Privacy protection in Natural Disaster Management

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.

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

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.***

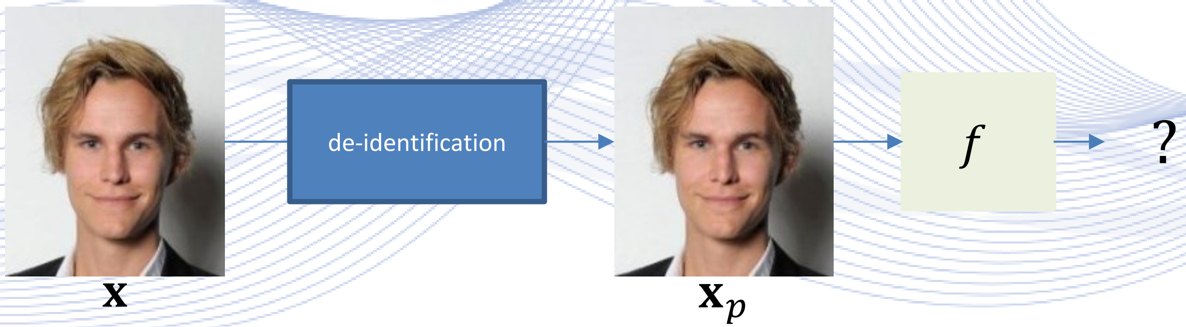
Face de-identification against humans

- "Traditional" privacy protection against **face recognition** aimed at hindering/disabling a **human identifier** from being able to distinguish a specific face in the image.
- Disadvantages:
  - These approaches were not designed to fool machines (automated face identification)
  - They typically deteriorate significantly image quality and produce "ugly" noisy images with minimal utility.
  - Some of them are ***completely naïve and fully inversible*** (e.g., image negation or darkening)

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

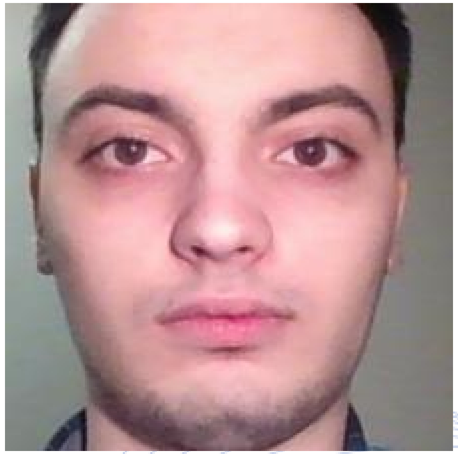
Formal face de-identification definition:

- 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: $f\left(\mathbf{x}_p; \boldsymbol{\theta}\right) \neq y$.

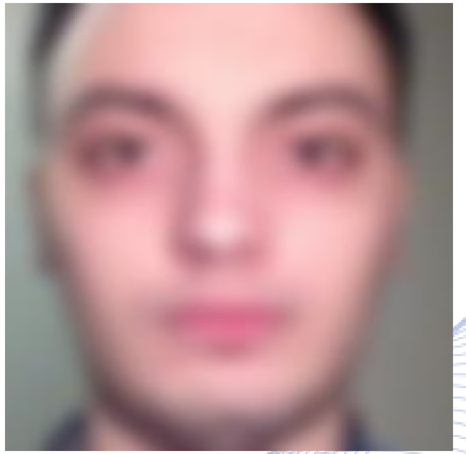
Face de-identification metrics

- Face de-identification performance against systems:
  - ***1-classification accuracy***.
- Face de-identification performance against humans.
- Similarity of the de-identified image with the original one:
  - e.g.: structural image similarity.
- Introduced image noise metrics (e.g., ***MSE***).
- Subjective image quality metrics:
  - perceived image quality, ***CW-SSIM,*** faceness, etc.

