A Quantitative Evaluation Framework of Video De-Identification Methods

Sathya Bursic*, Alessandro D'Amelio, Marco Granato, Giuliano Grossi, Raffaella Lanzarotti

PHuSe Lab, Department of Computer Science, University of Milan

Motivation

- We live in an era of privacy concerns which motivates a large research effort in face de-identification
- There is a general movement from hand-crafted to deep learning methods
- De-identification doesn't suffice as a measure of performance and utility
- We want the media to retain as much as possible structural information, thus preserving utility

Motivation

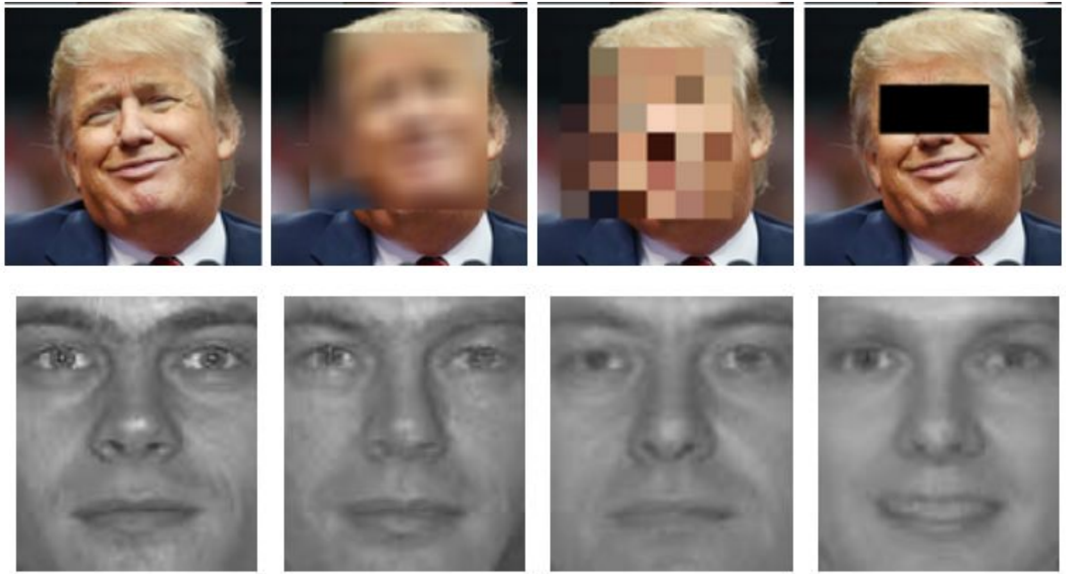


Figure 1: *First column*: original image. *First row*: naive methods, applying respectively blurring, pixelation, masking. *Second row*: results obtained by the k-same method.

Motivation

The three main requirements of a de-identification system:

- **The de-identification itself**, quantifiable as the capability of fooling face verification methods
- **Expression preservation**, measurable in terms of elicitation of the same Action Units (AUs) in both the original and the de-identified videos
- **The photo-reality safe-guard**, that we will measure in terms of feature preservation.

