Face De-identification for privacy protection summary

Dr. V. Mygdalis, E. Chatzikyriakidis, Prof. I. Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 3.3

Artificial Intelligence &
Information Analysis Lab

Face De-identification for privacy protection

- **Privacy and data protection**
- Classical face de-identification
- Autoencoder-based Face De-identification
- GAN-based de-identification
- Adversarial face de-identification
- K-anonymity attacks
- SVDD Adversarial Defense

Privacy and data protection

- Protection of personal data must be ensured in the acquired video and/or images.
- The EU's General Data Protection Regulation 2016/679), repealing the 1995 Data Protection Directive.
- "*Member States shall protect the fundamental rights and freedoms of natural persons and in particular their right to privacy, with respect to the processing and distribution of personal data.*"

Data protection issues in Autonomous Systems

- Public perceives AS as machines infringing privacy.
- No trespassing above private property.

- Distinguish between:
  - actors, spectators, crowd
  - public events, private events.

Artificial Intelligence & Information Analysis Lab

Data protection issues in drones

- Data protection issues for AV shooting:
  - broadcasting
  - creating experimental databases.

- Use of data de-identification algorithms when doing AV shooting.

Data anonymity requirements in AV data bases

- Data to be distributed must be ***anonymous***:
  - Any evidence that can be used to link acquired data to real people, is prohibited (e.g., address, names, etc.).
  - ***Facial images*** fall into the same category. They cannot be anonymous, since someone could link a facial image to a real person.
  - Soft biometric and non-biometric identifiers (fancy clothes, tattoos, skin marks, etc.) should be hindered as well.

Data anonymity requirements in AV data bases

- Image and video data collected by drones fall into the general data acquisition/shooting/distribution category.

- ***Consent forms must be collected for experimental AV data.***

- Standard AV shooting privacy-protection rules must be observed for AV data to be broadcasted.

Facial data protection approaches

- ***Face de-detection*** (Face detector obfuscation):
  - Apply image manipulations until face detection algorithms are no longer able to work
- ***Face de-identification*** (Face recognizer obfuscation):
  - Corrupt the facial region so that deep NN face classifiers fail.
- Developed methodology:
  - Simple/Naive approaches (additive noise, impulsive noise)
  - Reconstruction-based (SVD, PCA, hypersphere projections, auto-encoder-based) approaches.
  - Adversarial face de-identification.

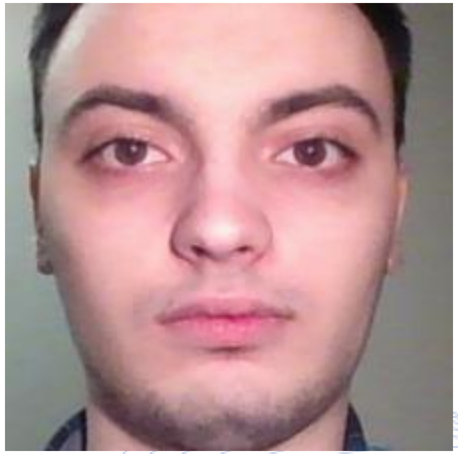
Personal image protection approaches

- Person de-detection
- Person de-identification
  - Human body images
- Personal object de-detection/de-identification
  - Car plates, car make.

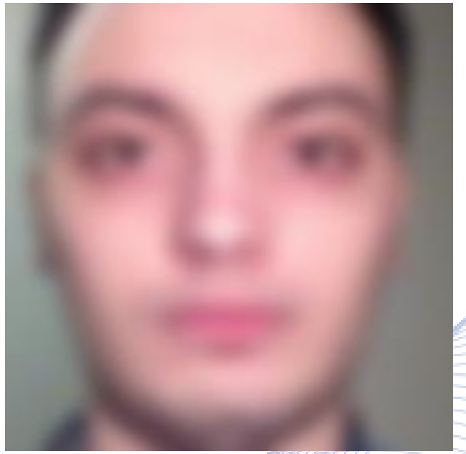
Face De-identification for privacy protection

- Privacy and data protection
- **Classical face de-identification**
- Autoencoder-based Face De-identification
- GAN-based de-identification
- Adversarial face de-identification
- K-anonymity attacks
- SVDD Adversarial Defense

Facial data protection approaches



Original Image



Gaussian blur with std. deviation of 5

Hypersphere projection with radius of 8

Face De-identification definitions

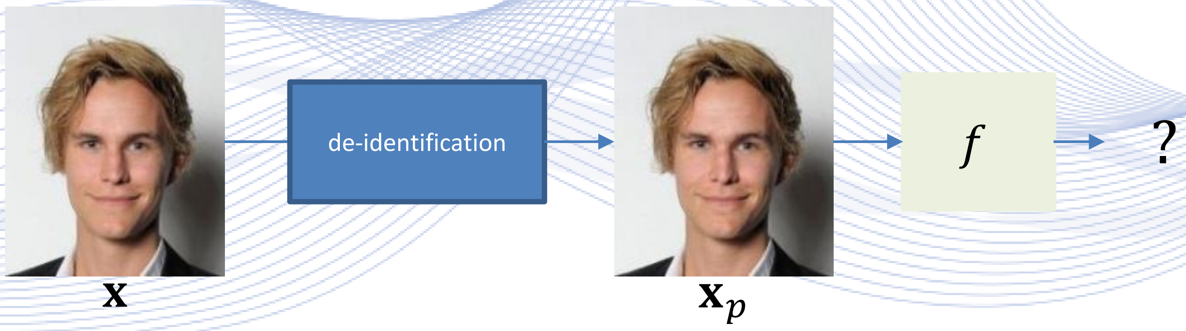
Face de-identification (DID) or ***Face recognition obfuscation*** tries to fool machine face recognition systems and/or face recognition by humans:

- Recognition by ***machines or humans*** (darkening, blurring, pixilation, additive noise methods, reconstruction-based methods, GAN-based methods)
- Machine recognition only (adversarial attacks).
- ***Focus on machine recognition obfuscation.***

Face De-identification definitions

Simple face de-identification definition:

- A trained face recognition system f take an input facial image $\mathbf{x}$ and predicts its corresponding identity label y: $f(\mathbf{x};\boldsymbol{\theta}) \rightarrow y$.
- Face de-identification methods aim to alter the original facial image $\mathbf{x}$ and produce a de-identified image $\mathbf{x}_p$ that can no longer be correctly identified: $f(\mathbf{x}_p;\boldsymbol{\theta}) \rightarrow ?$.