Acceptable Image Quality Issues



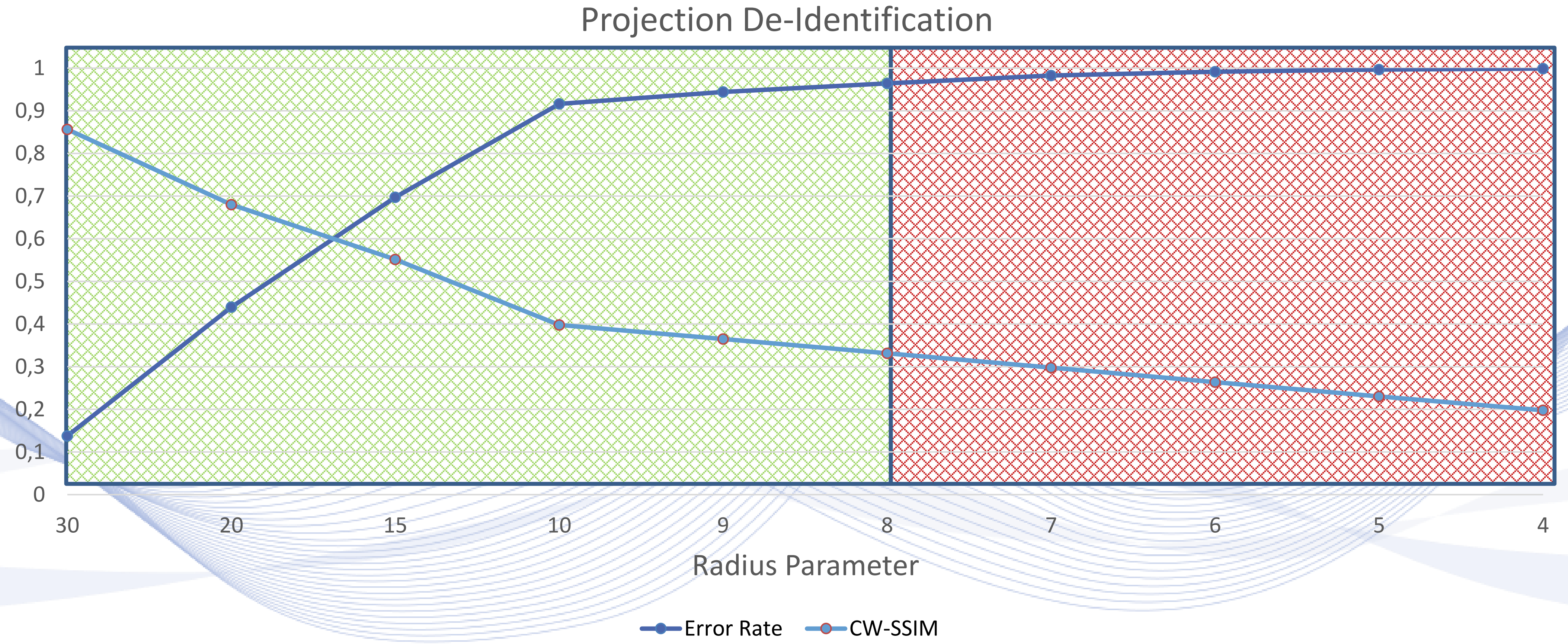
Original Image



Gaussian blur with std. deviation of 5

Hypersphere projection with radius of 8

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

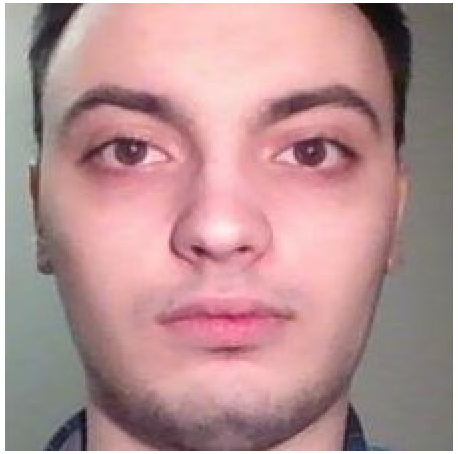


Face de-identification methods

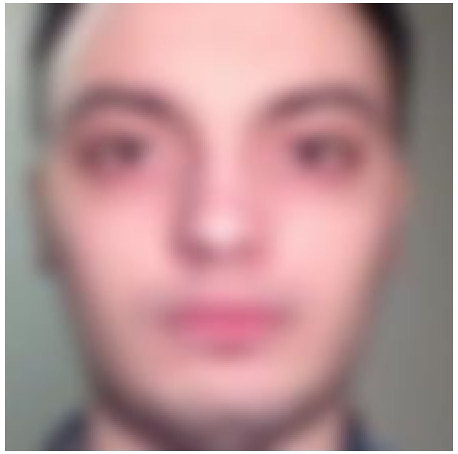
- Numerous face de-identification methods have been developed.
- Ad-hoc face de-identification methods:
  - masks on facial regions;
  - low-pass filtering or random noise addition;
  - swap face sub regions belonging to different individuals;
  - spatial subsampling resulting in facial region pixilation.

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

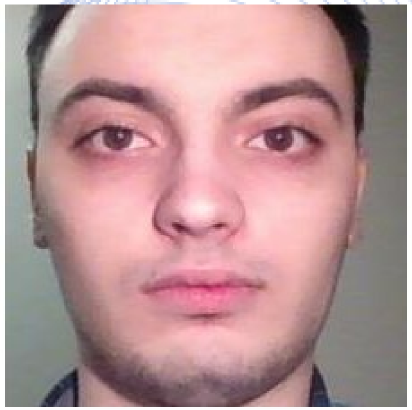
- Modified face reconstruction methods:
  - Reduce the number of eigenfaces used for reconstructing the de-identified facial images.

- Taking advantage of the particularities of specific face identification methods in order to defeat them:
  - blocking efficient feature extraction.

Face de-identification methods

Reconstruction-based face DID approaches:

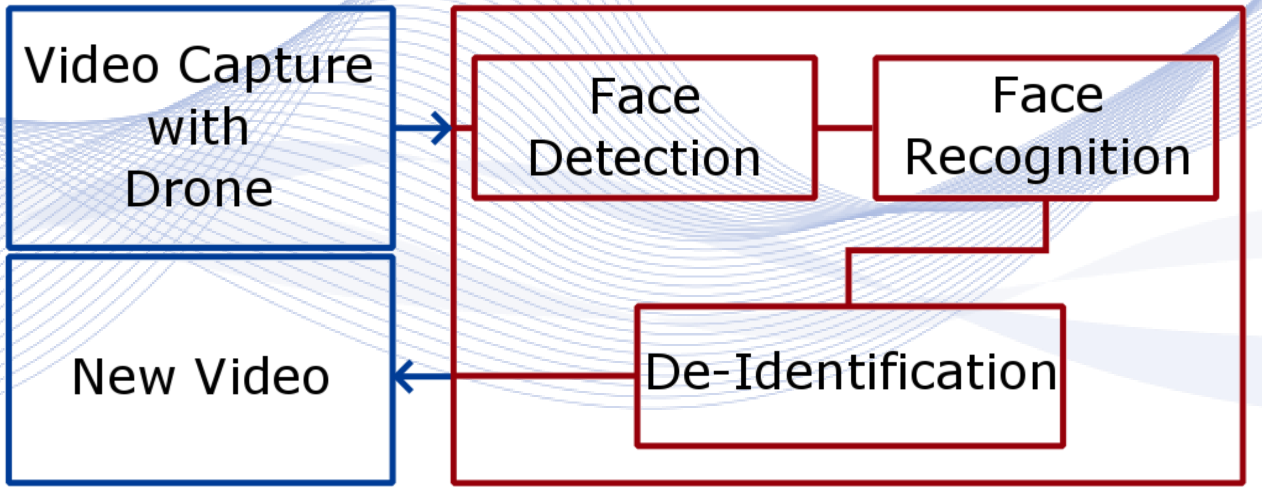
- Obtain facial image coefficients using some reconstruction method (e.g., PCA, SVD, Autoencoder).
- Apply modifications to these coefficients.
- Reconstruct a distorted facial image.



