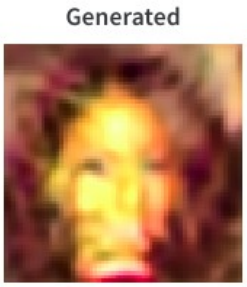
StarGAN v2 Report

First let me say a few words about the code. I implemented the networks as they are described in paper in the most simple way. In particular, I did not implement high pass. Instead, I trained both a U-net like architecture, and one without any skip connections between the encoder and decoder part of the generator. It turned out, that U-net like architecture, gives much better results visually. Secondly, as I was using 64x64 images, the networks had to be downscaled. I used the downscaling that was implicitly suggested by the code, making the nets shorter and wider. Next, when I was implementing the objectives, I made a mistake in the adversarial loss for the discriminator. The term including the fake images had the wrong sign. So the net was learning something (while debugging, I tried a lot of different tricks – such as gradient clipping, removing the diversity loss, nothing helped of course, although the learning curves did become more smooth), but it was very strange, here are some examples:

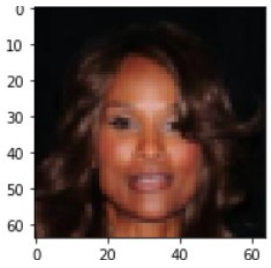
male

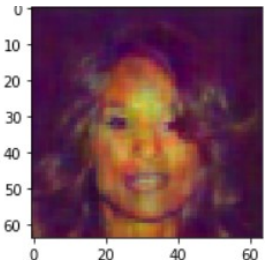


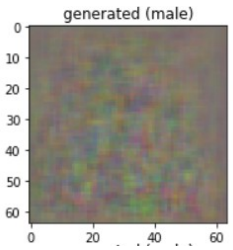
female

In order to track progress, I displayed each component of the loss, and also displayed images from the dataset along with their generated versions (the style was sampled using the mapping network). Because training took several days, it was very handy to use wandb to display the losses and images. The reports from wandb are also available.

Now I am going to describe some observations for the learning process of U-net and no skip-connection architectures. First of all, U-net started from basically identity transformations, while the other from noise.





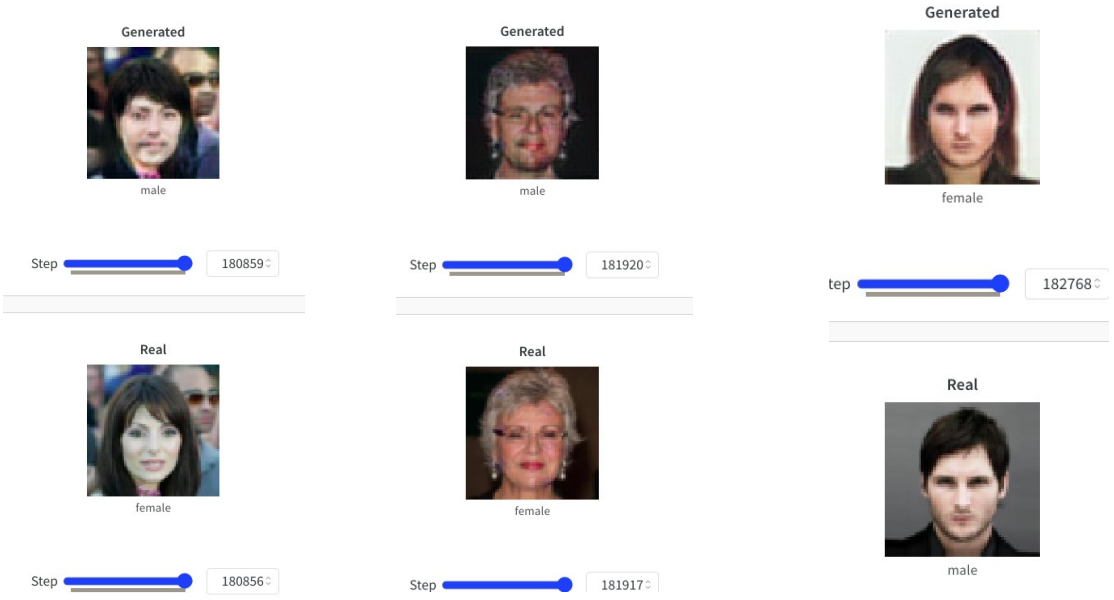


Next, for Unet, grid-like artifacts began to appear at the same time as the model was learning to change the domain and style.