Face Swap Methods

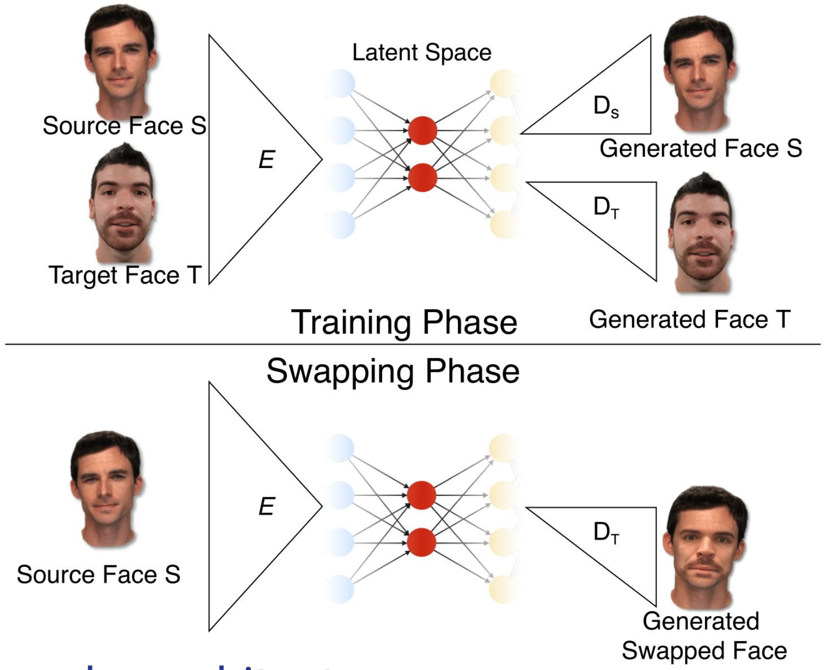


Figure 2: The autoencoder architecture

Face Swap Methods

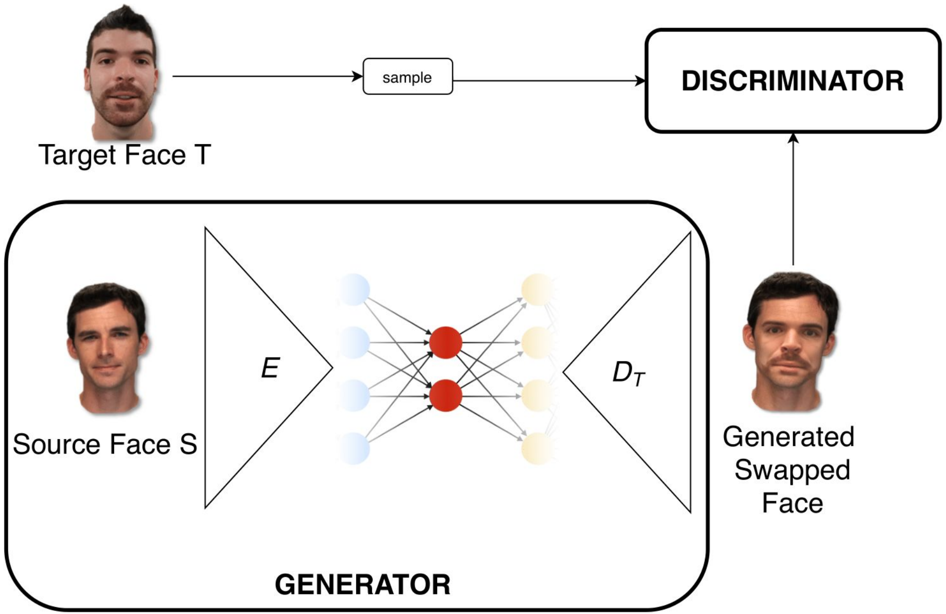


Figure 3: The GAN architecture

Methodology

We consider and compare four open-source methods:

- **Dfaker** (https://github.com/dfaker/df)
- **DeepFaceLab** (https://github.com/iperov/DeepFaceLab)
- **FaceSwap** (https://github.com/deepfakes/faceswap)
- **FaceSwap-GAN** (https://github.com/shaoanlu/faceswap-GAN)

on the RAVDESS dataset (5 male actors, 5 female actors) and train for 100000 iterations, taking snapshots every 5000 iterations and monitoring three metrics.

Methodology

1. **De-identification**: after calculating facial descriptors for the source, target and swapped subjects, respectively, we calculate the mean distances between the source/swapped, and target/swapped.
2. **Expression Preservation**: under the FACS framework, we extract AUs with OpenFace and produce PCC and RMSE over video frames.
3. **Photo-Reality**: we calculate the Fréchet Inception Distance (FID) between the original and swapped videos