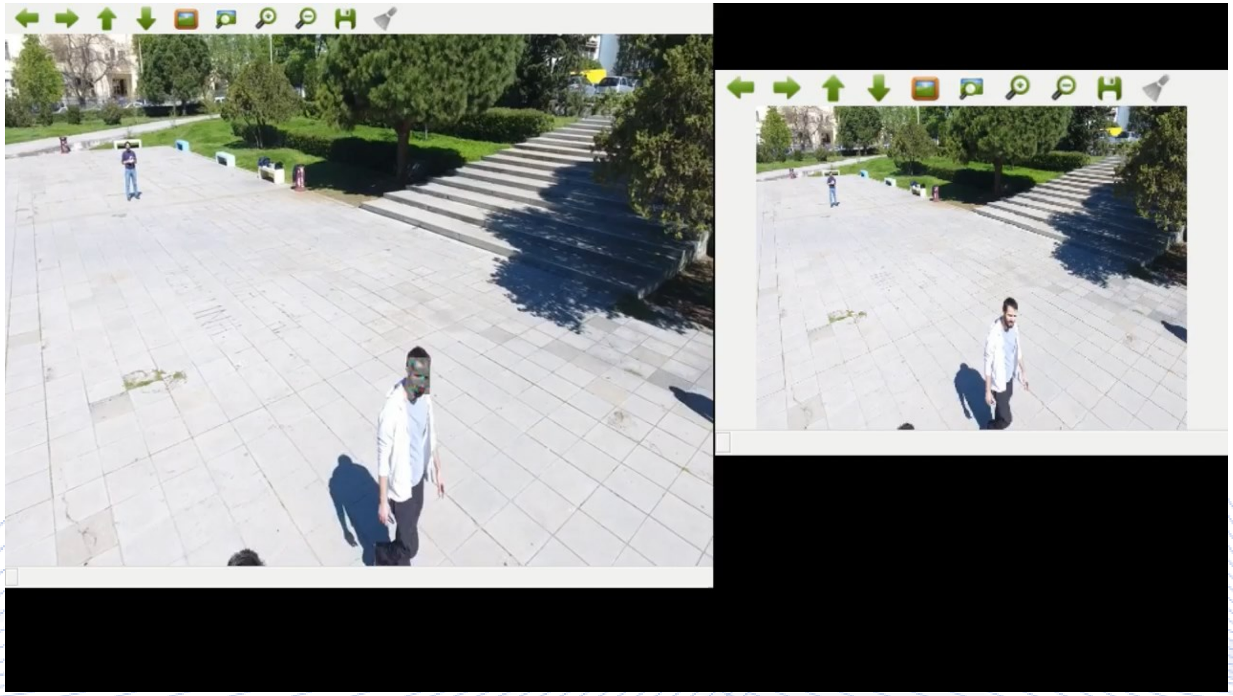
Original Image

Hypersphere projection with radius

Face de-identification on drone videos



Face de-identification on drone videos

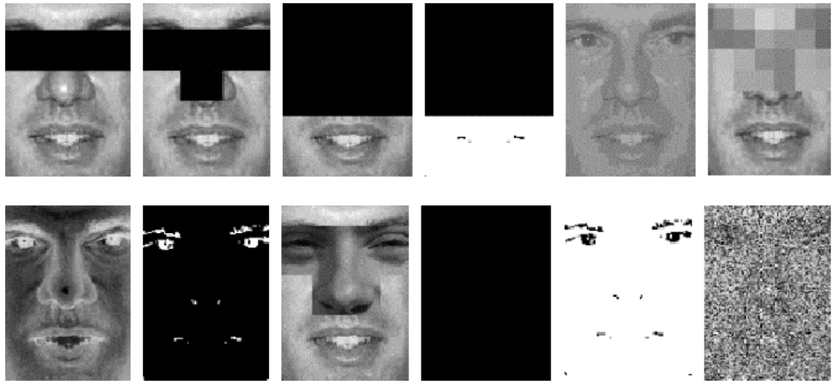


SVD-DID face de-identification in a drone video.

Face de-identification methods

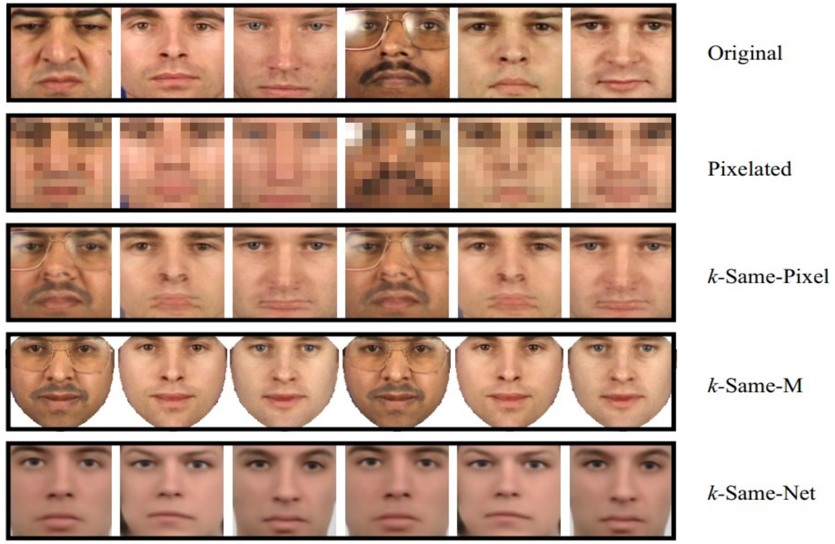
Drawbacks of previous face DID methods:

- 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).



