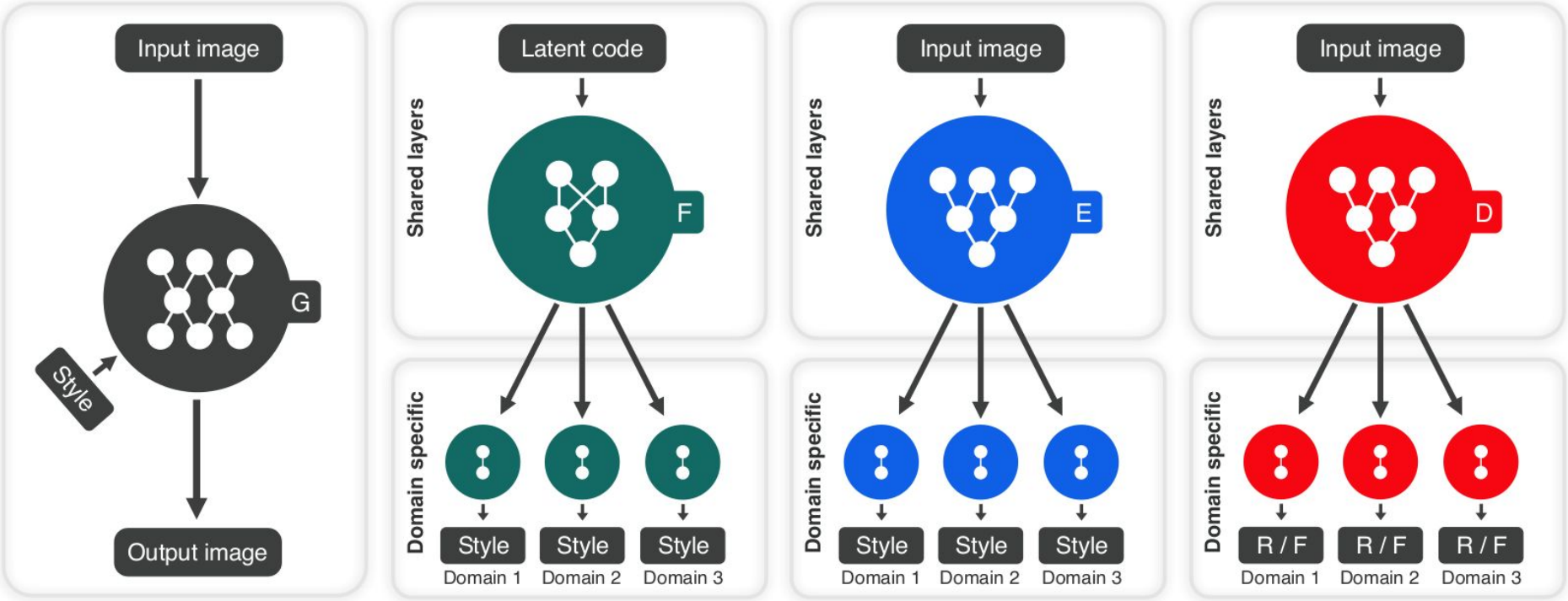
Text Guided Image Generation

Fakhriddin Tojiboev

Yandex School of Data Analysis

StarGAN v2



(a) Generator (b) Mapping network (c) Style encoder (d) Discriminator

Yunjey Choi et al. StarGAN v2: Diverse Image Synthesis for Multiple Domains

Proposal

StarGAN v2 controls the attributes of the input image using the reference image. Moreover it has the mapping network that converts random noise to style vector to change the attributes of the input image.