Results - De-identification

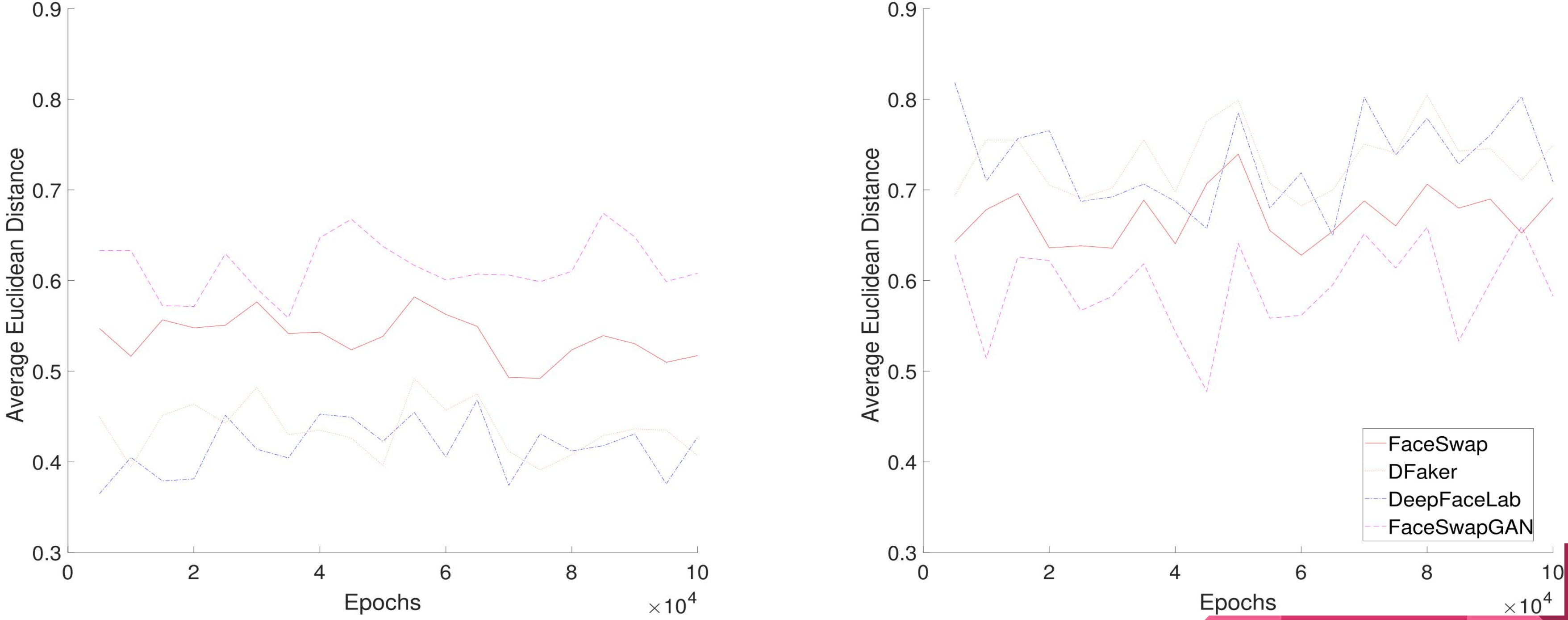
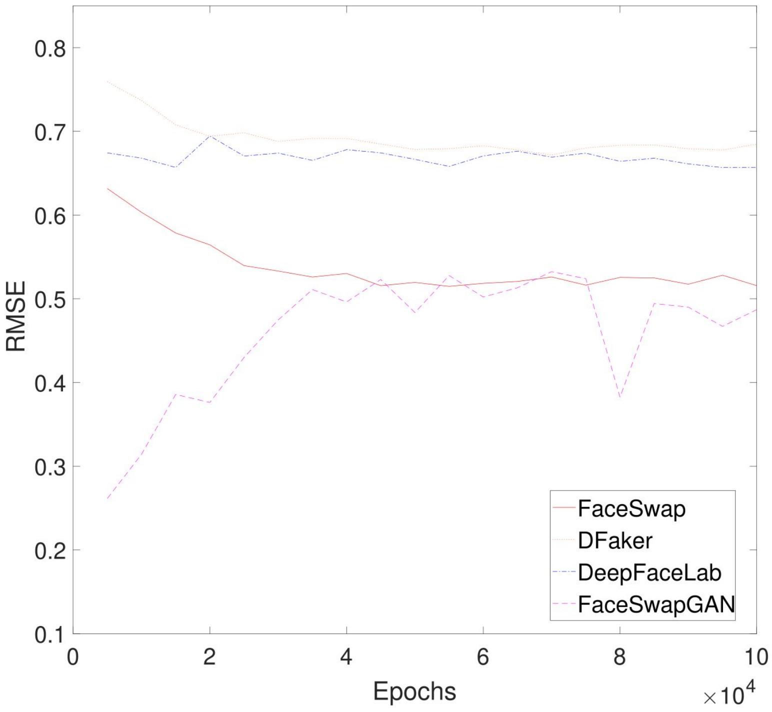


Figure 4. - The mean face descriptor distances for the source (left) and target (right) across epochs