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).

Autoencoder-based Face De-identification

- An image dataset $\mathcal{D} = \{\mathbf{x}_\mathrm{i} \in \mathbb{R}^n, i = 1, \dots, N\}$ is employed to train an autoencoder $\mathbf{x}_\mathrm{p} = \boldsymbol{G}(\mathbf{x}; \boldsymbol{\theta})$, $\mathbf{x}_\mathrm{p} \in \mathbb{R}^n$.
- Let $\mathbf{x}_i$ be a facial image and $\mathbf{z}_i$ be its encoded feature vector, learnt by an autoencoder.
- The reconstruction $\mathbf{x}_\mathrm{p}$ represents a lossy version of the original image, preserving similarity with $\mathbf{x}_i$.
- Information loss is enough to greatly lower face identification accuracy.

Autoencoder-based De-identification

To produce visibly different facial identities, the autoencoder is disintegrated to its encoder and decoder parts, focusing on finetuning the encoder, using the following loss function:

$$J(\mathbf{z}_i, \mathbf{t}_i) = |\mathbf{z}_i - \mathbf{t}_i|_2^2.$$

- $\boldsymbol{t}_i$ is the generic target facial image representation (features).
- Its choice depends on the desired properties to be preserved/discarded on the reconstructed facial image.

Artificial Intelligence & Information Analysis Lab

Supervised Attribute Preserving Face DID

- In order to obtain $\mathbf{t}_i$, we first define the intermediate target $\boldsymbol{s}_i$:

$$\mathbf{s}_i = (1 - a)\mathbf{z}_i + a\mathbf{P}_i\mathbf{Z},$$

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N].$$

- where a is a trade-off parameter.

Supervised Attribute Preserving Face DID

- $\mathbf{P}_i \in [0,1]^{N\times N}$ is an ***attraction matrix*** encoding the data indices that contain desired properties to be preserved (e.g., ethnicity, mood, gender):

$$P_{ij} = \begin{cases} \frac{1}{|\mathcal{D}_i|}, & \text{if } \mathbf{x}_j \in \mathcal{D}_i \\ 0, & \text{otherwise.} \end{cases}$$

- $\mathcal{D}_i$: sets containing related facial images.

Supervised Attribute Preserving Face DID

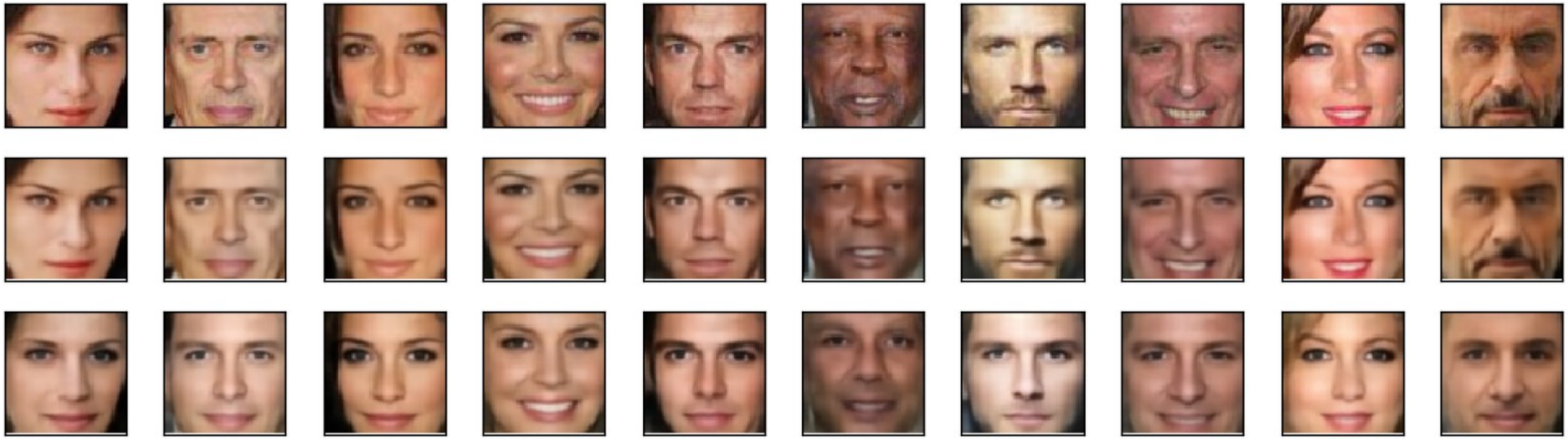
- In a similar fashion, a ***repulsion matrix*** $\mathbf{Q}_i$ is defined:

$$Q_{ij} = \begin{cases} \frac{1}{|\mathcal{U}_i|}, & \text{if } \mathbf{x}_j \in \mathcal{U}_i \\ 0, & \text{otherwise.} \end{cases}$$

- $\mathcal{U}_i$: sets containing opposing facial images.
- $\mathbf{Q}_i$ encodes undesirable properties (e.g., same-person facial images).
- The final reconstruction weight target is defined as follows:

$$\mathbf{t}_i = (1+\beta)\mathbf{s}_i - a\mathbf{Q}_i\mathbf{S}, \qquad \mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_N].$$

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.

Supervised Attribute Preserving Face DID

This method is evaluated in terms of the following performance metrics [NOU2019] :

- Faceness (FCNS), De-identification performance, Output diversity (DIV).