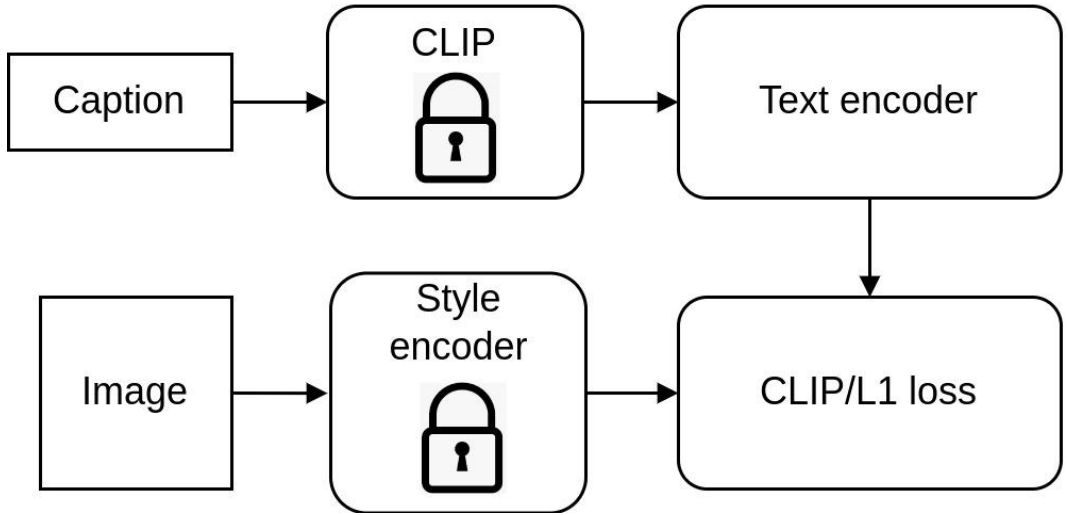
Aim:
Add a text encoder module to enable StarGAN v2 to change the attributes of an input image using text.

The list of tasks:
1) Build a text encoder that extracts style vector from text
2) Train the text encoder in the similar manner as the mapping network

Training text encoder

Text encoder architecture

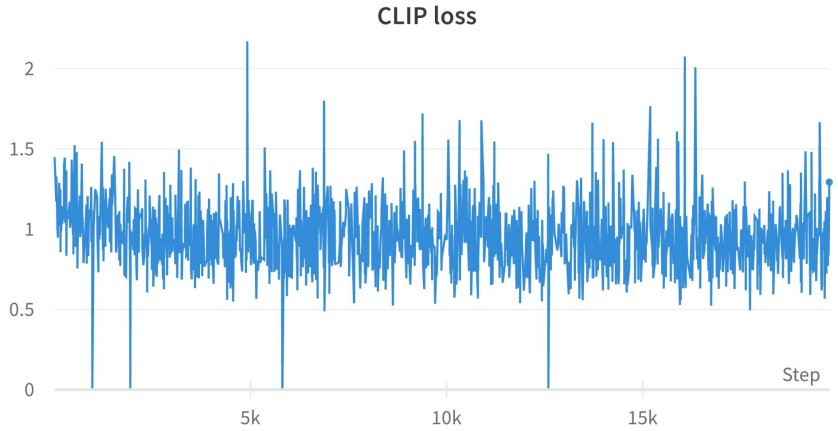
Transformer encoder (d_model=512, n_heads=8, num_layers=1)
ReLU
Linear(512, 512)
ReLU
Linear(512, 256)
ReLU
Linear(256, 128)