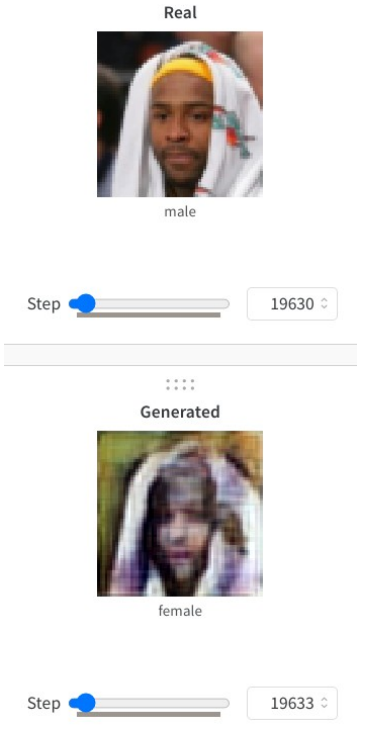
female

male

The model was slowly learning to change the source, without distorting it much. In particular, at this stage it wouldn't take of glasses or jewelry.



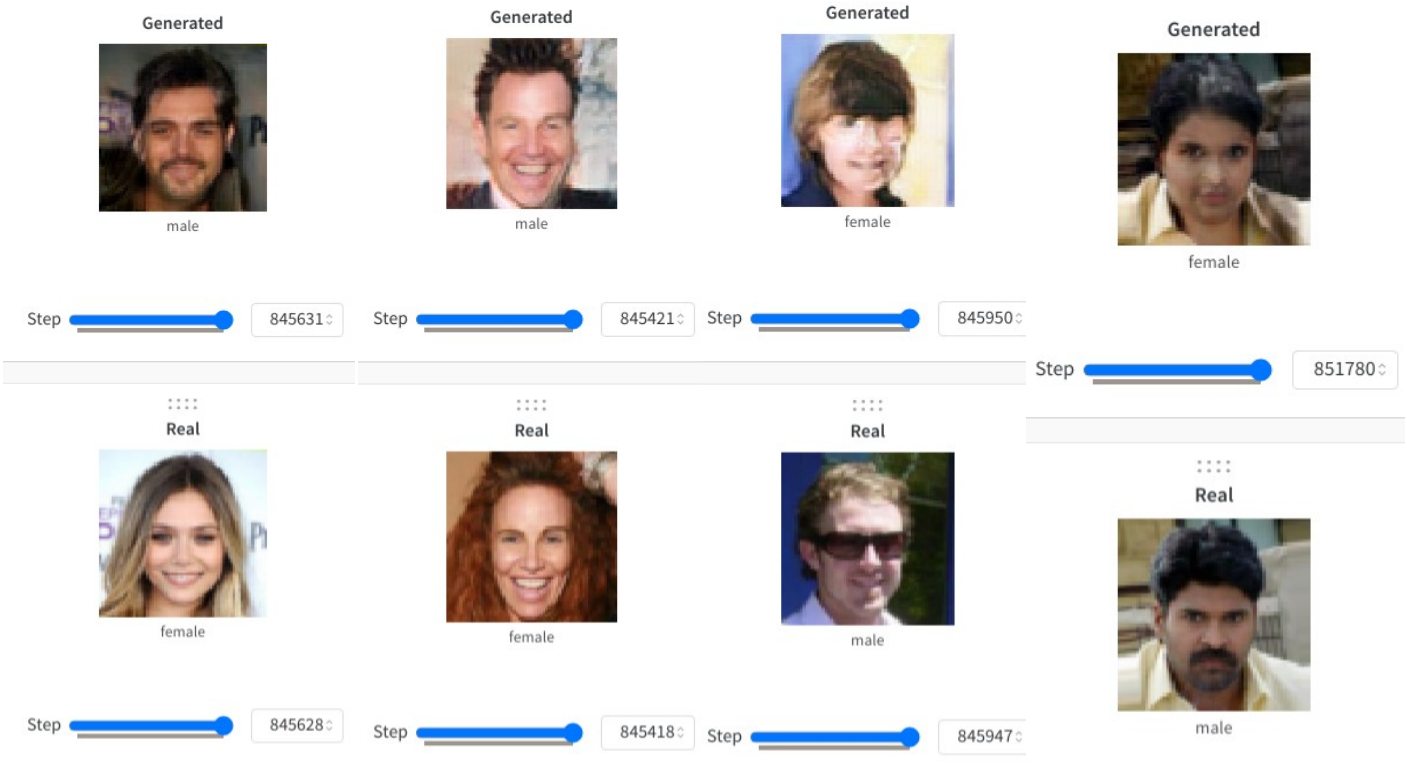
In the meantime, the other architecture, was just learning to draw faces.



