



Reconstruction of encrypted faces for presentation attacks on a face recognition scheme.

Armin Niedermüller, Ahmet Bozkurt



Goals

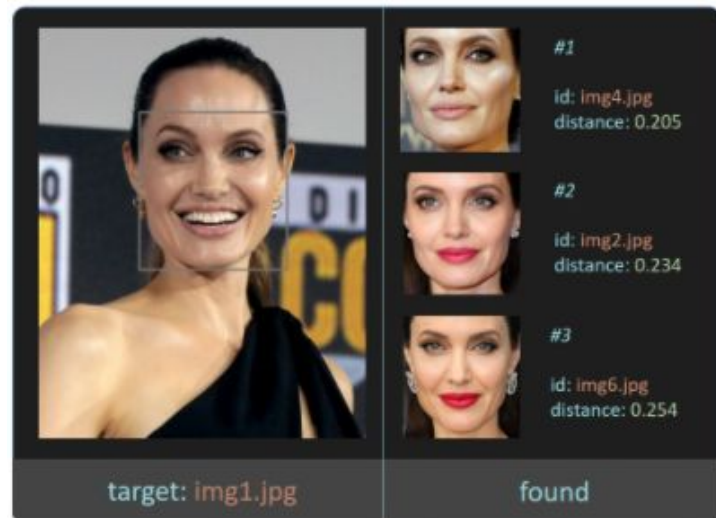
1. Set up a face recognition scheme - in our case: DeepFace using Google FaceNet
2. Test different models to enhance and / or generate plain faces from encrypted faces.
 - a. Use pretrained model weights to generated results
 - b. Train the model weights with our own data (encrypted and plain faces)
3. The goal is to trick a face recognition scheme into recognizing our data as valid results.

Face Recognition Benchmark - DeepFace FaceNet

Link: <https://github.com/serengil/deepface>

"A face recognition and facial attribute analysis (age, gender, emotion and race) framework for python. It is a hybrid face recognition framework wrapping state-of-the-art models. Google FaceNet is used for recognition."

This framework outputs a face distance between two images. A distance ≤ 0.4 means that both faces are from the same person.



Attack Method 1: Pixel2Style2Pixel (pSp) - Super Resolution

Link: <https://github.com/eladrich/pixel2style2pixel>

“A generic image-to-image translation framework, consisting of different submodules for different tasks.”

We tried the submodule “Super-Resolution”. Our hypothesis is, that the visually good-looking results of blurred-image-reconstruction (image on the right) can be transferred to encrypted-image-reconstruction.

The network will be trained with our own data, consisting of plain faces and their encrypted counterparts.