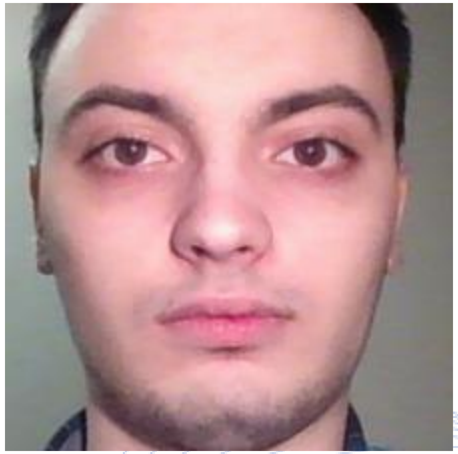
Face De-identification definitions

Formal face de-identification definition:

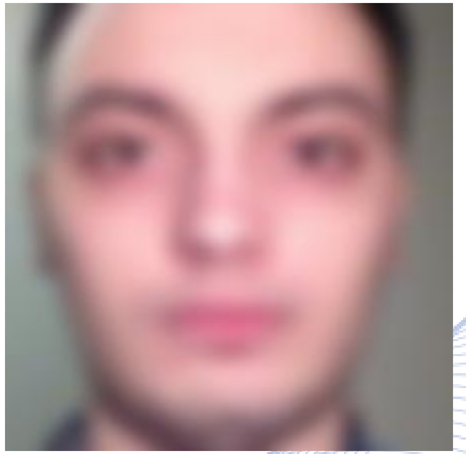
- Let $\mathbf{x} \in \mathbb{R}^n$ be a vector containing e.g., a ***facial image Region of Interest*** (ROI) representation with $y \in \{C_1, \dots, C_m\}$ its label. Function $f(\mathbf{x}; \boldsymbol{\theta}) = y$ is the ML recognizer/classifier.
- Face de-identification is about manipulating input vector $\mathbf{x}$ in some way, such as:
  - Perturbation: $\mathbf{x}_p = \mathbf{x} + \mathbf{p}$ (e.g., noise, pixelation, blurring, adversarial attacks)
  - Transformation: $\mathbf{x}_p = \mathbf{S}\mathbf{x} + \mathbf{p}$ (e.g., reconstruction methods)
  - Generative mapping function: $\mathbf{x}_p = \boldsymbol{G}(\mathbf{x}; \boldsymbol{\theta}_G): \mathbb{R}^n \mapsto \mathbb{R}^n$, (AE, GANS)
- They all force the face identifier to fail: $f(\mathbf{x}_p; \boldsymbol{\theta}) \neq y$.

Artificial Intelligence & Information Analysis Lab

Acceptable Image Quality Issues



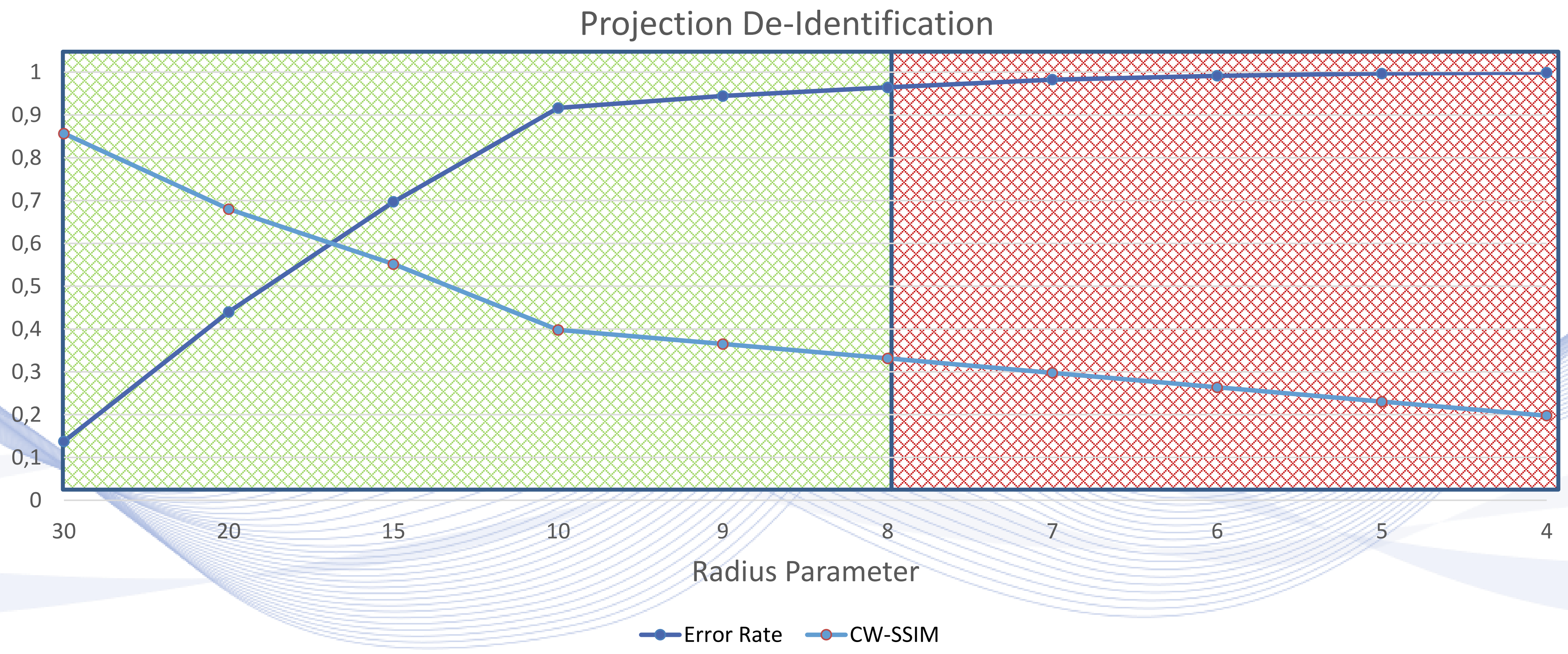
Original Image



Gaussian blur with std. deviation of 5

Hypersphere projection with radius of 8

Trade-off between de-identification performance and facial image quality

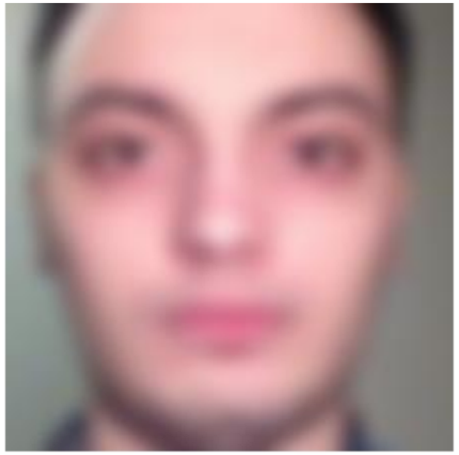


Artificial Intelligence & Information Analysis Lab

Face de-identification methods

- Naïve face de-identification refers to applying additive noise (e.g., Gaussian, impulse) to or blur the (detected) input facial image region, until the system fails to detect/classify the face.