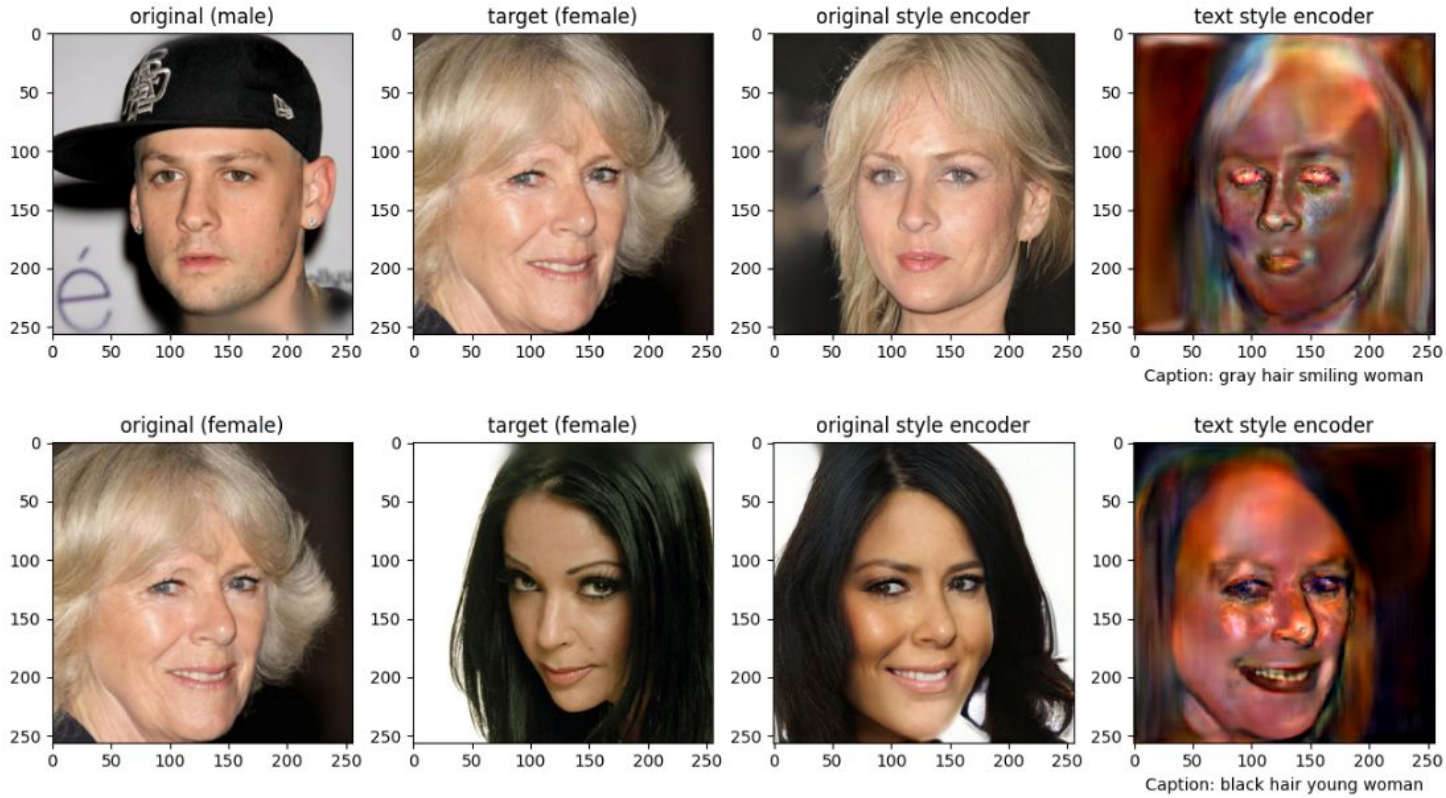
Alec Radford et al. Learning Transferable Visual Models From Natural Language Supervision

Results with CLIP loss

$$\mathcal{L}_{\text{CLIP}} = \mathbb{E}_{(\mathbf{x},\mathbf{c}),y}\left[-\log \frac{\exp\left(T(\mathbf{c}_0)^{\mathrm{T}} E_y(\mathbf{x}_0)\right)}{\sum_n \exp\left(T(\mathbf{c}_0)^{\mathrm{T}} E_y(\mathbf{x}_n)\right)}\right]$$