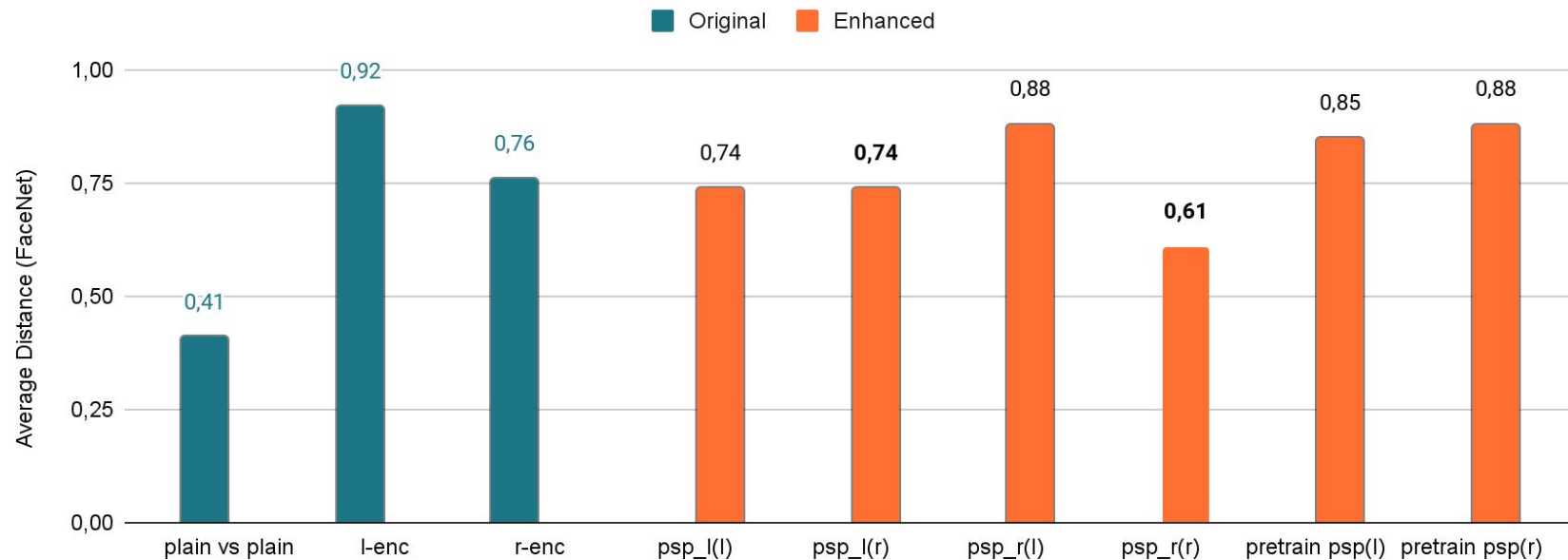
pSp - Methodology

- Training:
 - 175k iterations per encryption method 40 Hours (20 hours for each dataset) on:
 - RTX 3090
 - 32 GB DDR3 3600 Mhz
 - 5950 16 Cores / 32 Threads
- Data:
 - Test: 50 faces / 1.392 images
 - Train: 10522 faces / 489.231 images
- Comparison on FaceNet:
 - Intra class with the last 15 faces of each test set with 5 images each = 75 faces











Results pSp



Intra Class Face Comparison - PSPGAN



Results pSp - Images

<p>P: 0,41</p>  <p>plain</p>	<p>L: 0,72 / R: 0,76</p>  <p>l_{enc}</p>	<p>L: 0,85 / R: 0,88</p>  <p>$psp_pre\{l_{enc}\}$</p>	<p>L: 0,74 / R: 0,74</p>  <p>$psp_l(l_{enc})$</p>	<p>L: 0,88 / R: 0,61</p>  <p>$psp_r(l_{enc})$</p>
 <p>plain</p>	 <p>r_{enc}</p>	 <p>$psp_pre\{r_{enc}\}$</p>	 <p>$psp_l(r_{enc})$</p>	 <p>$psp_r(r_{enc})$</p>

Attack Method 2: : GFPGAN

Link: <https://github.com/TencentARC/GFPGAN>

"GFPGAN aims at real-world face restoration."

As an alternative to face generation / inpainting (pSp), we wanted to try a network which is specialized on denoising and restoration. Looking at the examples, the network not only seems capable of deblurring but also denoising.

The network will be trained with our own data, consisting of plain faces and their encrypted counterparts.



