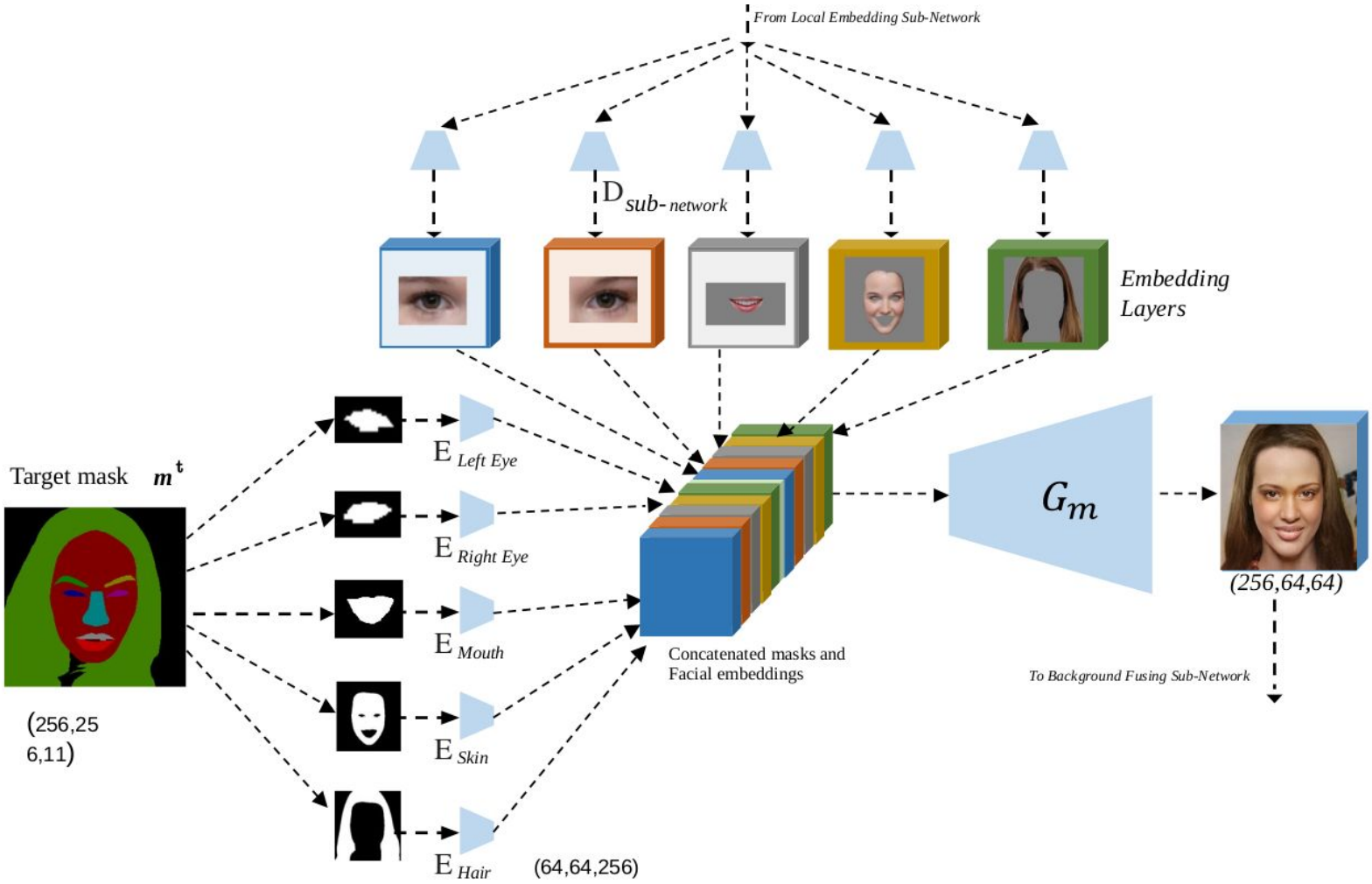

Figure 17: Multi Target Mask Encoder: "$E_{\text{Left Eye}}$", "$E_{\text{Right Eye}}$","E_{Mouth}", "E_{Skin}", and "E_{Hair}" are the five local encoders. The shape of the output of each of the local encoders is (64,64,256)

Dataset

Dataset:

- CelebAMask-HQ
- 30K HD images of Celebrities.
- Shape: (1024,1024,3).
- 40 attribute annotations.
- (512,512) Image Masks with 19 Classes.

Figure 9: Sample Image and mask from the training dataset

Fréchet Inception Distance Score(FID):

- Metric to evaluate the performance of GAN network.
- Measuring the distance between feature vectors(mean and variance) calculated for real and generated images.
- It uses Inception V3 model as a feature extractor with 2048 activations from the second-to-last layer of the Inception model.
- Tested on 2000 test images from the CelebAMask-HQ dataset.

Experiment Method	FID Scores
Original Method	26.54
Changing the Eye mask	26.04
Multi Target Mask Encoder	28.48

Table 1: FID scores for different Experiments.

Conclusion

- The framework is capable of explicitly editing facial components, style transfer from one image to another, generate new faces using a fixed target mask.
- The FID score is used to evaluate the quality of the generated images.
- The framework is capable of generating realistic images by keeping the target mask fixed and changing the source image and vice versa.
- The framework is also capable of adding glasses to the face, change eyebrow, eyeball color, add and remove bangs, change emotions on the face, increase and reduce hair, grow and remove beard from the face, and change lip color.
- A series of experiments were conducted to understand the capabilities and limitation of this technique.
  - Increasing the learning rate helped the framework in learning facial embeddings faster, but the style transferability of the output image was not good
  - Changing the normalization from batch to instance normalization for the encoder network and adding dropout to the CNN layers was not very helpful in generating good images
  - The target mask encoder was split into individual masks and the model was trained with four additional encoders, the output images with this method were as good
  - Changing the window size for the eye using the eye mask improved the controllability of the generated images when compared with generated images.

Future Scope

- The fine details in hair, shadows due to lighting, high resolution style image generation can improve further.

- This method needs a reference image to edit the features, unlike VAE where we have control over the latent space and change things like age of the subject.

- Training time for such a model is huge, more work can be done in this field.

Thank You

Questions?

- [1] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2414–2423. IEEE, 2016.
- [2] M. He, J. Liao, L. Yuan, and P. V. Sander. Neural color transfer between images. arXiv preprint arXiv:1710.00756, 2017.
- [3] H. Chang, J. Lu, F. Yu, and A. Finkelstein. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
- [4] B. Dolhansky and C. Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7902–7911, 2018.
- [5] Z. Shu, E. Shechtman, D. Samaras, and S. Hadap. Eye opener: Editing eyes in the wild. ACM Transactions on Graphics (TOG), 36(1):1, 2017.
- [6] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In Proceedings of the European Conference on Computer Vision (ECCV), pages 818–833, 2018.
- [7] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K.Nayar. Face swapping: automatically replacing faces in photographs. In ACM Transactions on Graphics (TOG), volume 27, page 39. ACM, 2008.
- [8] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast faceswap using convolutional neural networks. In The IEEE International Conference on Computer Vision, pages 3697–3705, 2017
- [9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. arXiv preprint arXiv:1711.11585, 2017.
- [10] Ref: M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014
- [11] O. Ronneberger, P. Fischer and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
- [12] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In CVPR, volume 3, page 7, 2017
- [13] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen and L. Yuan, “Mask-guided portrait editing with conditional gans,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3436–3445.