Appendix for FRR-Net: A Real-Time Blind Face Restoration and Relighting Network

Samira Pouyanfar, Sunando Sengupta, Mahmoud Mohammadi, Ebey Abraham
Brett Bloomquist, Lukas Dauterman, Anjali Parikh, Steve Lim, and Eric Sommerlade
Microsoft
{sapouyan,susengup,mahmoha,ebeyabraham,brettbl,ludauter,anjalip,stlim,ersomme}@microsoft.com

1. Overview

In this supplementary material, we present more experimental results and analysis.

- We evaluated the importance of distortion classifier.
- We evaluated the importance of training the model with segmentation mask and dice loss.
- We evaluated the importance of different losses used to train FRR-Net.
- We evaluated the importance of different image size in order to achieve a good trade-off between accuracy and speed.
- Finally, we showed more evaluation results for screen illumination and face occlusions.

2. Importance of Classifier:

FRR-Net includes a distortion classifier which is trained in parallel with the autoencoder. In order to show the importance of this component, we experiment the model without using the classifier. Figure 1 shows the comparison between FRR-Net with classifier and no-classifier with their corresponding masks. These samples shows that FRR-Net performs better with class prior information and generates more accurate mask compared to the no-classifier model.

3. Importance of Segmentation Mask

Figure 2 shows the comparison results between mask and no-mask models during the training on the validation set. From this figure, one can conclude the importance of the mask in enhancing face restoration.

Facial Mask Results FRR-Net generates two outputs: 1) enhanced face region, and 2) facial mask segmentation. Figure 3 shows a few samples of the input and outputs of the model. The facial mask is used during training so the model only focuses on the face and discards the background. During testing, this generated mask can be used to smoothly blend the image to its original background without the need for a separate face segmentation model.

Img size	PSNR ↑	LPIPS ↓	SSIM ↑	GPU (ms)
224*224	26.74	0.25	0.84	14
336*336	29.00	0.23	0.0.87	15
448*448	29.60	0.23	0.89	23

Table 1. Results of FRR-Net on various image size (trained for 150k iterations and validated on StyleGAN data).

4. Importance of losses:

Figure 4 shows the importance of each loss on each metric during the training. We first start with only reconstruction loss L_{rec} (green line). After adding perceptual loss (L_{per}), LPIPS is significantly enhanced, however, the SSIM and PSNR are not improved. Adding Angular loss L_{ang} increases both PSNR and SSIM but slightly reduces LPIPS. Finally, adding style loss (L_{style}) provides the best trade-off between all metrics. We also tried adversarial loss (L_{adv}), but it did not provide any performance boost to these metrics and therefore is discarded in this paper.

5. Importance of image size:

In this work, we use smaller image size compared to the existing face restoration models (512 and 1024 is the popular image size used by the state-of-the-art models) to lower the computational cost and have a more efficient model. Table 1 shows the difference in accuracy and inference time for training the model for 150k iterations on various image size (224, 336, and 448). Our results shows the 336*336 input gives the best trade-off between accuracy and speed.

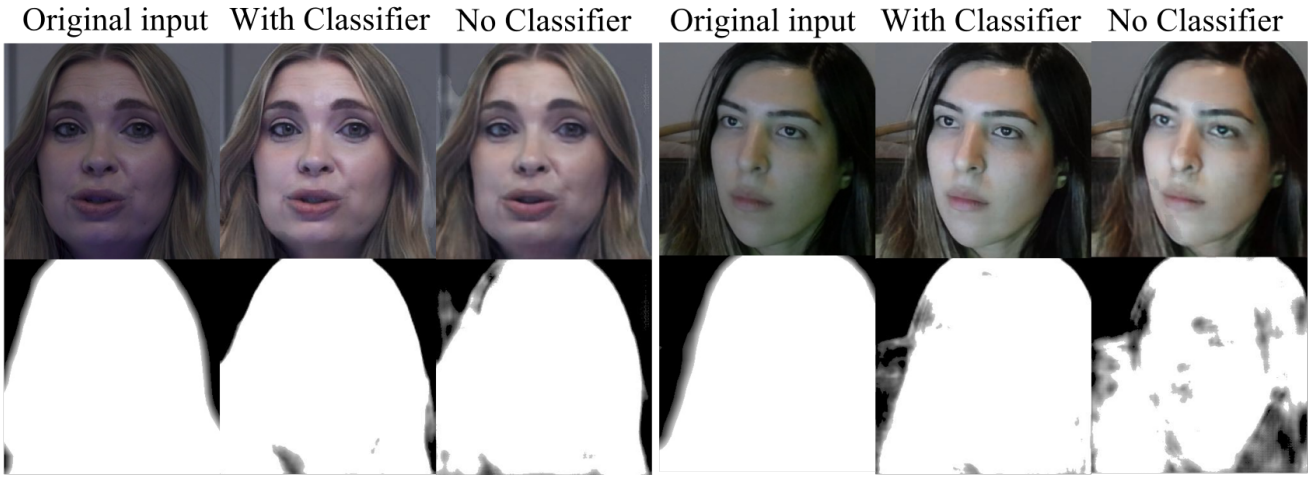

Figure 1. FRR-Net with classifier and no-classifier and the generated mask from each model.