Shift Method	*FCNS*	*DEID*	*DIV*
Ethn+Mood+Mth	**99.85**	92.65	*48.25*
Ethn+Mth+Mood	99.78	92.56	48.96
Mood+Ethn+Mth	99.71	92.74	39.60
Mood+Mth+Ethn	*99.76*	*92.94*	**51.28**
Mth+Mood+Ethn	99.52	93.02	41.08
Mth+Ethn+Mood	99.62	**93.22**	34.00

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.

GAN-based face de-identification

Live face de-identification in video [GAF2019].

Conditional Identity Anonymization GAN results [MAX2020].

GAN body image de-identification

- Generative adversarial networks attempting to generate synthetic body image samples from the distribution of all possible body images that were generated from true segmented body images.
- Synthetic body images should be de-identified ones.
- Extending face de-identification.
- It removes soft biometric (e.g., tattoos) and non-biometric identifiers (e.g., cloth color).

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.

Targeted adversarial attacks

For a given image $\mathbf{x} \in \mathbb{R}^n$, target label $t \in \mathcal{C} - \{y\}$, targeted adversarial attacks solve the following box-constrained optimization problem:

$$\text{Minimize } \|\mathbf{p}\|_2$$

subject to: $f(\mathbf{x}_p; \boldsymbol{\theta}) = t$ and $\mathbf{x}_p = \mathbf{x} + \mathbf{p} \in \mathbb{R}^n$.

An additional stopping condition of this optimization problem could be just:

$$f(\mathbf{x}_p; \boldsymbol{\theta}) \neq y.$$

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.

Artificial Intelligence & Information Analysis Lab

Adversarial Face De-Identification

- Alternative update equation of the I-FGSM:

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

- $\text{sign}(\cdot)$ function returns the sign of a real number.

Adversarial Face De-Identification

P-FGVM face de-identification method:

- Another face de-identification method based on adversarial samples.
- Penalized Fast Gradient Value Method (P-FGVM).
- Inspired by the adversarial attack method I-FGVM.
- It combines an adversarial loss term and a ‘realism’ loss term in the objective function.

Adversarial Face De-Identification

- Gradient descent update equations of the P-FGVM:

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

- λ is a weight coefficient for the proposed "realism term" $\mathbf{x}_p^i - \mathbf{x}$.
- It pushes the solution of the optimization problem towards images $\mathbf{x}_p$ that lie close to the original image $\mathbf{x}$, in terms of distance.

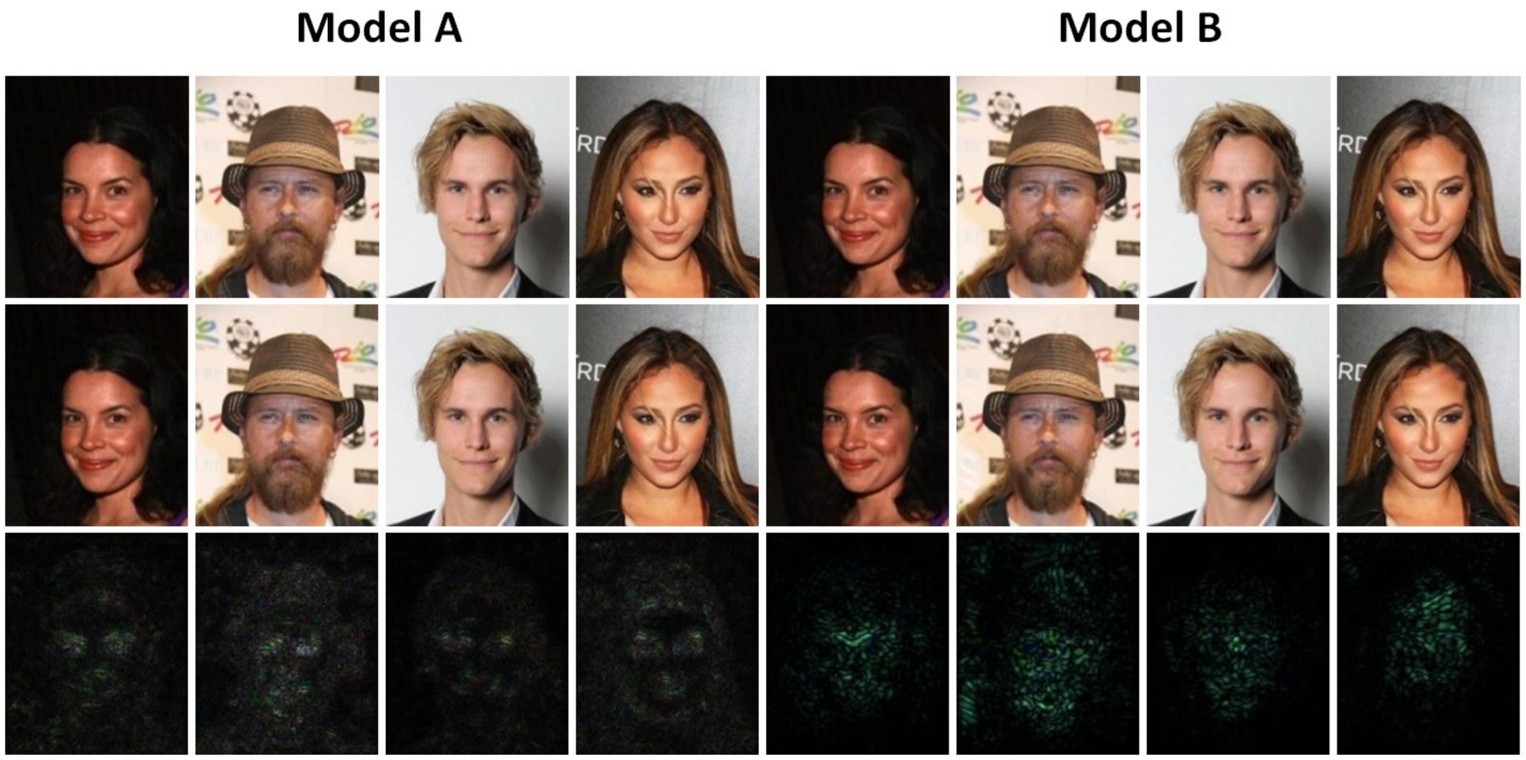
Adversarial Face De-Identification

Comparison of P-FGVM with I-FGVM and Iterative Fast Gradient Sign Method I-FGSM face DID methods.