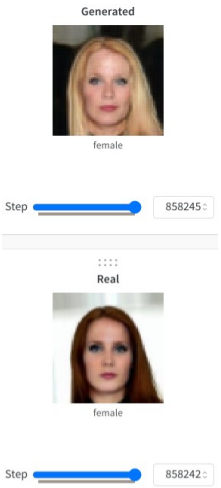
Anyway, both nets were learning. Not much to say about their losses apart from the fact that they decrease (of course, diversity loss increases, because that is our objective). Unet losses on the left, and no unet on the right. Also, I separated the losses that were calculated using reference images for style from the losses with synthesized styles. Not much difference. Perhaps the biggest difference is in adversarial loss between two architectures. Of course, since at the beginning unet is almost identity, discriminator has no way of telling it apart from real, so adversarial loss is almost zero for

unet at the beginning. For the no unet, it is more constant throughout, meaning that generator and discriminator have roughly the same intelligence.

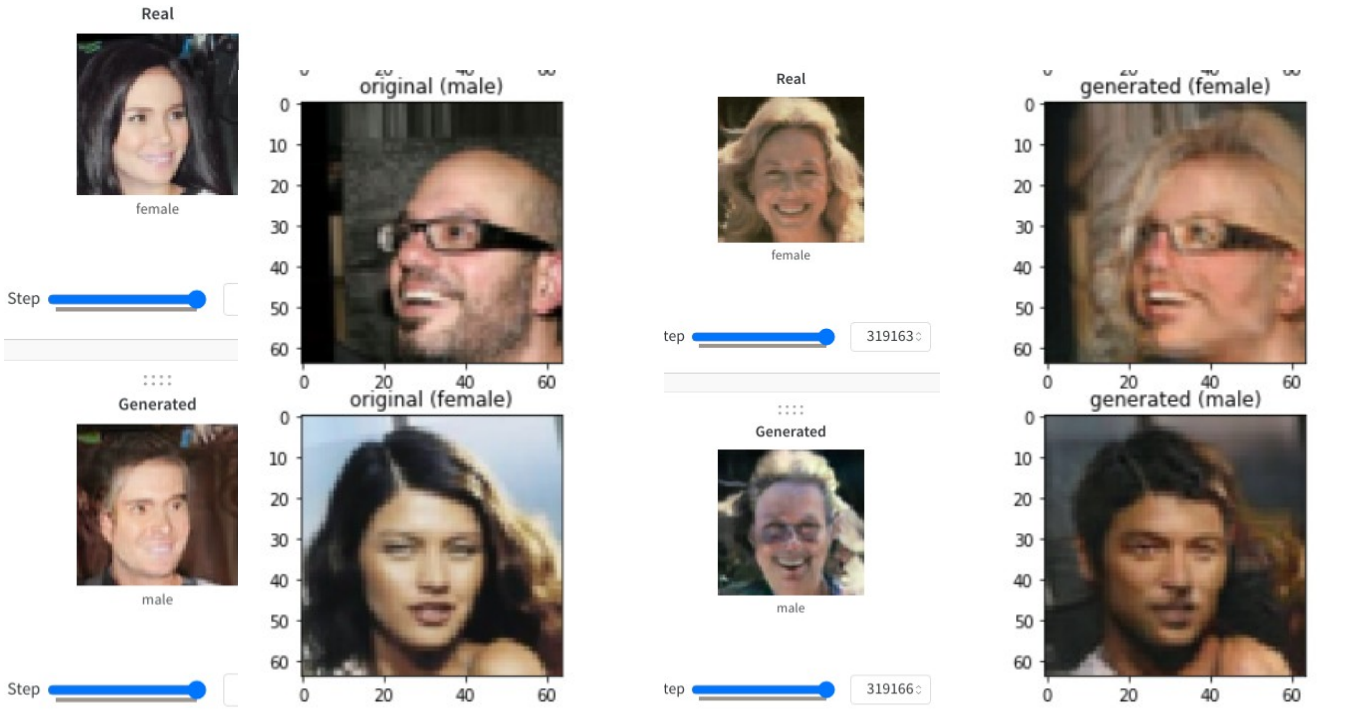
Enough said about training, let's come to the evaluation phase. First of all, some pictures generated by unet. The ability to remove hair and draw ears really fascinates me. It has also learned to some extent to remove glasses (not perfect, but still, pretty impressive I think). Also, when turning man into woman, unet primarily removes facial hair and covers ears, adding makeup sometimes. I think it's a little lazy on drawing long hair like shown in the original paper. Note how in the third example, the net removed glasses (I will explore on that ability below: it's a little tricky once the net has trained longer, it works better at earlier stages).