Original Image

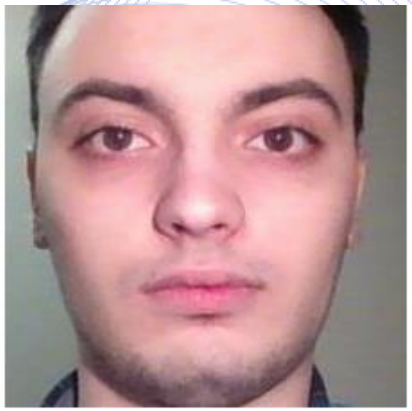


Gaussian blur with std. deviation of 5

Face de-identification methods

Reconstruction-based face DID approaches:

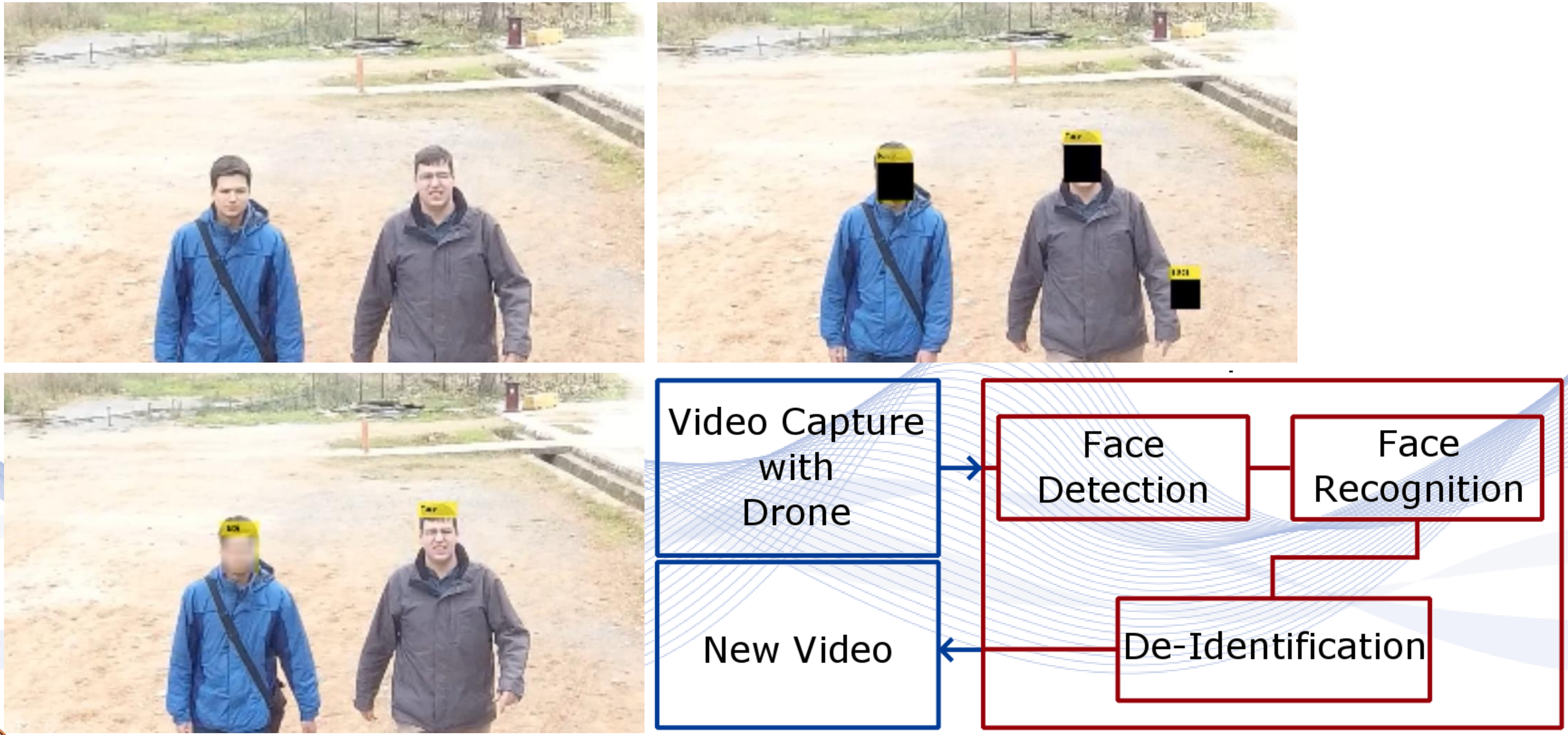
- Obtain facial image coefficients using some reconstruction method (e.g., PCA, SVD, Autoencoder).
- Apply modifications to these coefficients.
- Reconstruct a distorted facial image.