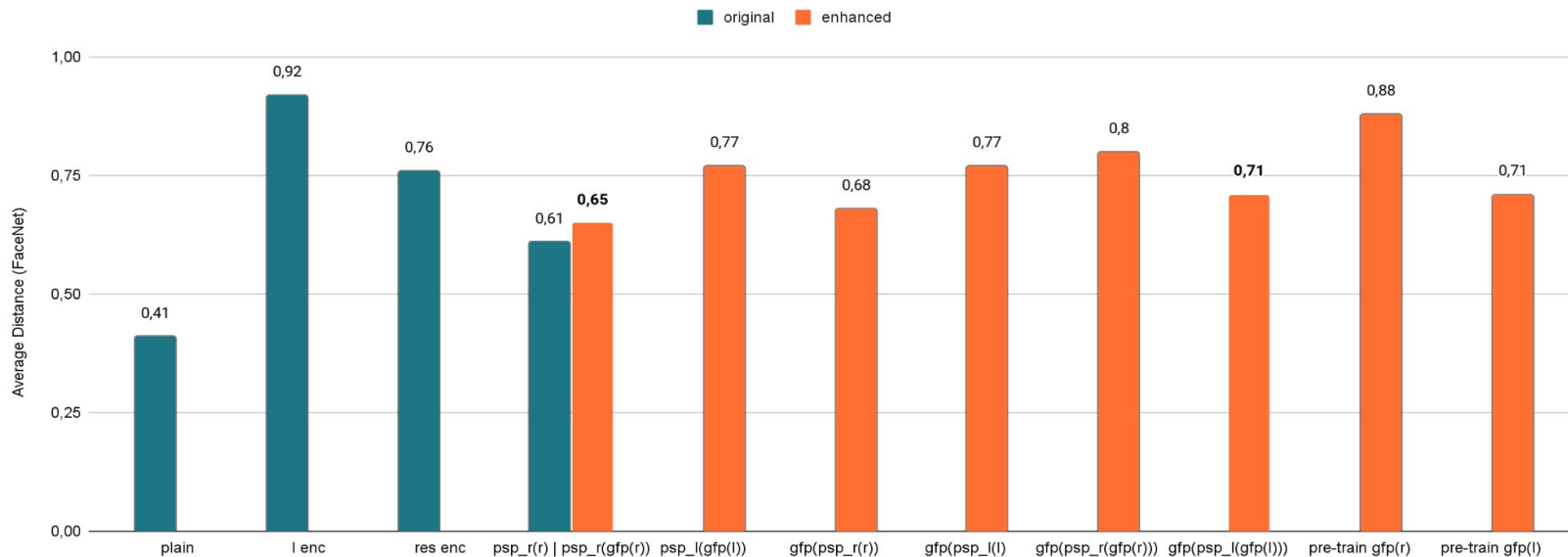


GFP GAN - Methodology










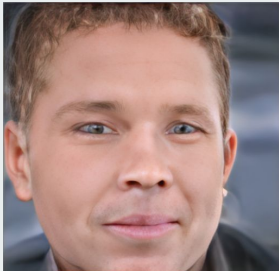


- Training did not work due to not solvable problems in their code
- Even if it would work, the specific distortions from the encrypted data cannot be trained, since GFP GAN creates distorted images from the plain data automatically
- However, GFP Gan showed some very good results in denoising and was able to generate features such as glasses, which pSp was not. Furthermore, blurry faces from old family pictures did really look like the person and not like some famous person.
- Hypothesis: GFP GAN could be able to improve our results, even with a pretrained model and thus we wanted to test it for post and preprocessing of the PSP Network
- Comparison on FaceNet:
 - Intra class with the last 15 faces of each test set with 5 images each = 75 faces

GFPGAN Results (only pretrained)

Intra Class Face Comparison - GFP GAN / PSP GAN



GFPGAN Results (only pretrained) - Images

<p>P: 0,41</p>  <p>plain</p>	<p>L: 0,72 / R: 0,76</p>  <p>l_enc</p>	<p>L: 0,71 / R: 0,88</p>  <p>gsp_pre{l_enc}</p>	<p>L: 0,77 / R: 0,68</p>  <p>gsp{psp_l(l_enc)}</p>	<p>L: 0,77 / R: 0,65</p>  <p>psp_l[gsp{l_enc}]</p>	<p>L: 0,71 / R: 0,80</p>  <p>gsp(psp_l[gsp{l_enc}])</p>
 <p>plain</p>	 <p>r_enc</p>	 <p>gsp_pre{r_enc}</p>	 <p>gsp{psp_r(r_enc)}</p>	 <p>psp_r[gsp{r_enc}]</p>	 <p>gsp(psp_r[gsp{r_enc}])</p>