A really fascinating aspect I have found, is that it improves the image visually: lighting, colour etc. And also it can color black and white images (I will show these abilities by manually feeding photos – just a little difficult to capture them in train time, besides, we want to compare the two nets)!
Here is an example of unet adding light (while also changing hair color):

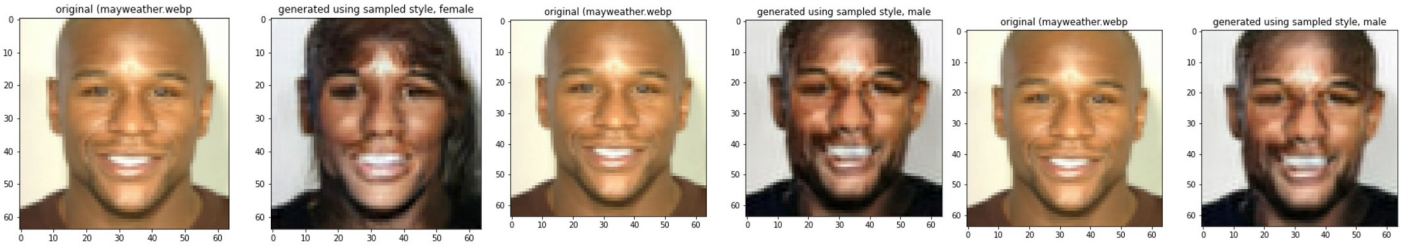


For now, let's see the average results with nounet. We can see that it can also do the things unet does, though personally, I think it distorts the images more, and produces less attractive results.

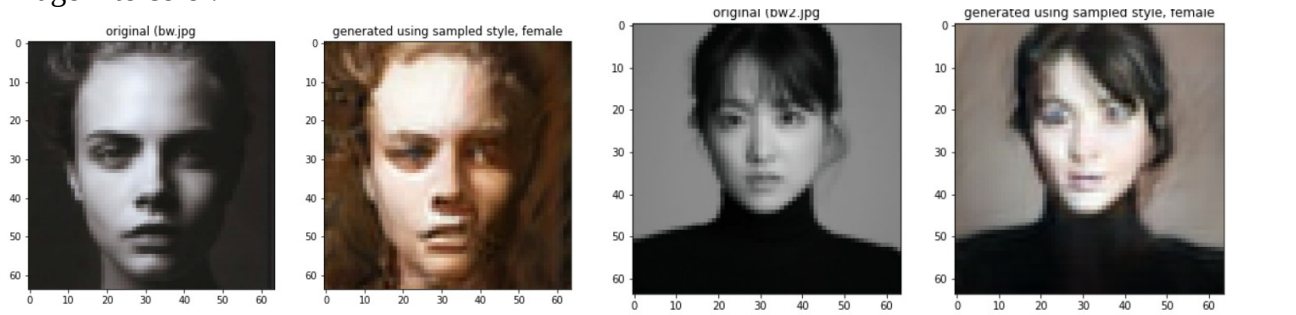


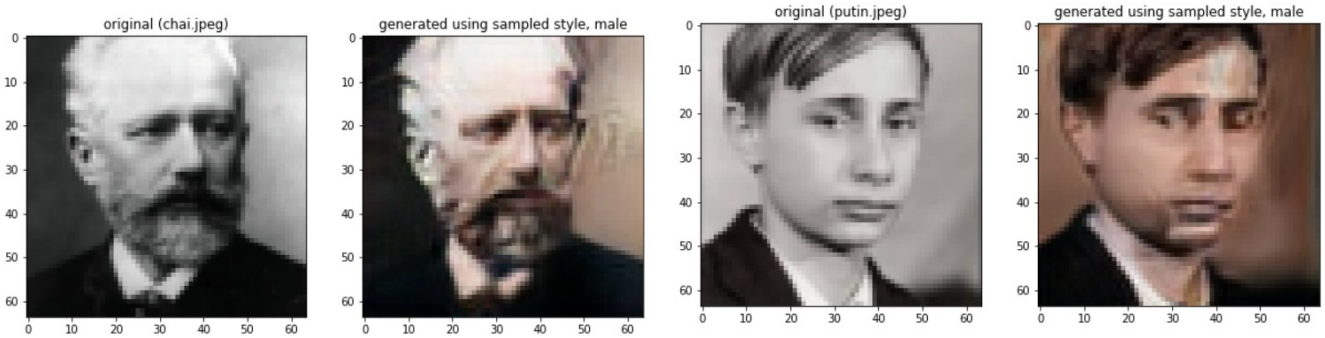
Here is an example of nounet removing shadows:

Anyway, let’s pass to the evaluation. LPIPS score for the nounet was approximately 4.9 for reference based generation and 4.4 for latent samples. I will just show some samples for nounet. First, let’s see how good it is with different races.