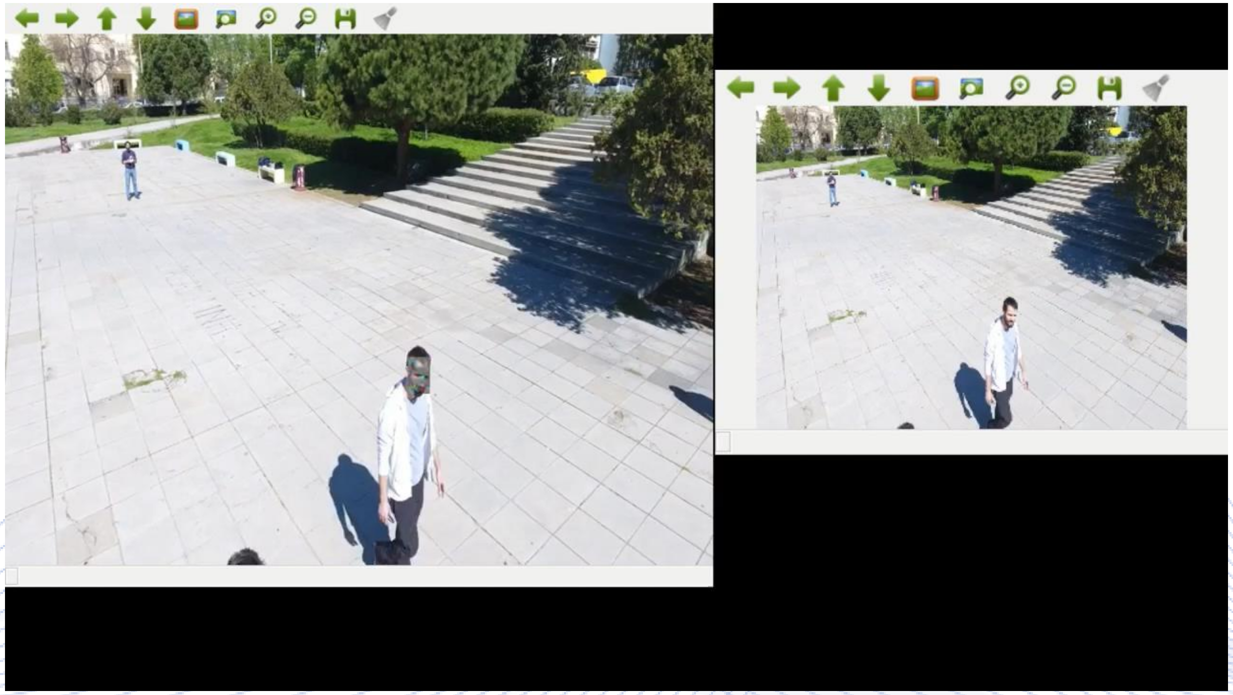
Original Image

Hypersphere projection with radius

Face de-identification on drone videos



Face de-identification on drone videos

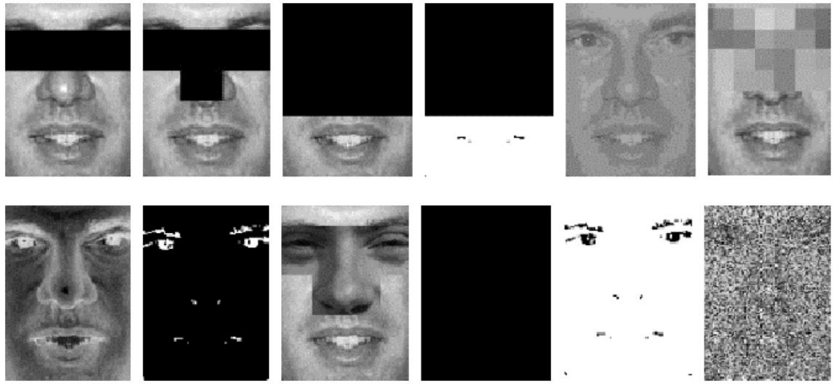


SVD-DID face de-identification in a drone video.

Face de-identification methods

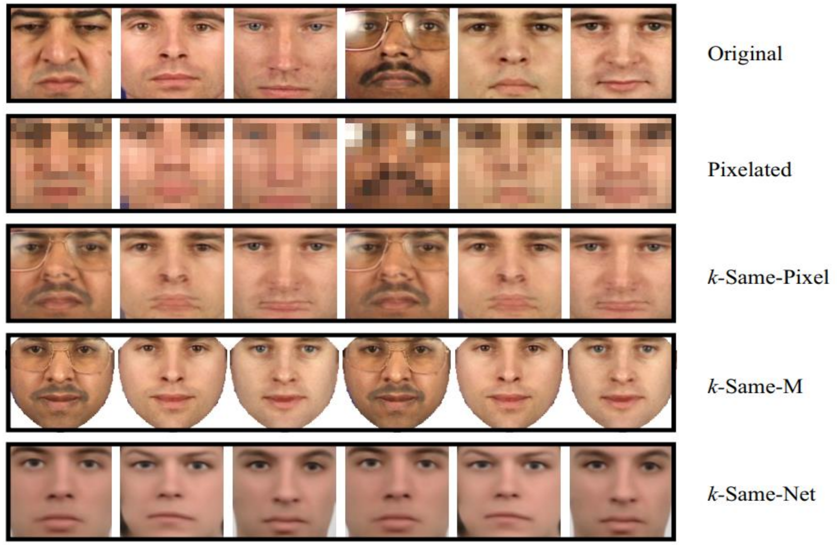
Drawbacks of previous face DID methods:

- They strongly alter original facial images.

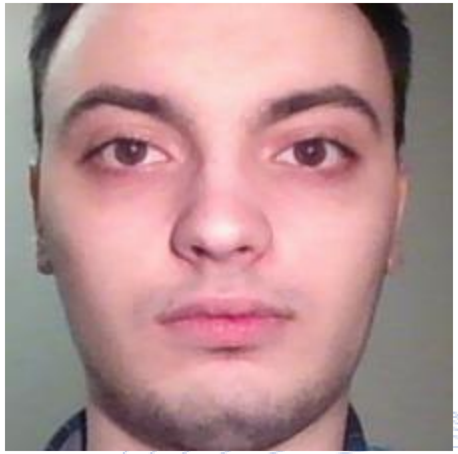


Desirable face DID method properties against machines:

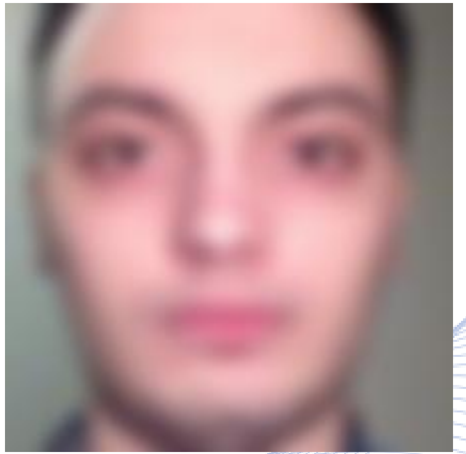
- De-identified image should retain the unique original facial image unique characteristics (e.g., race, gender, age, expression, pose).