Attack Method 3: BSRGAN

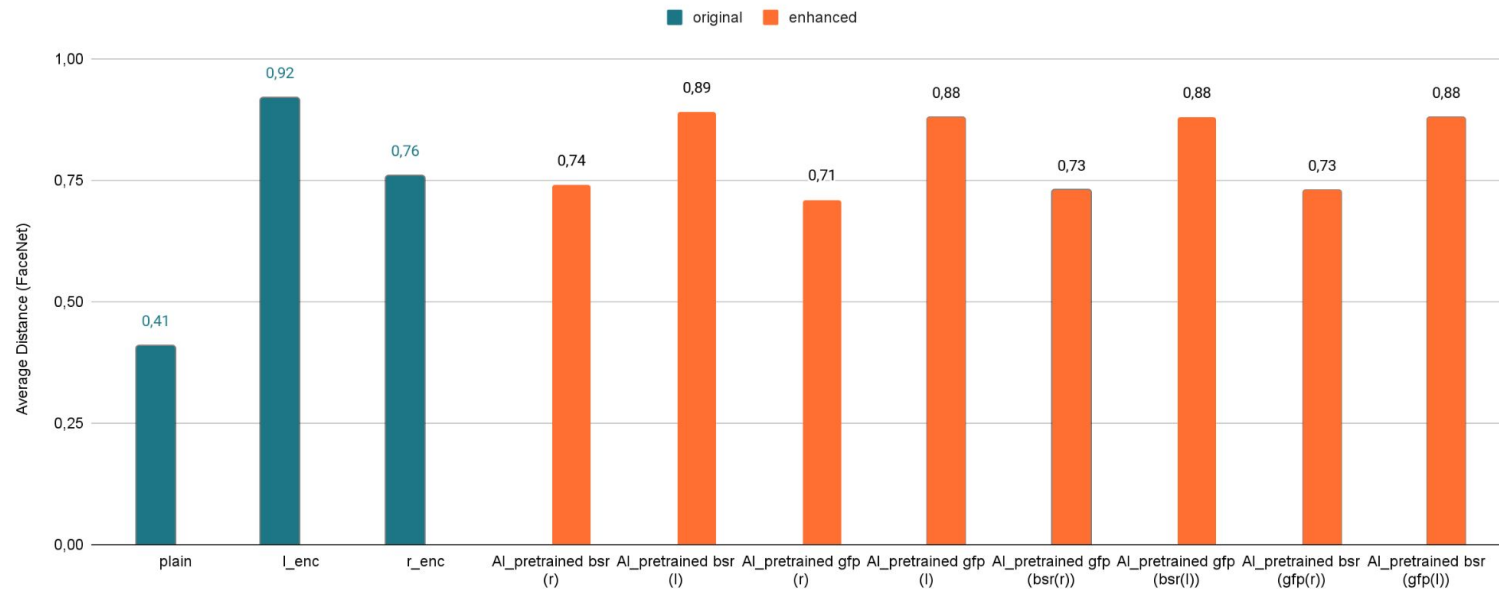
Link: <https://github.com/cszn/BSRGAN>

“Designing a Practical Degradation Model for Deep Blind Image Super-Resolution”





BSRGAN Results (only pretrained)

Intra Class Face Comparison - BRS GAN / GFP GAN



BSRGAN Results (only pretrained) - Images

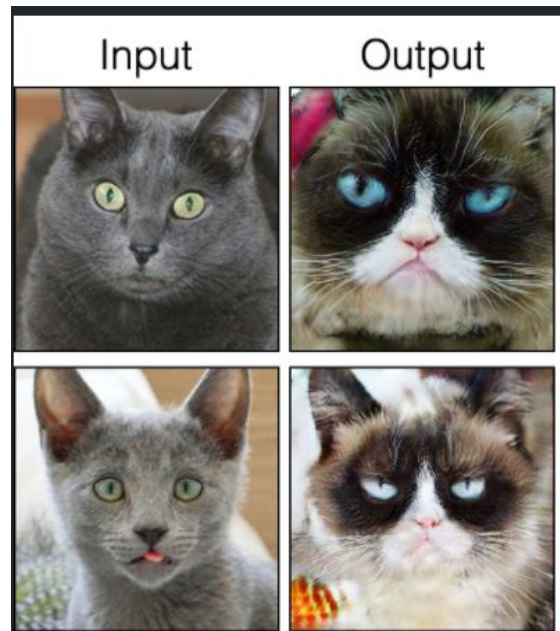
<p>P: 0,41</p>  <p>plain</p>	<p>L: 0,92 / R: 0,76</p>  <p><code>l_enc</code></p>	<p>L: 0,89 / R: 0,74</p>  <p><code>bsr_pre{l_enc}</code></p>	<p>L: 0,88 / R: 0,73</p>  <p><code>bsr(gfp{l_enc})</code></p>	<p>L: 0,88 / R: 0,73</p>  <p><code>gfp(bsr{l_enc})</code></p>
 <p>plain</p>	 <p><code>r_enc</code></p>	 <p><code>bsr_pre{r_enc}</code></p>	 <p><code>bsr(gfp{r_enc})</code></p>	 <p><code>gfp(bsr{r_enc})</code></p>

Attack Method 4: CUT

Link:

<https://github.com/taesungp/contrastive-unpaired-translation>

"We provide our PyTorch implementation of unpaired image-to-image translation based on patchwise contrastive learning and adversarial learning. No hand-crafted loss and inverse network is used. Compared to CycleGAN, our model training is faster and less memory-intensive. In addition, our method can be extended to single image training, where each "domain" is only a single image."