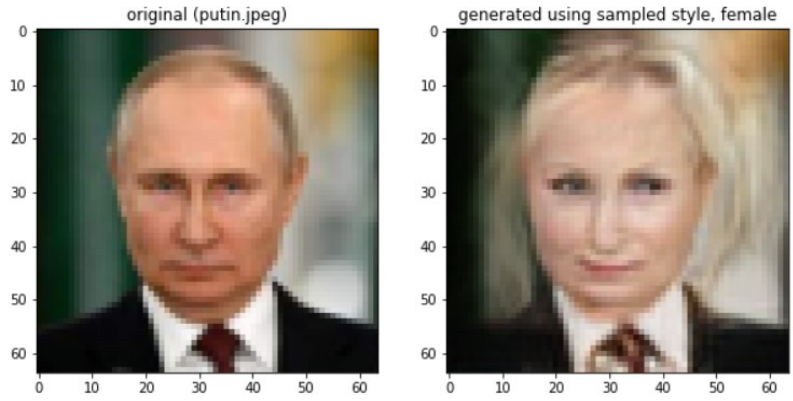


It doesn’t really wanna change race, and also, the celeba dataset doesn’t seem to be balanced on different races, so it doesn’t work very well for other races. Let’s see if it turns black and white image into color.



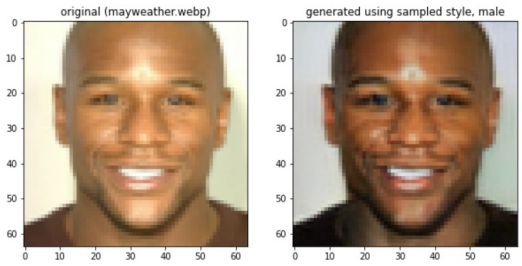


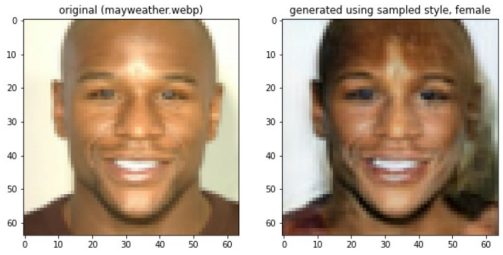
While we are at it



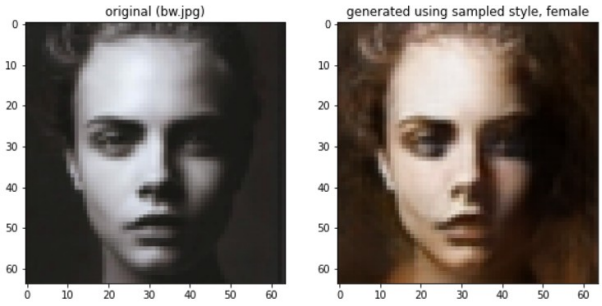
Let’s do do the YSDA-curators test.

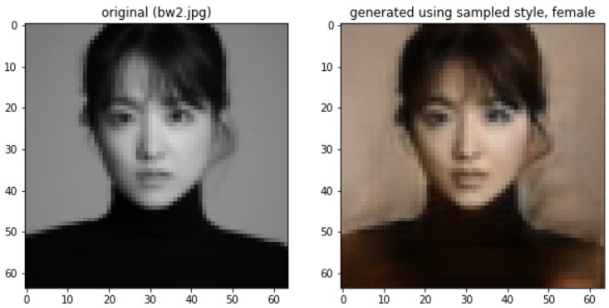
Not very successful frankly. Now let’s do similar experiments for unet. LPIPS value is approximately 2.3 (for both reference and latent). That might sound like unet is worse, but as we saw above, nounet achieves diversity by distorting images, which isn’t very good.

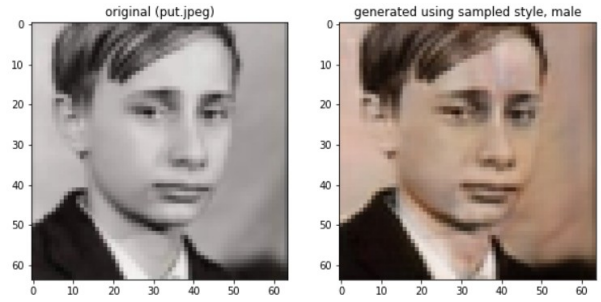


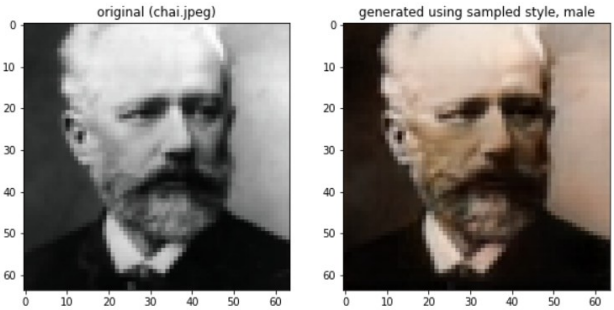


As we can see, unet can handle other races much better. Let’s see how well it can restore colour.