Acceptable Image Quality Issues



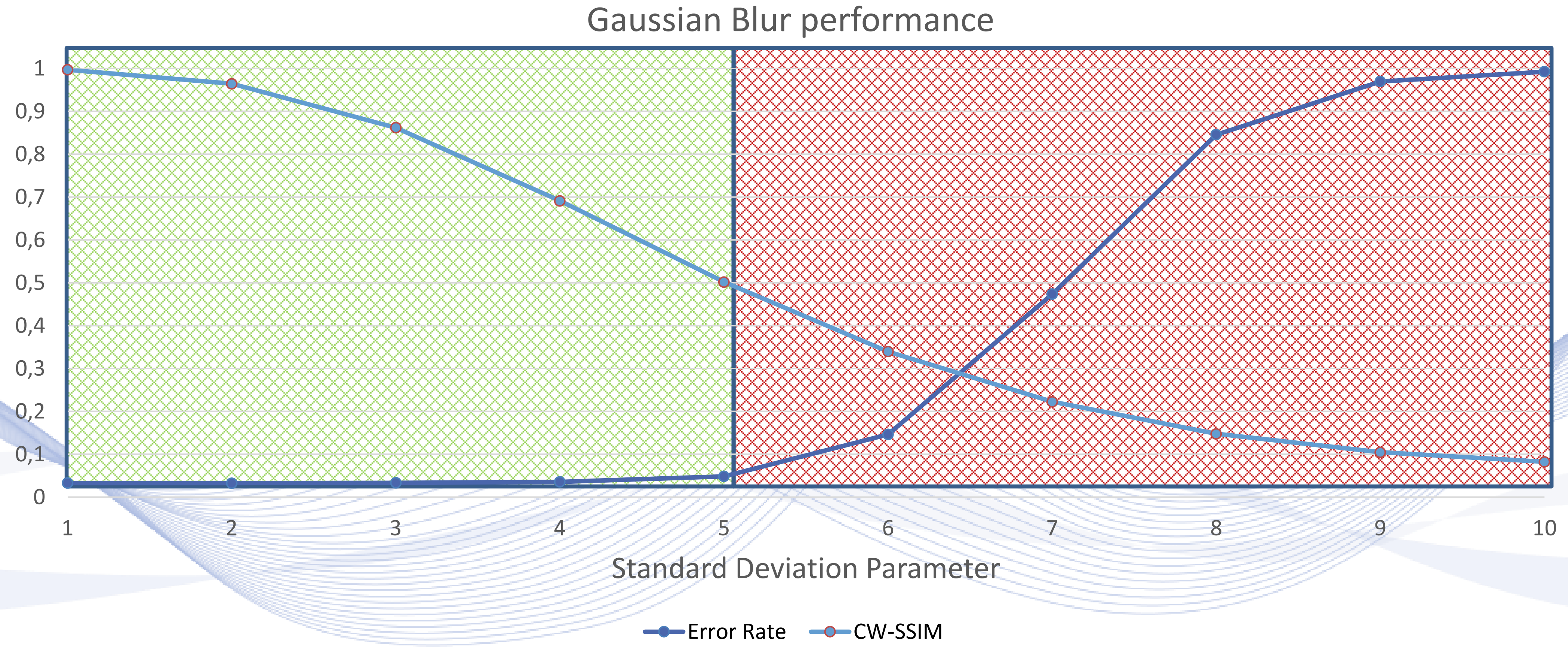
Original Image



Gaussian blur with std. deviation of 5

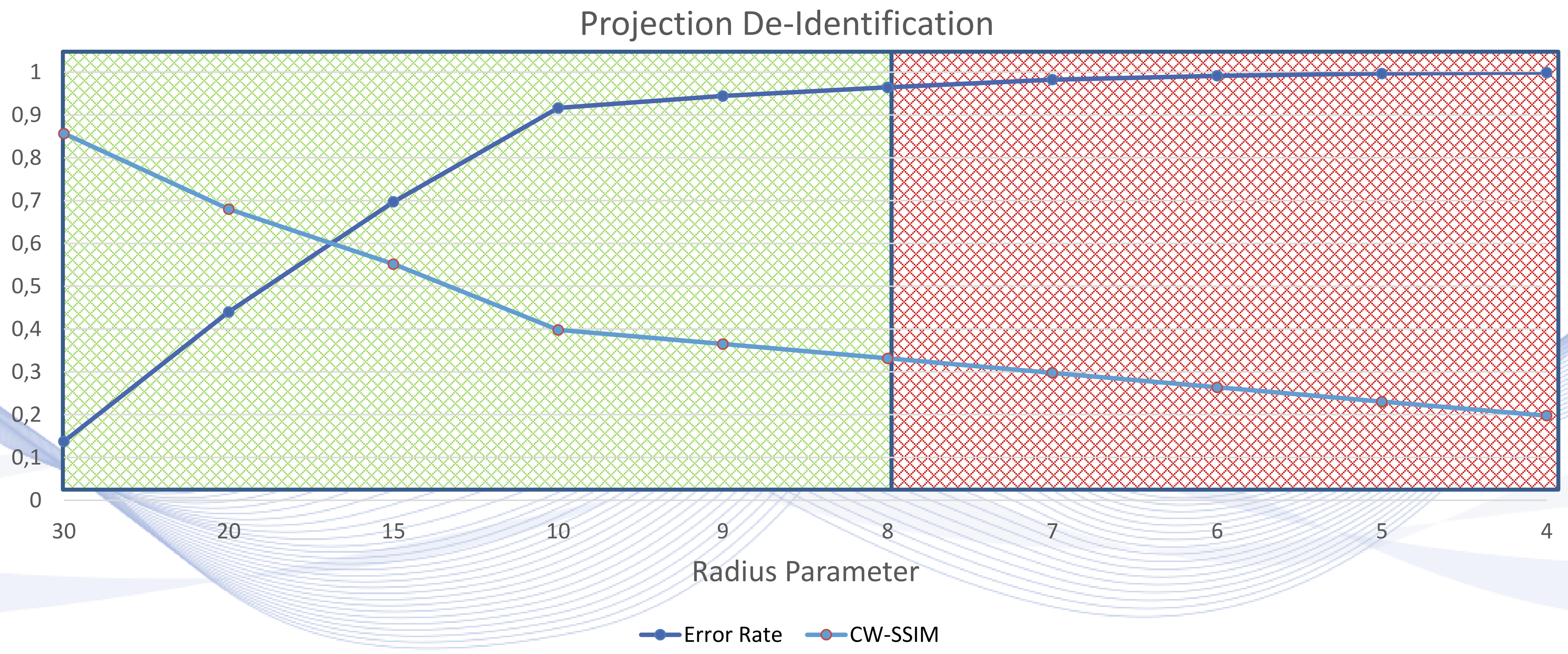
Hypersphere projection with radius of 8

Trade-off between de-identification performance and facial image quality



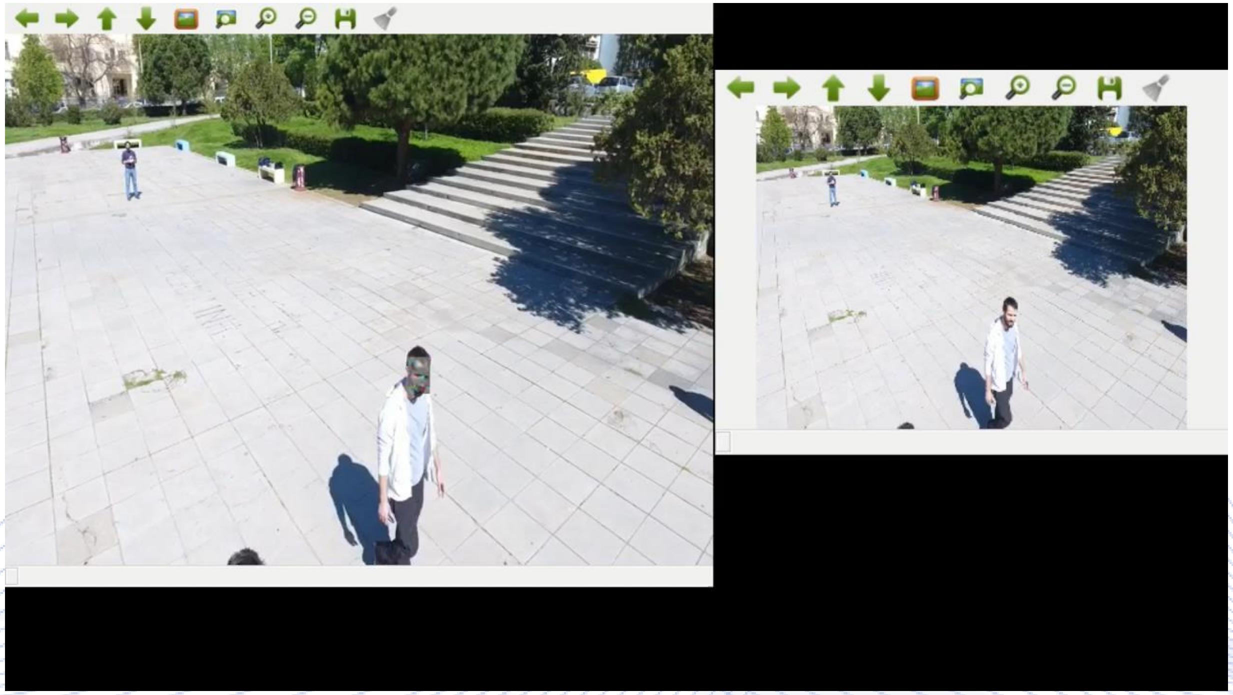
Artificial Intelligence & Information Analysis Lab

Trade-off between de-identification performance and facial image quality



Artificial Intelligence & Information Analysis Lab

SVD-DID



- Face de-identification in video.