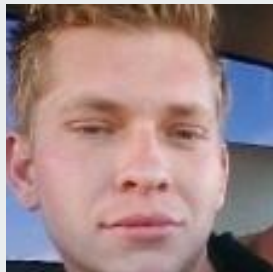
CUT Methodology

- Training:
 - 8k iterations per encryption method (approx. 1h) on: Each iteration takes 8 min.
 - Google Colab - Nvidia Quadro GPU / 11 GB RAM
- Data:
 - Train: 5 faces / 25 images
- Comparison on FaceNet:
 - Intra class with the last 15 faces of each test set with 5 images each = 75 faces

CUT Results - Images



P: 0,41



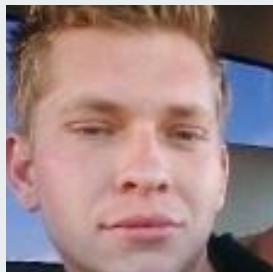
plain

R: 0,76



l_enc

R: 0,77



plain



r_enc



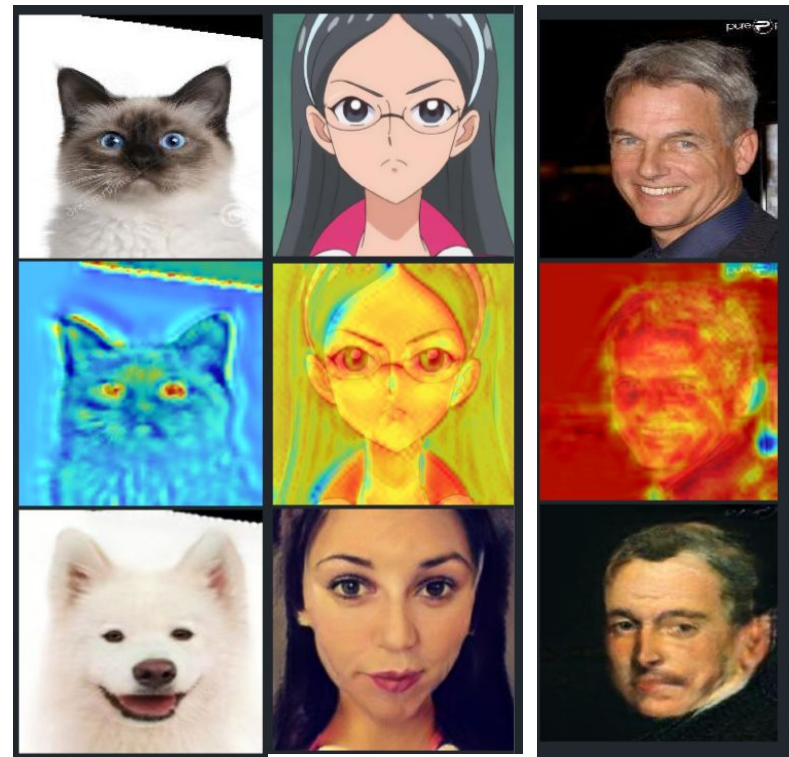
cut{r_enc}

Attack Method 5: U-GAT-IT

Link: <https://github.com/znxlw/UGATIT-pytorch>

"We propose a novel method for unsupervised image-to-image translation, which incorporates a new attention module and a new learnable normalization function in an end-to-end manner."