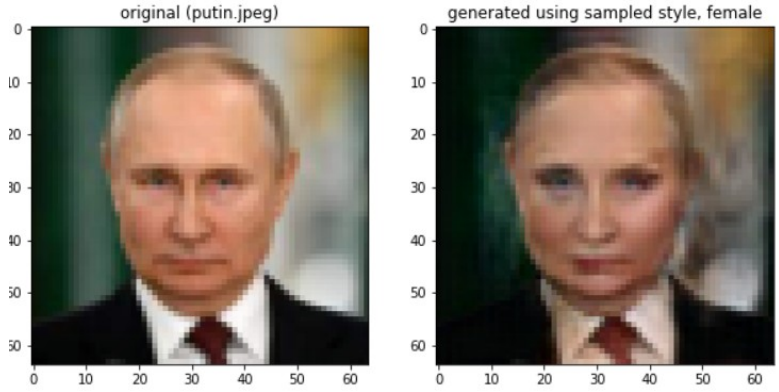




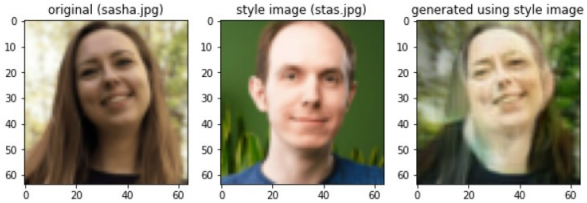


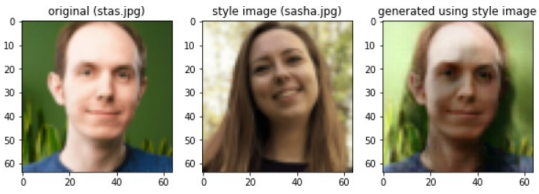


As we can see, it preserves the photo much better and produces much more coherent and visually pleasing results. Let’s keep going.

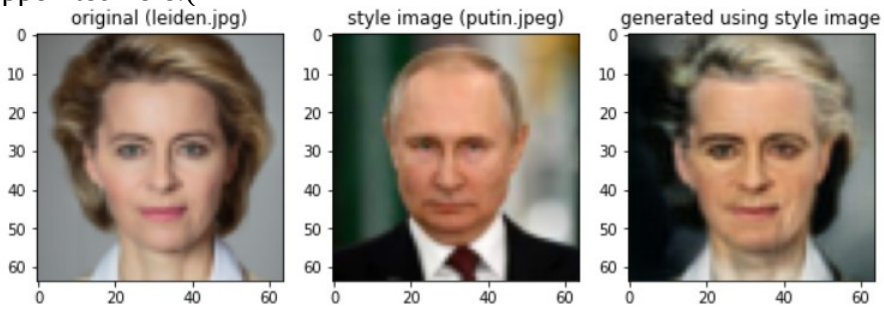


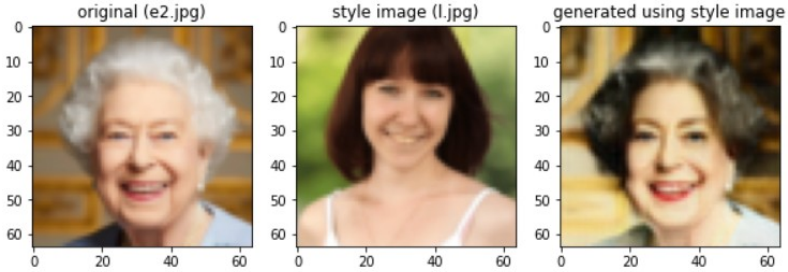
Again, we see much less deviation from the original photo, while feminine characteristics are somehow added, but very subtly. Now the curators test.



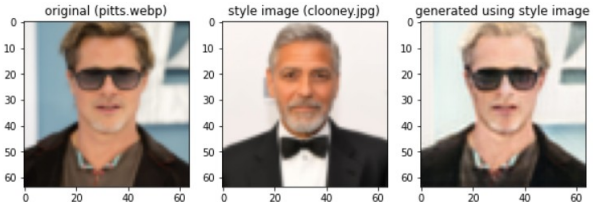


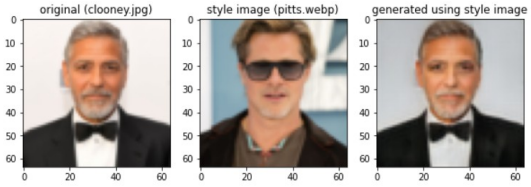
To be honest, I was a bit disappointed here:(
It does sort of work sometimes. I think it can capture age quite well.