- Performance metrics:
  - L2 Distance (L2), CW-SSIM (SI) and de-identification performance (MR).

Model A			Model B		
L2	SI	MR	L2	SI	MR
Experimental Results					
P-FGVM					
3.38	0.986	99.6%	2.11	0.995	96.0%
I-FGVM					
5.31	0.963	99.4%	2.67	0.993	93.2%
I-FGSM					
5.68	0.962	98.9%	5.74	0.968	94.4%
Percentage Improvement					
I-FGVM					
36.3%	2.3%	0.2%	20.9%	0.2%	3.0%
I-FGSM					
40.4%	2.4%	0.7%	63.2%	2.7%	1.7%

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-Anonymity-inspired adversarial attack

- Replacing the initial face with a face from another person.
- k-anonymity model:
  - de-identified images can be misclassified as belonging to at least k original individuals;
  - recognition rates are guaranteed to be lower than $1/k$.
- The core problem of k-same de-identification is to find the optimal selection of faces from the original face set consisting of $\mathcal{C} = \{C_1, \dots, C_m\}$ facial classes ($m \gg k$) to form the clusters of k faces.

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

A NN classifier label output $\hat{y}=f(\mathbf{x};\boldsymbol{\theta})$ is usually produced by finding the arguments of the maxima of the final layer:

$$f(\mathbf{x};\boldsymbol{\theta}) = argmax(\boldsymbol{g}(\mathbf{x};\boldsymbol{\theta})),$$

- $\boldsymbol{g}(\mathbf{x};\boldsymbol{\theta}) \in \mathbb{R}^m$ contains the network output values corresponding to the number of classes supported by the model.
- It has been observed that adversarial samples are usually classified correctly, only by obtaining the 2nd maximum ranking position instead of the 1st.
- Thus, network activations $\boldsymbol{g}(\mathbf{x};\boldsymbol{\theta})$ for a sample $\mathbf{x}$ in the final layer may in fact act as Quasi-Identifiers, along with the output label $\hat{y}$.

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

Let $r_{\mathbf{x}}(i) \in \mathcal{C},\ i = 1, \dots, m$, be a function associated with $\boldsymbol{g}(\mathbf{x}; \boldsymbol{\theta})$, outputting the $i-th$ most probable label of sample $\boldsymbol{x}$, ranked as follows:

$$r_{\mathbf{x}}(1) = argmax\left(\boldsymbol{g}(\mathbf{x}; \boldsymbol{\theta})\right) = f(\mathbf{x}; \boldsymbol{\theta}).$$

- For every adversarial sample in a dataset, we demand that:

$$r_{\mathbf{x}_p}(1) \neq y,$$

$$\mathrm{P}\left(r_{\mathbf{x}_p}(i) = y\right) \leq \frac{1}{k}, \qquad i = 1, \dots, m.$$

- The first term is the adversarial attack constraint.
- In the second term, $P(\cdot)$ is a probability function and k denotes the desirable "k-anonymity protection level" property for sample $\mathbf{x}$.

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

The solution can be obtained by solving the following optimization problem:

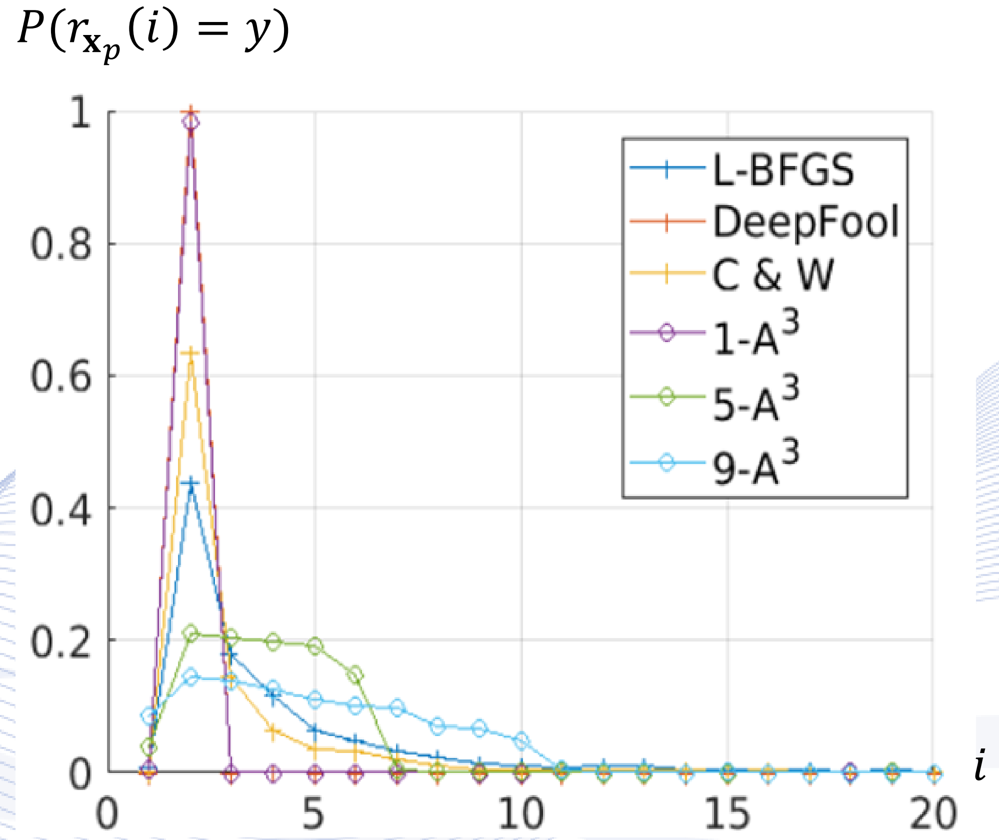
$$\min_{\mathbf{p}} ||\mathbf{p}||_2 + \left(d - s\left(\mathbf{x} - \mathbf{x}_p\right)\right) + \sum_{i=2}^{k} J_i \left(f\left(\mathbf{x}_p; \boldsymbol{\theta}\right), \mathbf{t}_i\right)$$