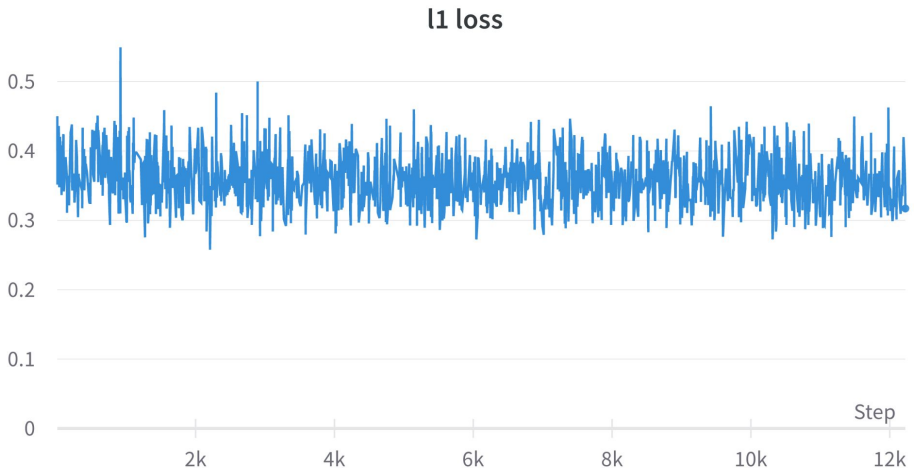


Results with CLIP loss



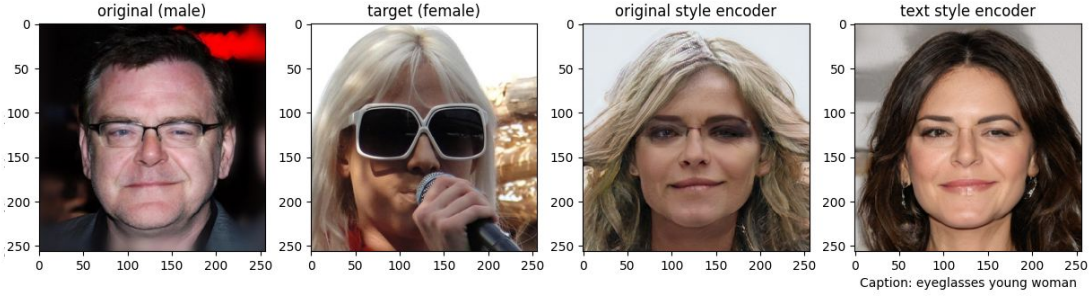
Results with L1 loss

$$\mathcal{L}_{\mathrm{L}_1} = \mathbb{E}_{(\mathbf{x},\mathbf{c}),y} \left\| T(\mathbf{c}) - E_y(\mathbf{x}) \right\|_1$$