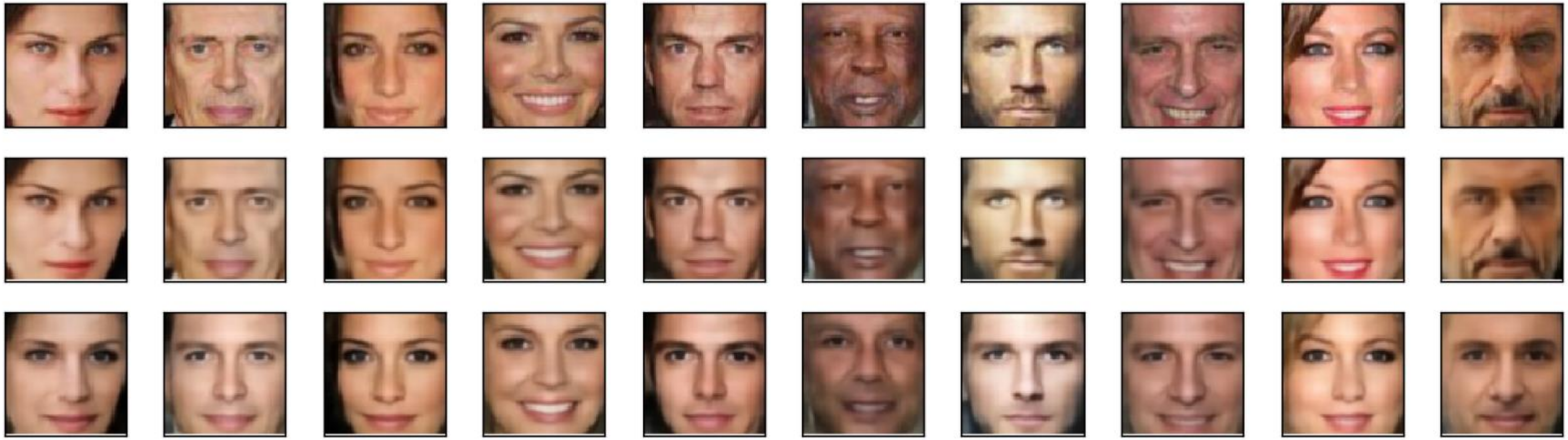
Face De-identification for privacy protection

- Privacy and data protection
- Classical face de-identification
- **Autoencoder-based Face De-identification**
- GAN-based de-identification
- Adversarial face de-identification
- K-anonymity attacks
- SVDD Adversarial Defense

Autoencoder-based Face De-identification

- Originating from reconstruction-based methods.
- Leverage deep autoencoders or even GANs for generating “fake” image content, that is recognizable neither by machines and humans.
- The de-identified facial image is produced by reconstruction, using a neural Autoencoder (AE).

Supervised Attribute Preserving Face DID



First row: original images; second row: images reconstructed by a standard AE, third row: Images reconstructed by Supervised Attributed Preserving DID.

Face De-identification for privacy protection

- Privacy and data protection
- Classical face de-identification
- Autoencoder-based Face De-identification
- **GAN-based de-identification**
- Adversarial face de-identification
- K-anonymity attacks
- SVDD Adversarial Defense

GAN-based face de-identification

GAN-based face de-identification extends AE-DID, by employing a Generator-Discriminator GD network pair, trained in an adversarial fashion. Given:

- source facial image $\mathbf{x}$ to be de-identified and its true label y.
- target 'wrong' facial image $\mathbf{t}$,

$\boldsymbol{G}$ calculates a reconstruction $\mathbf{x}_p = \boldsymbol{G}(\mathbf{x}, \mathbf{t}; \boldsymbol{\theta}_G)$ by:

  - minimizing the discrepancy between $\mathbf{x}_p$ and $\mathbf{t}$ or
  - by "learning the translation" of $\mathbf{x}$ to $\mathbf{t}$.

GAN-based face de-identification

- $\hat{d} = D(\mathbf{x}_p; \boldsymbol{\theta}_D)$ is a binary discriminator of whether $\mathbf{x}_p$ follows the distribution of $\mathbf{t}$, or not.
  - $\mathbf{x}, \mathbf{t}$ could be images belonging to the same class, or even completely different ones.
- If we feed the de-identified image $\mathbf{x}_p$ to a trained face recognizer $f(\mathbf{x}_p; \boldsymbol{\theta})$, it should not be able to identify it correctly $f(\mathbf{x}_p; \boldsymbol{\theta}) \neq y$.
- This pipeline leads to even more realistic image generations, when compared to AE-based de-identification.

Artificial Intelligence & Information Analysis Lab

GAN-based face de-identification

Live face de-identification in video [GAF2019].

Conditional Identity Anonymization GAN results [MAX2020].

GAN body image de-identification

Generative Full Body and Face De-Identification [BRK2017].

Face De-identification for privacy protection

- Privacy and data protection
- Classical face de-identification
- Autoencoder-based Face De-identification
- GAN-based de-identification
- **Adversarial face de-identification**
- K-anonymity attacks
- SVDD Adversarial Defense

Adversarial attacks & Defenses

Adversarial Attacks modify facial images to be wrongly identified.

- They may be employed for privacy protection.

Adversarial Defenses modify face recognition pipeline modules to make the pipeline robust to adversarial attacks.

- They be employed for content protection against adversarial attacks (e.g., copyright protection systems).

Adversarial Face de-identification

- Such methods perform de-identification by applying adversarial attacks on trained deep NN face recognizers.
- Adversarial attacks may be:
  - Targeted or un-targeted.
  - White-box or black box.
  - Iterative or single-step.
  - Transferable to different NN architectures/classification methods.
- The de-identified image is produced by returning gradient from a trained NN to the input facial image directly.
- They produce imperceptible facial image perturbations by humans.