- J_i is a classification loss function.
- $\mathbf{t}_i$ is the target output vector class corresponding to output label $r_{\boldsymbol{x}}(i)$.
- $s(\cdot) < d$ is a similarity cost function for regularization purposes (related to SSIM).

This method extends the standard targeted Adversarial Attack optimization problem towards perturbations using k different classes in the dataset, instead of just 1.

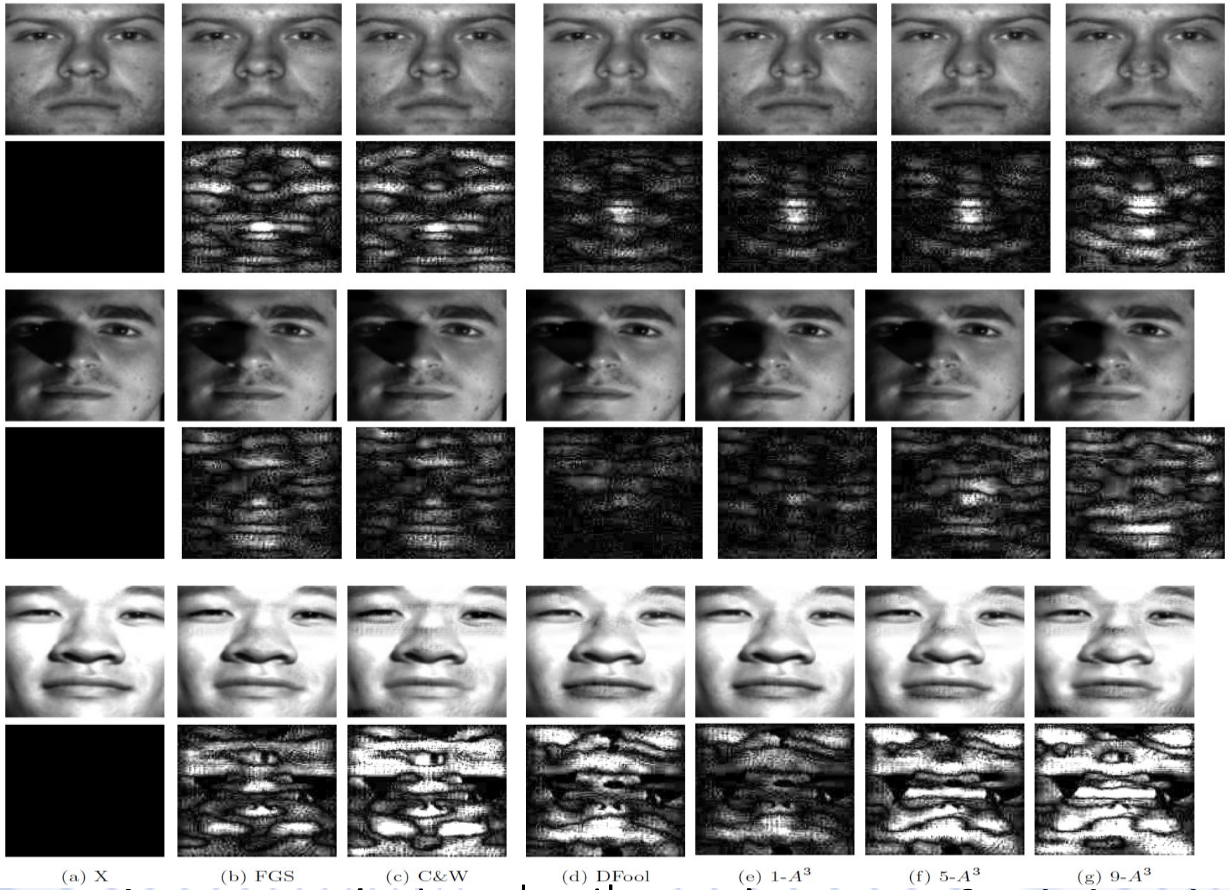
$k - A^3$ face de-identification method

- "Adversarial" datasets were created for each adversarial attack method, using 3 SoA methods and the proposed one.
- In most of the cases, the 2nd sorted ranked activations contain the "true" label of the adversarial samples.
- Only the $k - A^3$ method for $k = 5$, $k = 9$, satisfies the k-Anonymity Requirements.



Probability of obtaining the true face label, using the sorted ranked activations of the final layer.