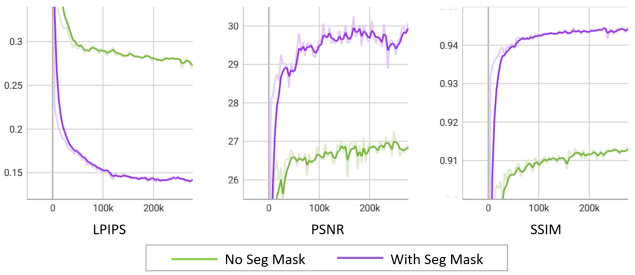

Figure 2. Validation results during training using segmentation mask (FRR-Net) vs no segmentation mask.

6. More Visulization Results

Screen Illuminations Results: In a video conference application, especially in low ambient lighting conditions, lighting from the screen may reflect on the face. However, there is no real-world public dataset for such scenarios. Therefore, in our degradation model, we apply jitter as well as other light/exposure distortions to synthetically generate these cases. In Figure 5, several face samples with screen illuminations are shown. As can be seen from this figure, our FRR-Net model smoothly reduces the illumination and generate a natural skin color.

Face Occlusions: In this paper, we focused on single face restoration and similar to other work in this area, we did not cover occlusions or other challenges in face detection or segmentation as this is out of the scope of this work. However, in a real-world video application occlusion may happen. Therefore, we tested our model on some of these cases as shown in Figure 6. From these results, we can see that the model can enhance the whole detected regions (for example, two overlapped faces).

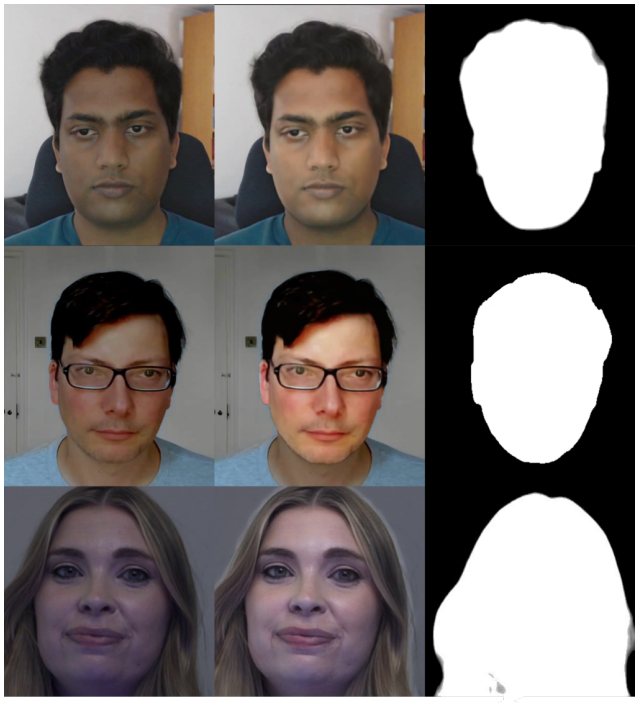

Figure 3. Samples of input (left) and outputs of FRR-Net including the enhanced face (middle), and facial mask (right).

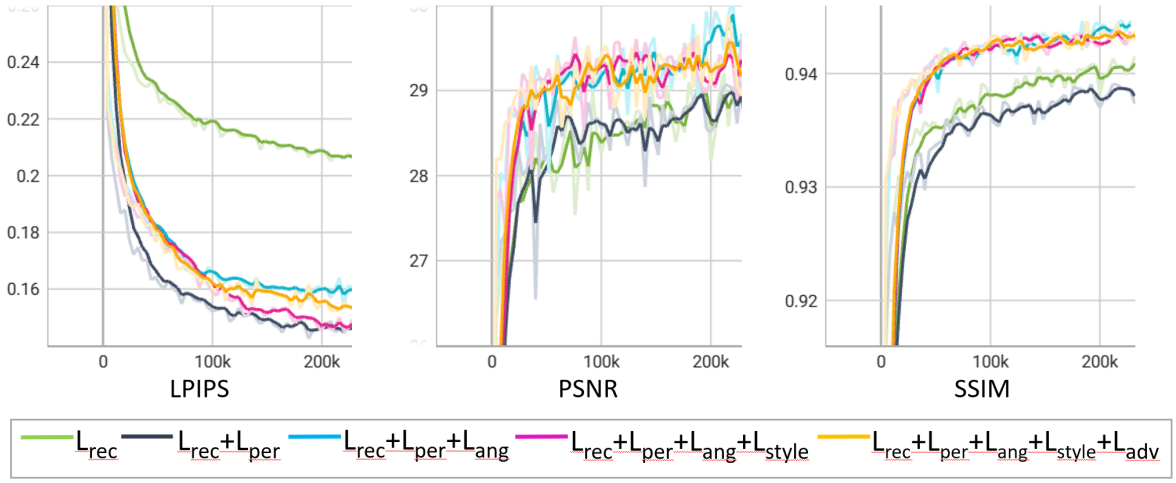

Figure 4. Importance of various losses on the evaluation metrics.

Figure 5. Real-world samples with screen illuminations.

Figure 6. Real-world samples with face occlusion.