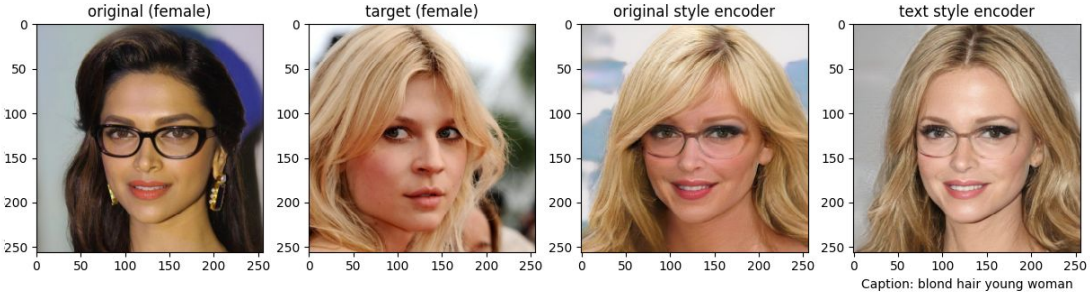


Results with L1 loss

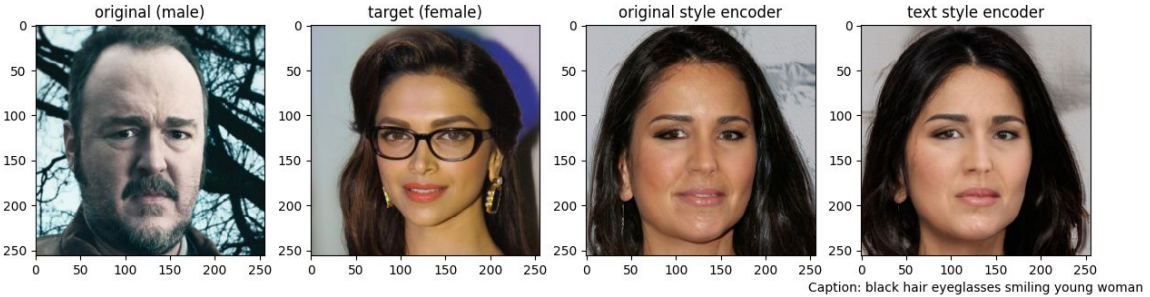
Both model can't put glasses



Both model can change hair color

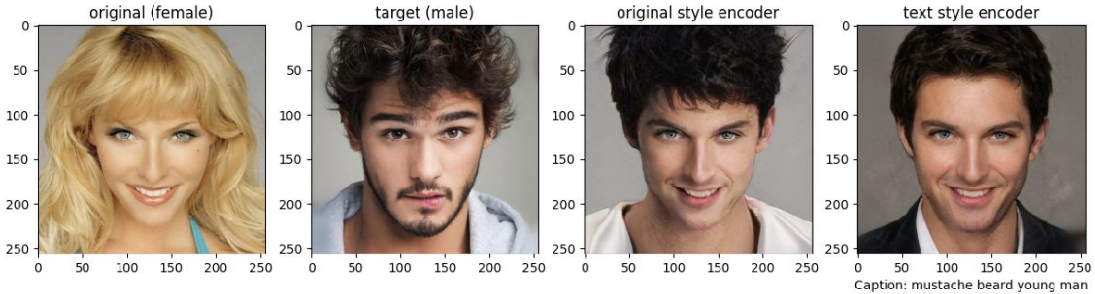


Both model can't put glasses

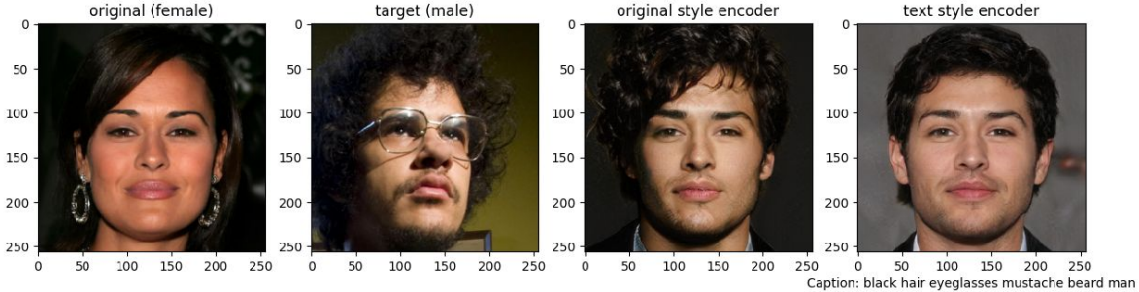


Results with L1 loss

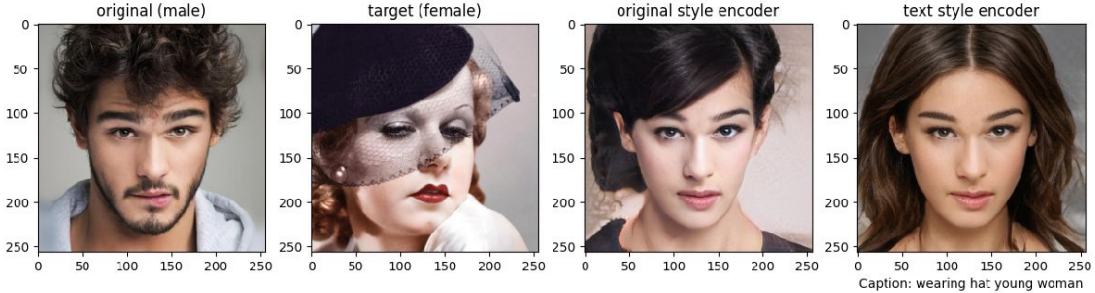
Both model can add beard



Both model can add beard but can't add glasses.