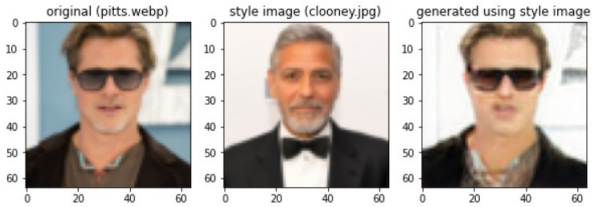


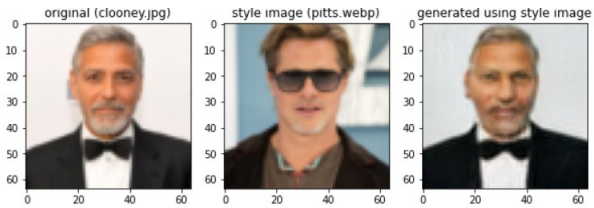
Also, let’s see how well our nets can add/remove glasses.



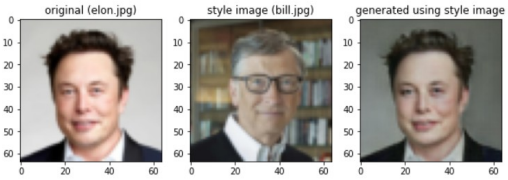


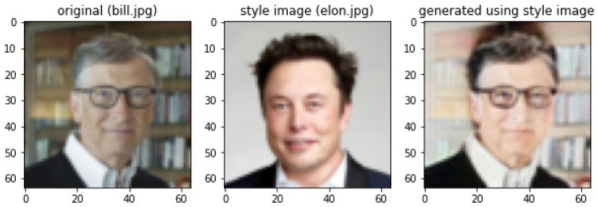
Well, unet seems to be really lazy. It’s very good at subtle changes, but when it comes to removing glasses – sorry, it doesn’t want to do so much work. What about nounet?



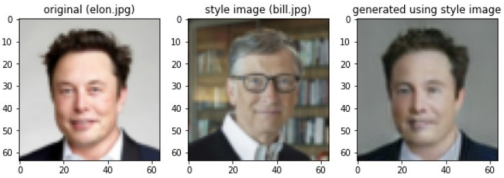


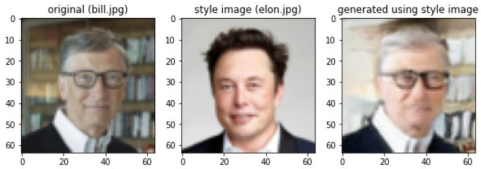
No, not with black sunglasses anyhow. Let’s try an easier task: removing transparent glasses.





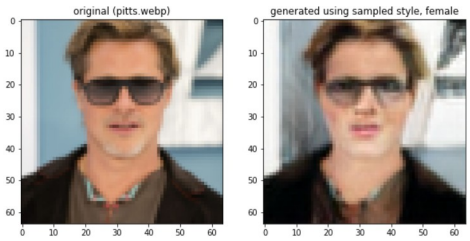
Unet above, it can’t with such thick glasses. What about nounet?

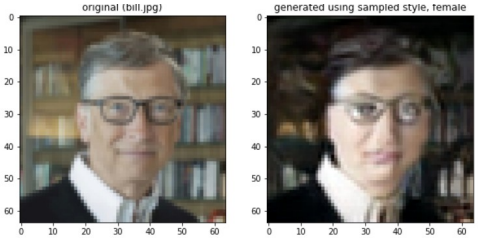




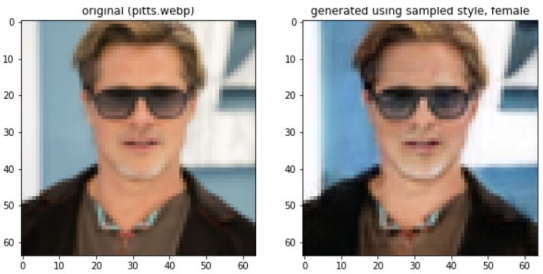
No success either, it just distorts the images. Curiously, what works, is changing the gender. For example although the glasses aren’t removed completely, they are brightened.

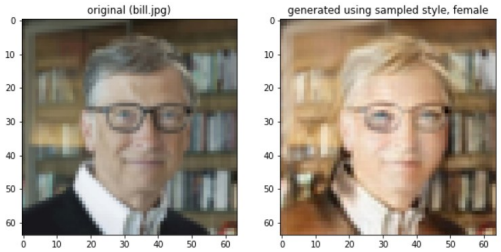
Here’s nounet, the Bill Gates glasses are “almost” removed. Or distorted if you prefer.