$k - A^3$ 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

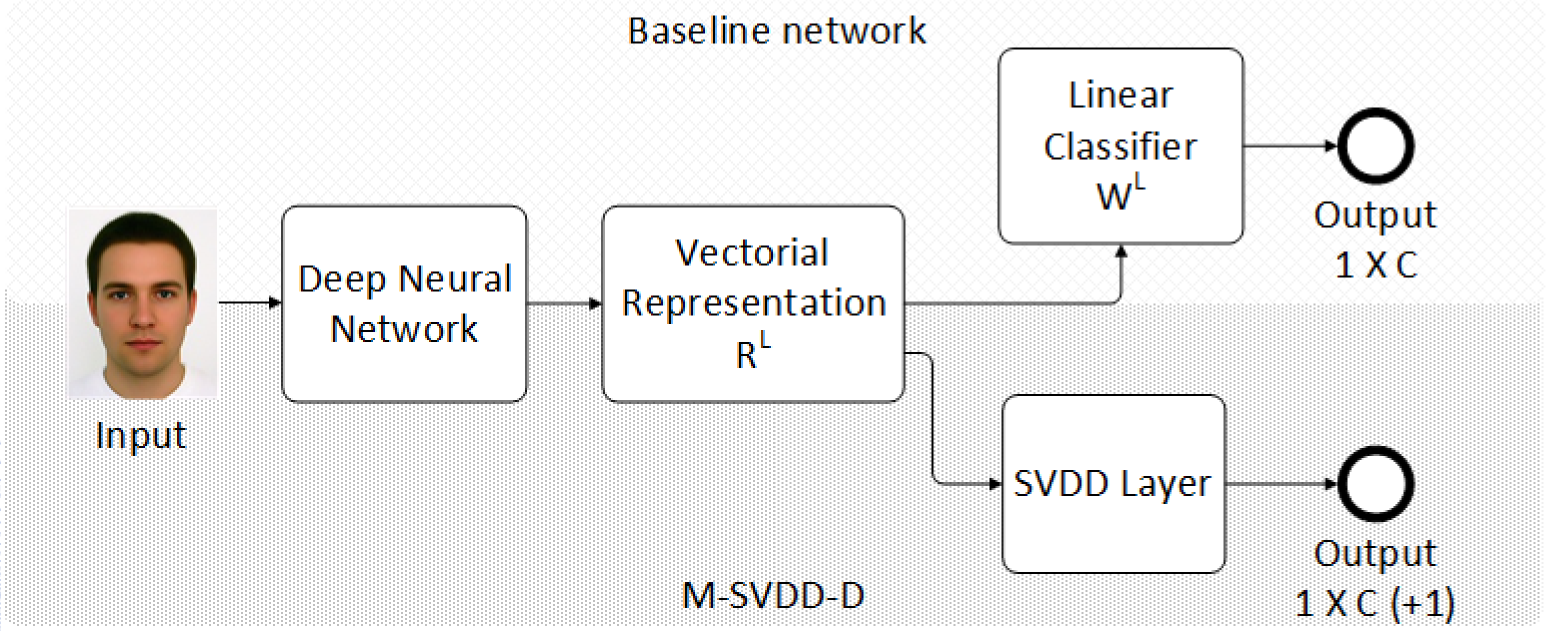
- Privacy and data protection
- Classical face de-identification
- Autoencoder-based Face De-identification
- GAN-based de-identification
- Adversarial face de-identification
- K-anonymity attacks
- **SVDD Adversarial Defense**

Adversarial Defense based on SVDD

- To protect face recognition NNs against adversarial attacks, we replace the NN classification layer with by m non-linear one-class classifiers (SVDD).
- We introduce the concept of ***minimum activation value*** $(T > 0)$, acting as an additional class ($m + 1$ class).
- Thus, the framework classifies $m + 1$ face classes, where m are the classes associated with one-class classifiers, and $m + 1$ is the adversarial class, using the following rule:

If $\boldsymbol{g}(\mathbf{x}; \boldsymbol{\theta}) < T\mathbf{1}$ for all m SVDD classifiers, then $\mathbf{x}$ is an adversarial example.

m-SVDD Adversarial Defense



m-SVDD Adversarial Defense

The effectiveness and noise required to fool the face recognition models before and after applying the proposed defense have been studied.

- The least noise is generated by the proposed $k - A^3$ method.
- SVDD defense methodology increases the robustness of the model.

m-SVDD Adversarial Defense

Experiment	Yale-LightCNN						
	Undefended			Defended			
Method/Metric	SSIM	MSE×10^4	ASR%	SSIM	MSE×10^4	ASR%, a=1	ASR% , a=0.97
L-BFGS	93.89	5.97	99.23	93.30	6.26	43.44 (F =19.36, D=37.19)	10.96 (F=8.81, D=80.22)
DeepFool	97.87	1.71	100	97.77	1.94	41.49 (F=47.43, D=11.06)	17.36 (F=37.85, D=44.77)
C & W	94.21	5.16	99.94	92.82	5.59	29.91 (F=24.64, D=45.44)	06.91 (F=12.70, D=80.37)
1-A^3	98.05	1.59	99.43	98.06	1.63	41.59 (F=48.82, D=9.57)	18.08 (F=38.11, D=43.80)
5-A^3	95.17	7.52	96.26	94.55	7.74	42.82 (F=18.80, D=38.37)	11.52(F= 12.09, D=76.38)
9-A^3	93.17	4.36	91.34	92.52	5.98	29.50 (F=15.47, D=55.02)	06.40(F=8.65, D=84.93)

Effectiveness and noise required to fool the face recognition models. Performance metrics: Attack Success Rate (ASR), (F: Failed Attacks, D: Detected Attacks). a is defense parameter [MYG2020].

Adversarial face DID Motivations

- Adversarial attacks minimally intervene with the original data, focusing only against automated analysis.
- Up to date, they are imperceptible by humans.
- It is a great tool for examining robustness of neural networks.
- They have the potential of fooling multiple neural networks.
- They expose AI weaknesses in critical applications, e.g., biometric identifiers, traffic sign classification.

Adversarial face DID Limitations

- A “host” pre-trained network is required to generate adversarial perturbations.
- Adversarial attacks are network specific, come with no guarantees and cannot be applied universally.
  - However, some attacks may be transferable between different architectures.
- Adversarial attacks provide no privacy protection against humans, thus will never become GDPR compliant.
- If the DNN ***adversarial robustness*** problem is solved in the future, adversarial attacks will not even work, at least not that well.
  - They will generate too much noise or will fail completely.

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