Results - Expression Preservation



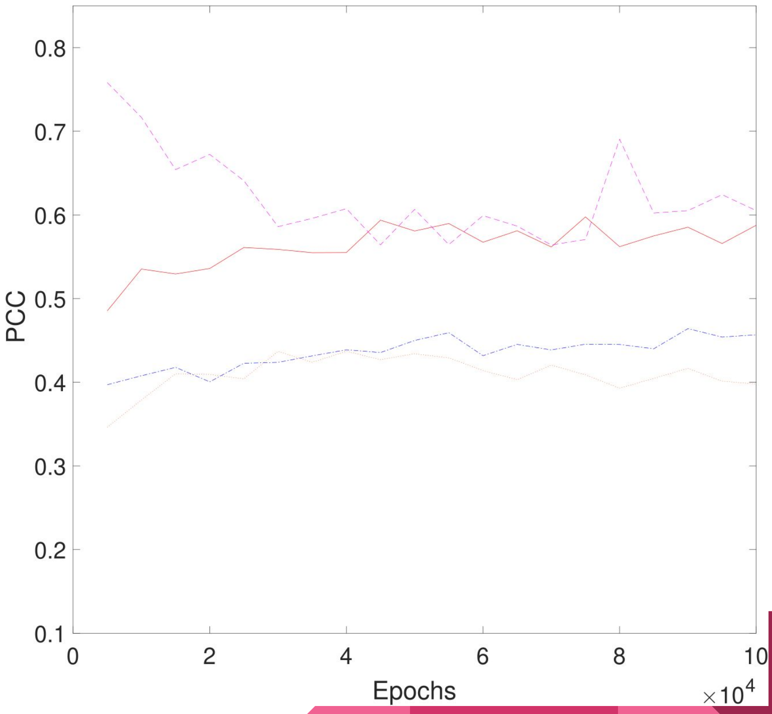


Figure 5. - The PCC and RMSE for AU intensity across epochs

Results - Photo-Reality

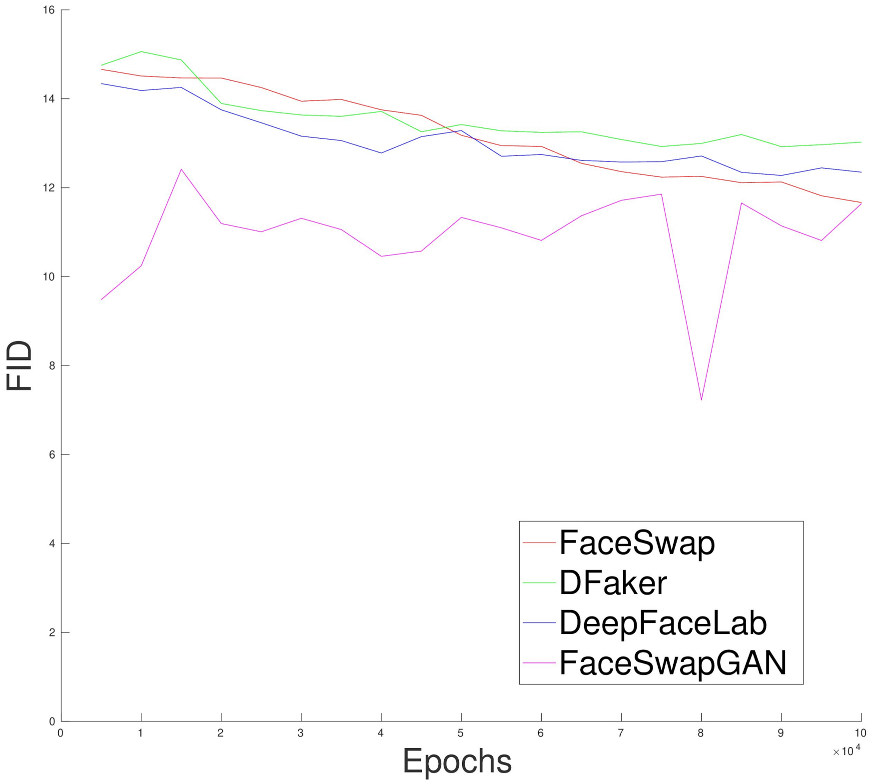


Figure 6. - The FID across epochs

Conclusions

- We introduced a quantitative evaluation framework for video de-id, and provide a baseline
- No one method is optimal according to the three metrics simultaneously, the objectives present a trade-off
- It is important to evaluate them jointly, in order to provide a complete picture of the method’s potential
- The last two metrics could be used as a stopping criteria for the training phase