Both model can't put hat

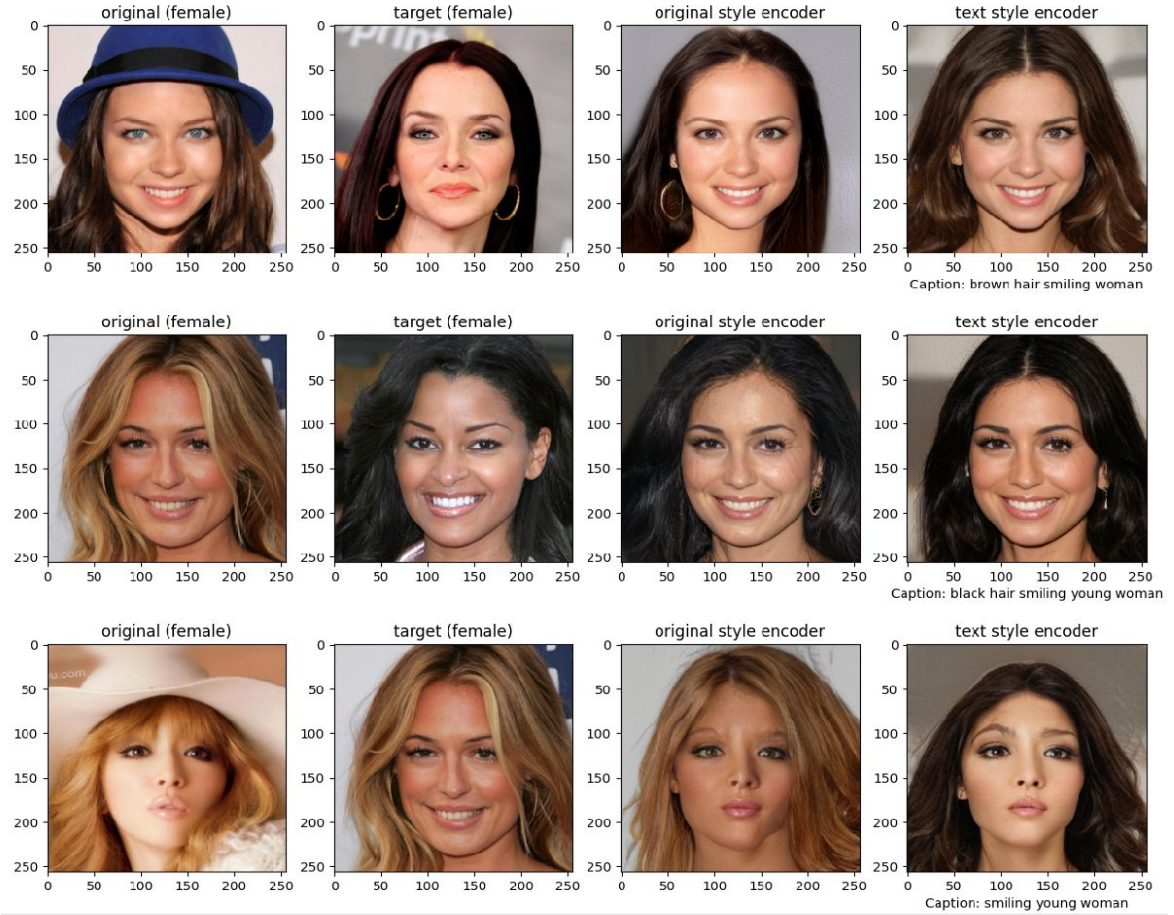


Results with L1 loss

Both model can change hair color

Both model can change hair color

Both model can’t make smiling woman



Conclusions

We see text encoder can change the hair color, gender, can add beard. If we don't specify the hair color then text encoder automatically changes it to black or brown by default. Text encoder fails when we want to put glasses or hat. However in these cases style encoder also fails. One of the possible solution is to train StarGAN v2 with balanced attributes from scratch and after that we might be able to change more attributes and it might solve the problem of text encoder as well.