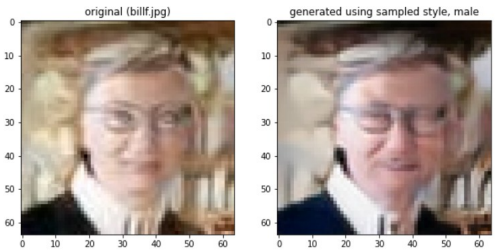


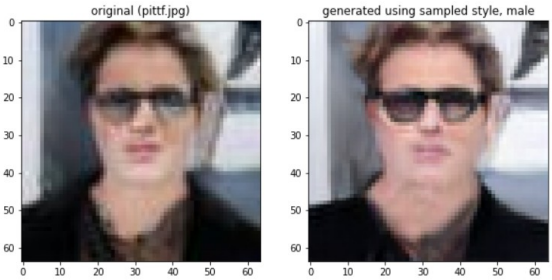
Here’s unet. As we can see, it doesn’t want to remove glasses.



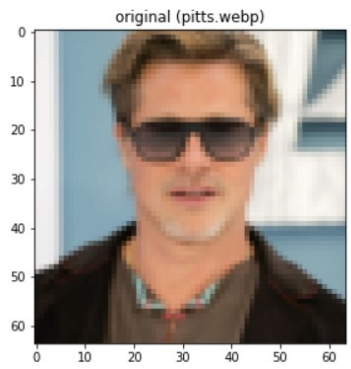


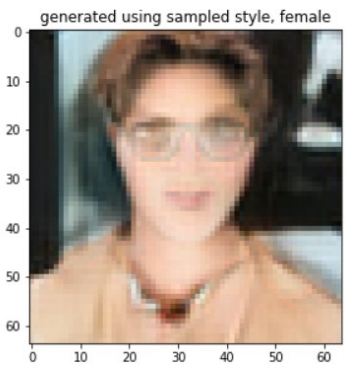
So, for nounet, let’s try this: we do several conversions male → female → male and see whether that will help.

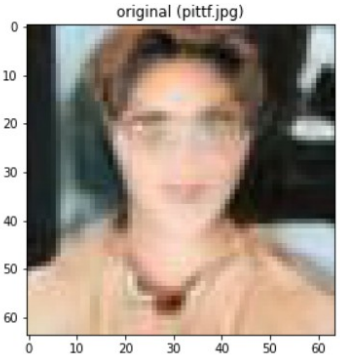


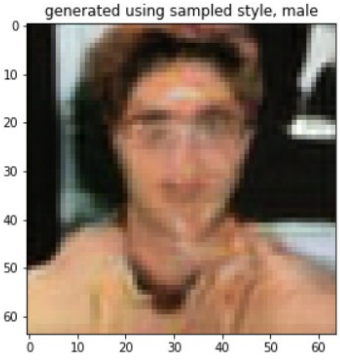


Not really. I think we can basically say, that the net has learnt to preserve glasses, although during earlier stages I saw it remove glasses. Let’s see, I’m going to load early checkpoint of unet, because I remember it did remove glasses moderately well.

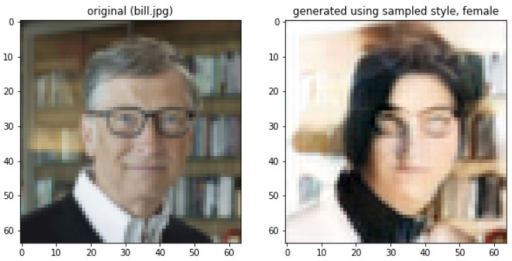


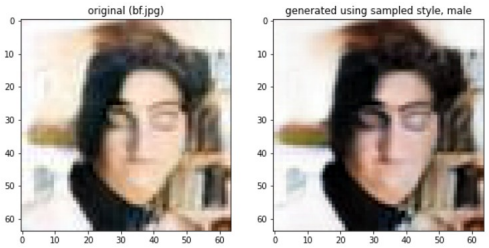






So, as we can see, after several iterations we can, indeed, remove glasses. But the image gets distorted.
Here is how Bill Gates glasses are removed (with early checkpoint, namely third epoch).





To conclude, I would like to say, that what I have noticed throughout the learning process and by comparing unet to nounet, is that there is a tradeoff between diversity of styles and intactness of images. As we have seen, nounet is more diverse and has a higher LPIPS, but it distorts the portraits more. Unet, on the other hand, is much more subtle and less diverse, has lower LPIPS, but produces more intact and realistic images.