Adversarial Face De-Identification

Iterative Fast Gradient Value Method (***I-FGVM***):

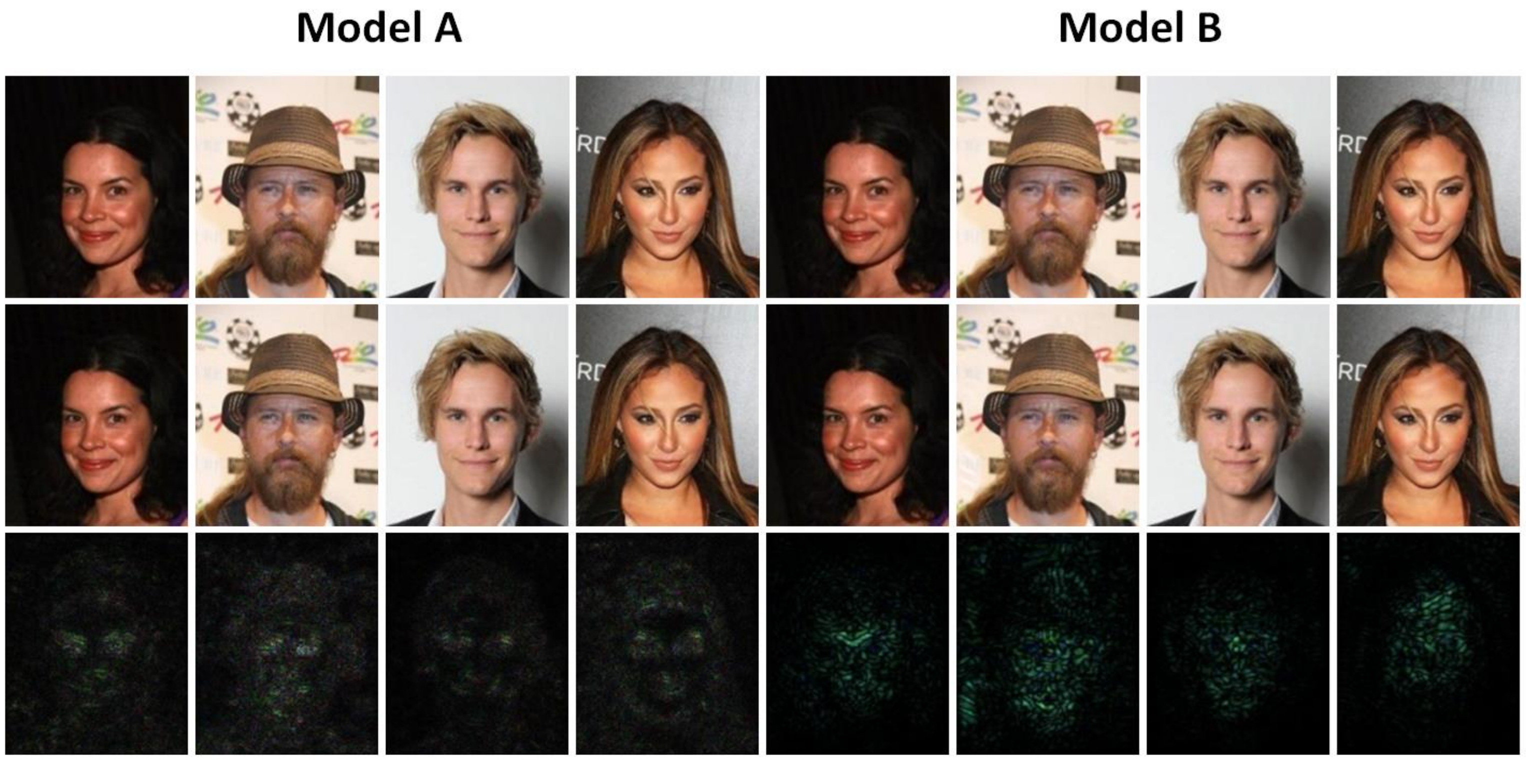
- Let images $\mathbf{x}$ have normalized pixel values in the domain $[0,1]$.
- The gradient descent update equations have the form:

$$\mathbf{x}_p^0 = \mathbf{x},$$

$$\mathbf{x}_p^{i+1} = \mathrm{clip}_{[0,1]}\left(\mathbf{x}_p^i - \alpha\nabla_{\mathbf{x}} J\left(\mathbf{x}_p^i, \mathbf{t}\right)\right).$$

- α is the step size, $\mathbf{x}$ is the original image, $\mathbf{x}_p^i$ is the adversarial image at step i,
- $J\left(\mathbf{x}_p^i, \mathbf{t}\right)$ is the adversarial loss,
- $\mathbf{t}$ is the target output vector class related to label target label t and
- $\mathrm{clip}_{[a,b]}$ is a constraint that keeps pixel values in the $[a, b]$ range.

Adversarial Face De-Identification



First row: original image; Second row: de-identified image. Third row: adversarial perturbation absolute value (x10) [CHA2019].

Face De-identification for privacy protection

- Privacy and data protection
- Classical face de-identification
- Autoencoder-based Face De-identification
- GAN-based de-identification
- Adversarial face de-identification
- **K-anonymity attacks**
- SVDD Adversarial Defense