Performs better as cycleGAN



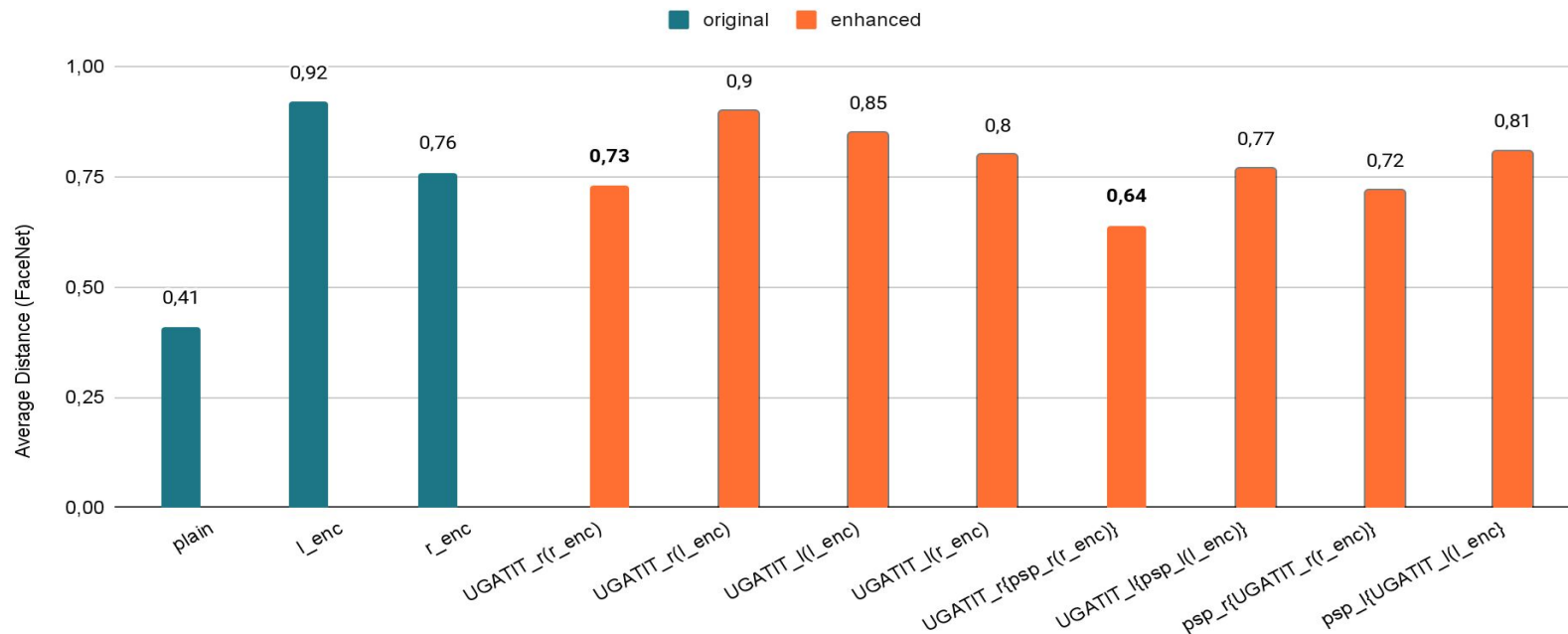


U-GAT-IT - Methodology

- **Training:**
 - 100k iterations per encryption method (2 x 20 hours = 40 hours of training) on:
 - RTX 3090 - 24 GB RAM
 - 32 GB DDR3 3600 Mhz
 - 5950 16 Cores / 32 Threads
- **Data:**
 - Test: 15 faces / 75 images
 - Train: 10522 faces / 489.231 images
- **Comparison on FaceNet:**
 - Intra class with the last 15 faces of each test set with 5 images each = 75 faces

U-GAT-IT Results

Intra Class Face Comparison - UGATIT / PSP



U-GAT-IT / PSP Results - Images

<p>P: 0,41</p>  <p>plain</p>  <p>plain</p>	<p>L: 0,92 / R: 0,76</p>  <p>l_{enc}</p>  <p>r_{enc}</p>	<p>L: 0,85 / R: 0,73</p>  <p>UGATIT_l(l_{enc})</p>  <p>UGATIT_r(r_{enc})</p>	<p>L: 0,9 / R: 0,8</p>  <p>UGATIT_r(l_{enc})</p>  <p>UGATIT_l(r_{enc})</p>	<p>L: 0,77 / R: 0,64</p>  <p>UGATIT_l($psp_l\{l_{enc}\}$)</p>  <p>UGATIT_r($psp_r\{r_{enc}\}$)</p>	<p>L: 0,81 / R: 0,72</p>  <p>$psp_l(UGATIT_l\{l_{enc}\})$</p>  <p>$psp_r(UGATIT_r\{r_{enc}\})$</p>
--	--	--	--	--	--

Final Results

	Encrypted Images	Generated by PSP (trained)	Generated by GFPGAN (pretrained)	Generated by U-GAT-IT (trained)	Generated by BSR GAN (pretrained)	Generated by CUT (trained)
Layer Encryption	0,92	0,74 (-0,18)	0,88 (-0,04)	0.85 (-0.07)	0.89 (-0.03)	-
Resolution Encryption	0,76	0,61 (-0,15)	0,71 (-0.05)	0.73 (-0.03)	0.74 (-0.02)	0.77 (+0.01)
Iterations	-	175k		100k	-	8k
Hardware	-	GPU: RTX 3090 / 24 GB RAM: DDR4 3600 Mhz CPU: 5950X 16 Cores	GPU: RTX 3090 / 24 GB RAM: DDR4 3600 Mhz CPU: 5950X 16 Cores	GPU: RTX 3090 / 24 GB RAM: DDR4 3600 Mhz CPU: 5950X 16 Cores	-	Google Colab Nvidia Quadro GPU / 11 GB
Training Time	-	40 Hours (20 hours for each dataset)	no training	40 Hours (20 hours for each dataset)	no training	1 Hour
Training Set	-	10522 faces / 489231 images	-	10522 faces / 489231 images	-	5 faces / 25 images
Validation Set	-	50 faces / 1392 images	-	-	-	-
Test Set	-	15 faces / 75 images	-	15 faces / 75 images	-	15 faces / 75 images