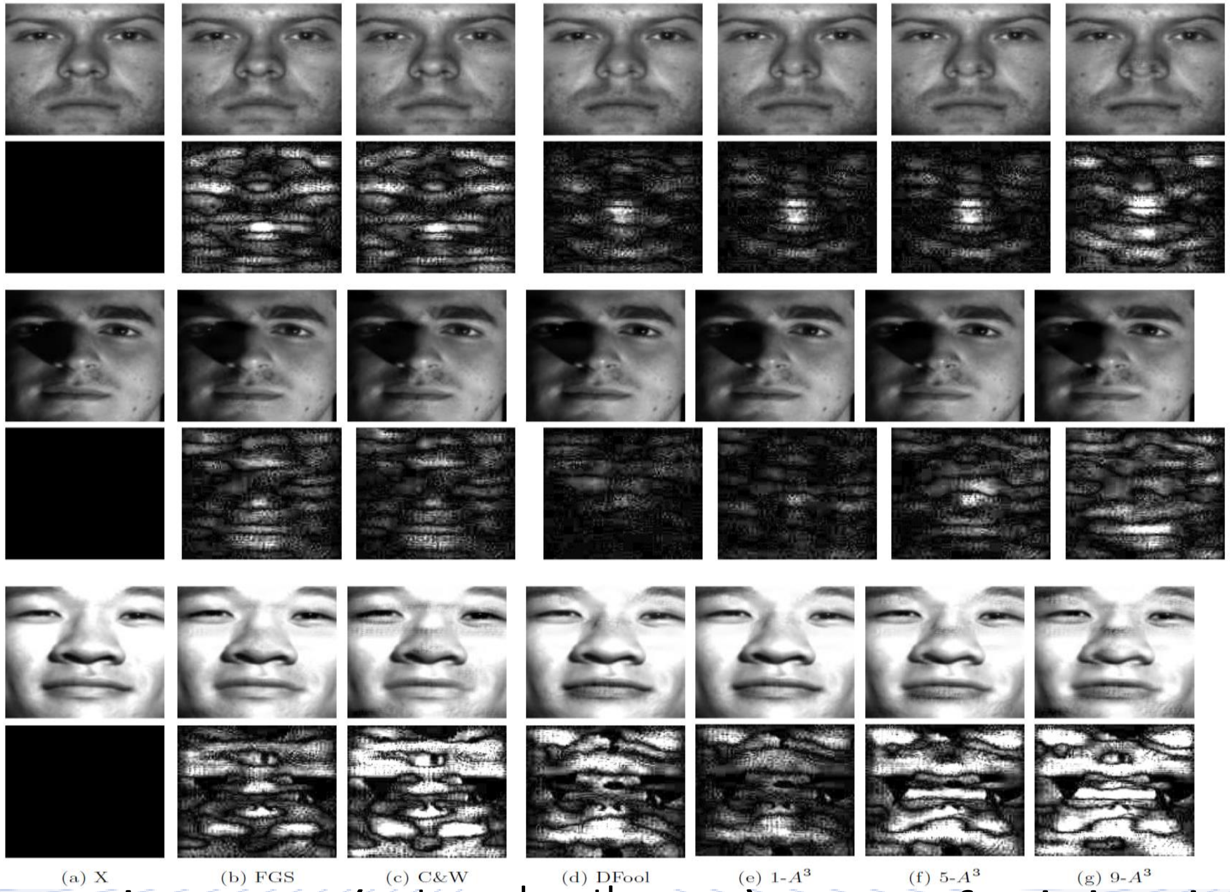
k-Anonymity-inspired adversarial attack

***k*-anonymity concept:**

- The maximum probability of retrieving a sample from a set must be less than $1/k$.
- Originally introduced in other research areas (e.g., Database research).

- In k-anonymity-inspired adversarial attack, the concept is altered as follows:
  - The maximum probability of retrieving the real person identity must be less than $1/k$, in every possible face classifier output ranking position.

$k-A^3$ face de-identification method

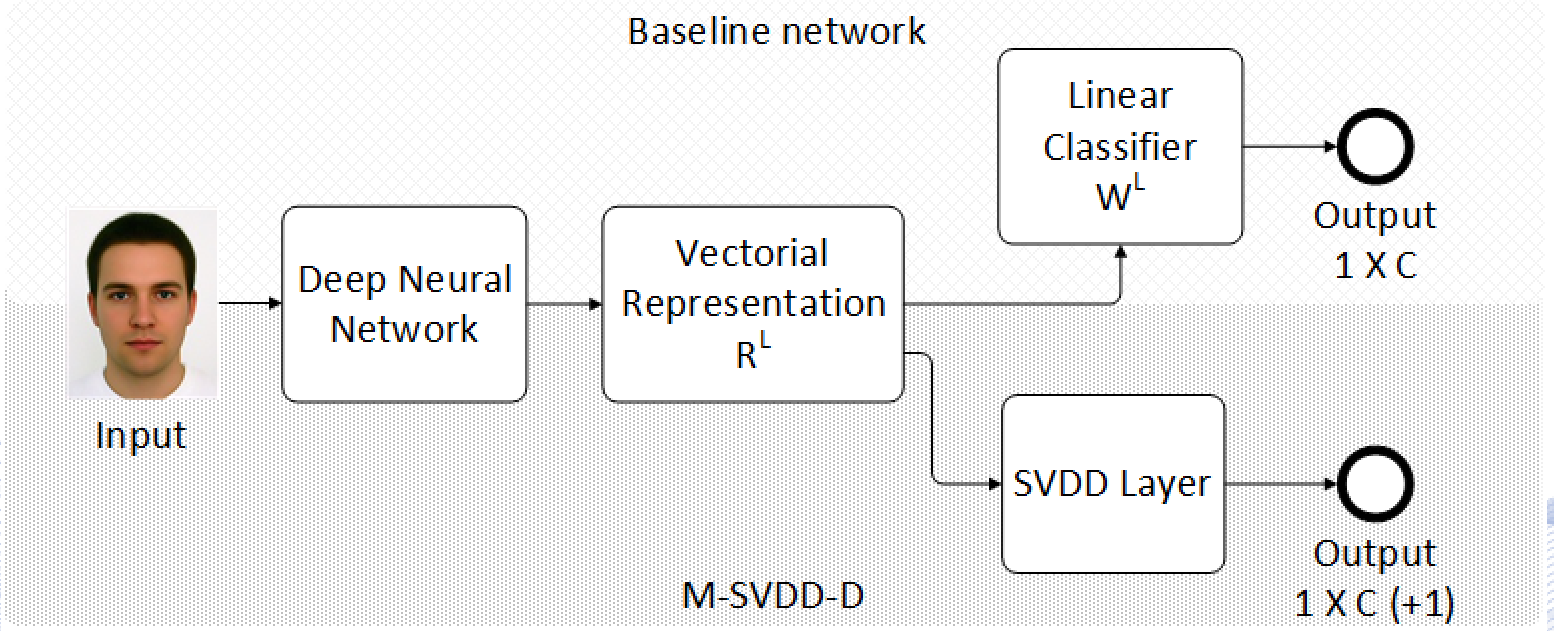


Face de-identification: original images (1st, 3rd, 5th row), magnified de-identification noise for various methods (2nd, 4th, 6th row, $k-A^3$ 3 right columns).

Face De-identification for privacy protection

- Privacy and data protection
- Classical face de-identification
- Autoencoder-based Face De-identification
- GAN-based de-identification
- Adversarial face de-identification
- K-anonymity attacks
- **SVDD Adversarial Defense**

m-SVDD Adversarial Defense



Bibliography

[MYG2020] V. Mygdalis, A. Tefas and I. Pitas, “*K-anonymity inspired Adversarial Attack and M-SVDD Defense*”, Neural Networks, Elsevier, vol. 124, pp. 296-307, 2020

[NOU2019] P. Nousi, S. Papadopoulos, A.Tefas and I.Pitas, "*Deep autoencoders for attribute preserving face de-identification*", Elsevier Signal Processing: Image Communication, 2019

[CHA2019] E. Chatzikyriakidis, C. Papaioannidis and I.Pitas, "*Adversarial Face De-Identification*" in Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019

[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.

[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).

[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.

[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Bibliography

[SZE2013] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199. 2013 Dec 21.

[BRK2017] Brkic, K., Sikiric, I., Hrkac, T., & Kalafatic, Z. *“I Know That Person: Generative Full Body and Face De-Identification of People in Images”,* Proc. CVPR, 2017

[GAF2019] Gafni, Oran, Lior Wolf, and Yaniv Taigman. "*Live face de-identification in video.*" Proc. IEEE International Conference on Computer Vision, 2019.

[MAX2020] Maximov, Maxim, Ismail Elezi, and Laura Leal-Taixé. "*CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks.*" Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr