Project Detective V2

Deepfake **E**vidence **T**racking and **E**valuation for **C**ontent **T**ransparency and **I**ntegrity **V**erification **E**fforts

Fondamenti di Visione Artificiale e Biometria

Anno accademico 2024/2025

Prof. Michele Nappi
Tutor Dott. Matteo Polsinelli

Che cos’è il DeepFake

- La manipolazione di immagini e video è una pratica esistente da diversi anni.

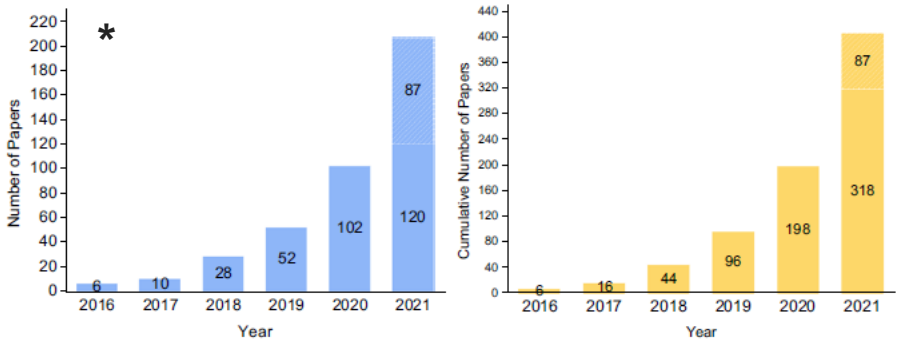
- Tuttavia, l’utilizzo dell'intelligenza artificiale introduce un nuovo aspetto: **il realismo**. I contenuti multimediali sintetici generati sono convincenti perché verosimili.

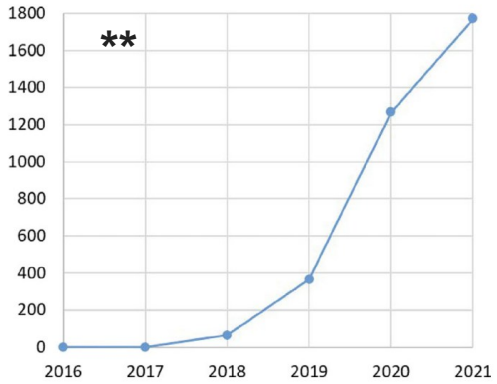
- I nuovi metodi di creazione di informazione possono sicuramente essere utilizzata per nobili scopi: migliorare l’offerta formativa, creare nuove forme di intrattenimento o migliorare quelle già esistenti etc.

- Allo stesso tempo, possono essere utilizzare, ad esempio, per creare frodi o fake news in quanto le persone possono essere facilmente ingannate.

Motivazioni del Progetto

1) Grande interesse da parte della comunità scientifica





2) Riscontro pratico nella vita di tutti i giorni, per via delle implicazioni pratiche

* Juefei-Xu, F. et al., 2022. Countering malicious deepfakes: Survey, battleground, and horizon. *International journal of computer vision*, *130*(7), pp.1678-1734.

** Nguyen, T.T., et al. 2022. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, p.103525.

Conosci il tuo avversario: il «falsario»

1. Vuole generare immagini con il minimo contenuto di artefatti nel dominio spaziale.

2. Non vuole lasciare **impronte** di qualsiasi tipo nell'immagine generata che siano la prova che si tratta di un fake. Se è costretto a lasciare delle impronte, vuole cancellarle in una fase di post-processing.

Falsificatore: strumenti utilizzati

- Tra i metodi più utilizzati per la generazione di DeepFake, come suggerisce la parola stessa, ci sono quelli di Deep Learning.

- **In particolare le GAN hanno dimostrato di essere tra i metodi che meglio soddisfano il primo requisito del falsificatore: minimo contenuto di artefatti nel dominio spaziale.***

- Inoltre, questi metodi sono facilmente reperibili online e pronti all'uso. *

Summary of notable deepfake tools.

Tools	Links	Key features
Faceswap	https://github.com/deepfakes/faceswap	- Using two encoder–decoder pairs. - Parameters of the encoder are shared.
Faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN	Adversarial loss and perceptual loss (VGGface) are added to an auto-encoder architecture.
Few-Shot Face Translation	https://github.com/shaoanlu/fewshot-face-translation-GAN	- Use a pre-trained face recognition model to extract latent embeddings for GAN processing. - Incorporate semantic priors obtained by modules from FUNIT (Liu et al., 2019) and SPADE (Park et al., 2019).
DeepFaceLab	https://github.com/iperov/DeepFaceLab	- Expand from the Faceswap method with new models, e.g. H64, H128, LIAEF128, SAE (DeepFaceLab, 2022a). - Support multiple face extraction modes, e.g. S3FD, MTCNN, dlib, or manual (DeepFaceLab, 2022a).
DFaker	https://github.com/dfaker/df	- DSSIM loss function (DSSIM, 2022) is used to reconstruct face. - Implemented based on Keras library.
DeepFake_tf	https://github.com/StromWine/DeepFake_tf	Similar to DFaker but implemented based on tensorflow.
AvatarMe	https://github.com/lattas/AvatarMe	- Reconstruct 3D faces from arbitrary "in-the-wild" images. - Can reconstruct authentic 4K by 6K-resolution 3D faces from a single low-resolution image (Lattas et al., 2020).
MarioNETte	https://hyperconnect.github.io/MarioNETte	- A few-shot face reenactment framework that preserves the target identity. - No additional fine-tuning phase is needed for identity adaptation (Ha et al., 2020).
DiscoFaceGAN	https://github.com/microsoft/DiscoFaceGAN	- Generate face images of virtual people with independent latent variables of identity, expression, pose, and illumination. - Embed 3D priors into adversarial learning (Deng et al., 2020).
StyleRig	https://gvv.mpi-inf.mpg.de/projects/StyleRig	- Create portrait images of faces with a rig-like control over a pretrained and fixed StyleGAN via 3D morphable face models. - Self-supervised without manual annotations (Tewari et al., 2020).
FaceShifter	https://lingzhili.com/FaceShifterPage	- Face swapping in high-fidelity by exploiting and integrating the target attributes. - Can be applied to any new face pairs without requiring subject specific training (Li et al., 2019a).
FSGAN	https://github.com/YuvalNirkin/fsgan	- A face swapping and reenactment model that can be applied to pairs of faces without requiring training on those faces. - Adjust to both pose and expression variations (Nirkin et al., 2019).
StyleGAN	https://github.com/NVlabs/stylegan	- A new generator architecture for GANs is proposed based on style transfer literature. - The new architecture leads to automatic, unsupervised separation of high-level attributes and enables intuitive, scale-specific control of the synthesis of images (Karras et al., 2019).
Face2Face	https://justusthies.github.io/posts/face2face/	- Real-time facial reenactment of monocular target video sequence, e.g. Youtube video. - Animate the facial expressions of the target video by a source actor and re-render the manipulated output video in a photo-realistic fashion (Thies et al., 2016).
Neural Textures	https://github.com/SSRSGJYD/NeuralTexture	- Feature maps that are learned as part of the scene capture process and stored as maps on top of 3D mesh proxies. - Can coherently re-render or manipulate existing video content in both static and dynamic environments at real-time rates (Thies et al., 2019).
Transformable Bottleneck Networks	https://github.com/kyleolsz/TB-Networks	- A method for fine-grained 3D manipulation of image content. - Apply spatial transformations in CNN models using a transformable bottleneck framework (Olszewski et al., 2019).
"Do as I Do" Motion Transfer	https://github.com/carolineec/EverybodyDanceNow	- Automatically transfer the motion from a source to a target person by learning a video-to-video translation. - Can create a motion-synchronized dancing video with multiple subjects (Chan et al., 2019).
Neural Voice Puppetry	https://justusthies.github.io/posts/neural-voice-puppetry	- A method for audio-driven facial video synthesis. - Synthesize videos of a talking head from an audio sequence of another person using 3D face representation. (Thies et al., 2020).

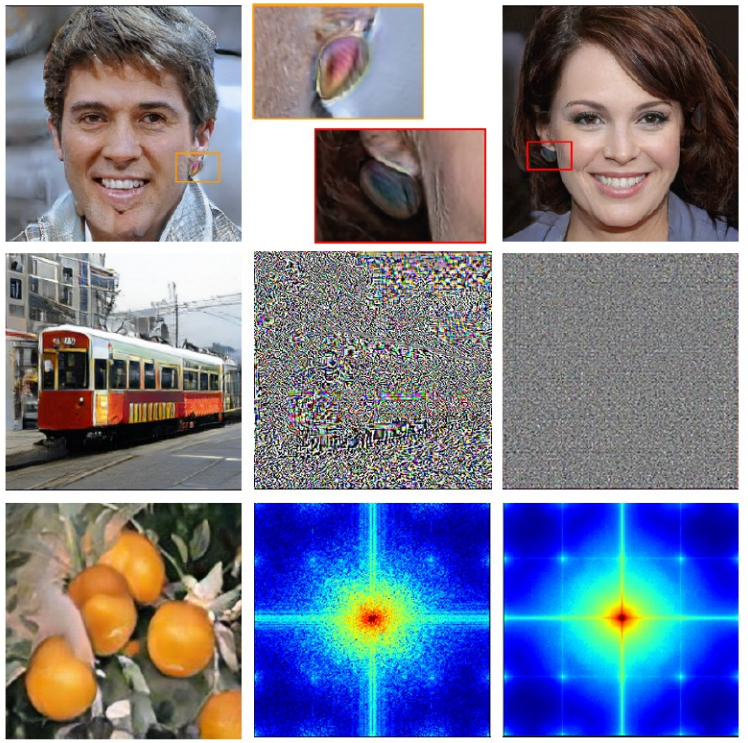
* Nguyen, T.T., et al. 2022. Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, p.103525.

GAN fingerprint

Le GAN aggiungono una fingerprint ben visibile (generalmente da un metodo automatico) sia nel dominio spaziale che nel dominio delle frequenze: **il secondo requisito del falsificatore non è rispettato *.**

Esempi nel dominio spaziale

Esempi nel dominio delle frequenze



Anomalie nel colore o mancaza di simmetrie. Sono artefatti visibili nel dominio spaziale ma stanno via via scomparendo

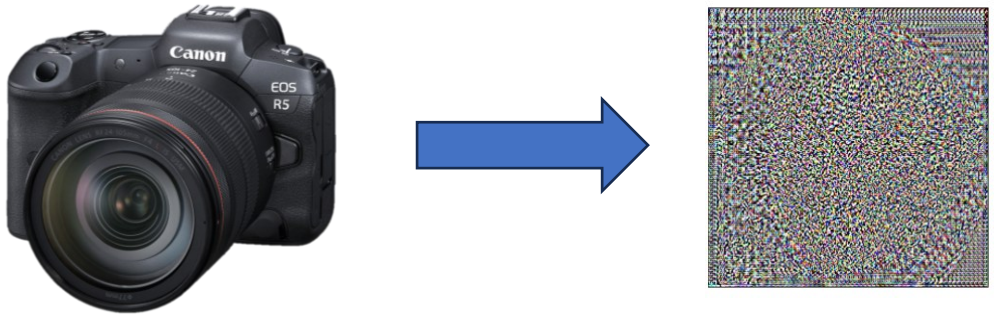
* Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G. and Verdoliva, L., 2021, July. Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. In 2021 IEEE international conference on multimedia and expo (ICME) (pp. 1-6). IEEE.

Visualizzare la fingerprint nel dominio spaziale con tecniche forensi già note

- Ciascun dispositivo di acquisizione di immagini, a causa di imperfezioni di produzione, lascia un'impronta unica e stabile su ciascuna foto acquisita, nota come photo-response nonuniformity (PRNU) pattern

- Il PRNU può essere stimato ed e trattato come una vera e propria device fingerprint

- Tale fingerprint può essere autilizzata per attribuire un immagine ad un particolare device, riconoscere e localizzare eventuali manipolazioni ed è uno dei metodi più utilizzati nell'analisi forense.



Stima della fingerprint nel dominio spaziale *

1) $\widehat{X}_i = f(X_i)$

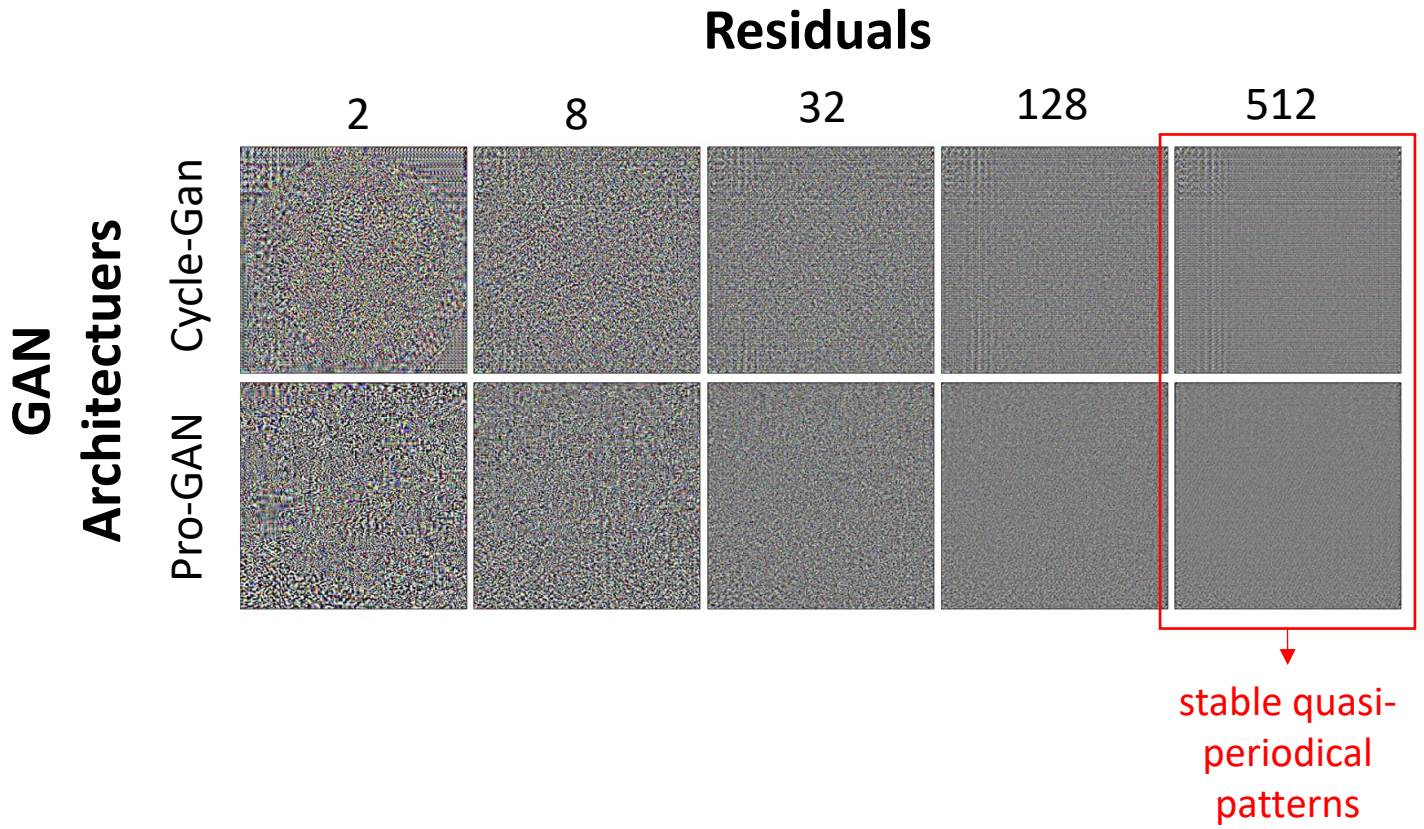
2) $R_i = X_i - f(X_i)$

3) $R_i = F + W_i$

non-zero deterministic component

random noise component

4) $\widehat{F} = \frac{1}{N}\sum_{i=1}^{N} R_i$



* Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019, March). Do gans leave artificial fingerprints?. In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 506-511). IEEE.

Visualizzare la Fingerprint nel Dominio Delle Frequenze (1)

- Previous work has already linked upsampling operations (deconvolution o transposed-convolution) to causing grid-like patterns in the image domain *

- Recognizing this, the architecture of both the generator-network and the discriminator-network shifted from using strided transposed convolution (DCGAN, CramerGAN, CycleGAN, MMDGAN, and SNDCGAN to using traditional upsampling methods—like nearest neighbor or bilinear upsampling followed by a convolutional layer (ProGAN, BigGAN, and StyleGAN)

- **While these changes addressed the problem in the spatial domain, the results show that the artifacts are still detectable in the frequency domain** **

*Odena, A., Dumoulin, V. and Olah, C., 2016. Deconvolution and checkerboard artifacts. *Distill*, *1*(10), p.e3.

**Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D. and Holz, T., 2020, November. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning* (pp. 3247-3258). PMLR.

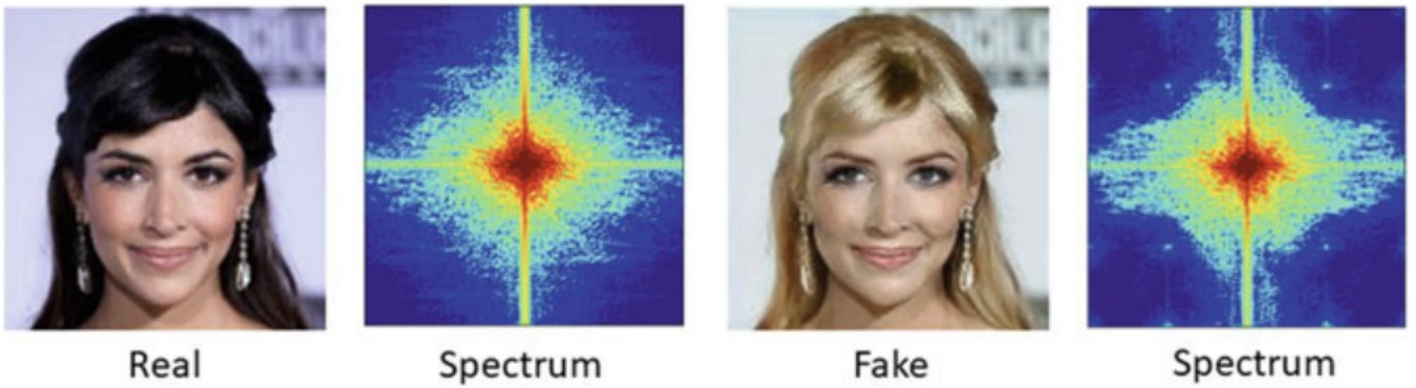
Visualizzare la Fingerprint nel Dominio Delle Frequenze (2)

- Il motivo principale del successo di questi metodi sta nel fatto che le fingerprint sono “direttamente” ben visibili nel domino delle frequenze e quindi è più semplice fare la detection

- Le anomalie che si riscontrano nel dominio delle frequenze sono di due tipi: “patterns anormali” e discrepanze nella Power Distribution.

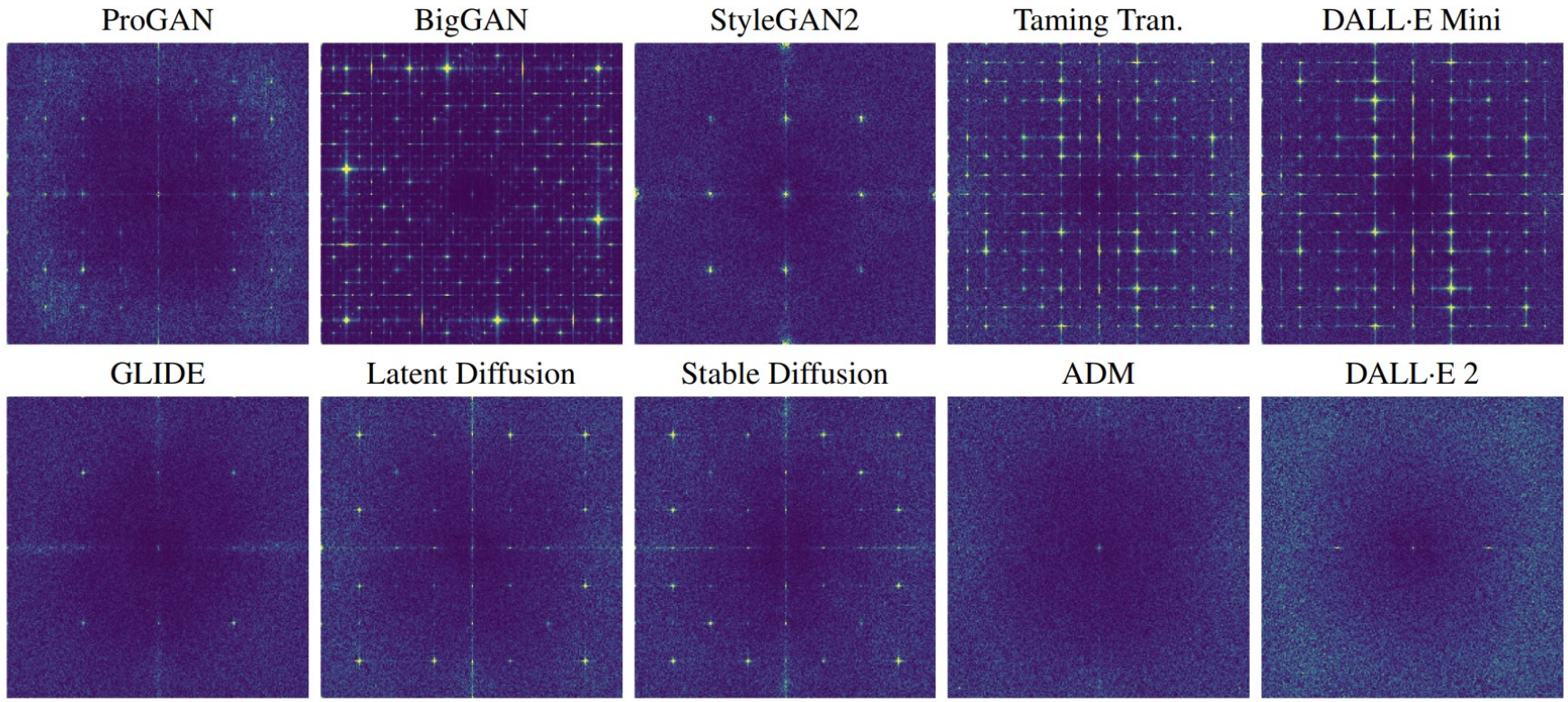
Visualizzare la Finger nel Dominio Delle Frequenze (3)

- Esempi di pattern anormali, come ad esempio linee e punti, sono più frequenti nello spettro delle immagini generate da CycleGAN [47] and StarGAN [13].



- Nello spettro delle frequence delle immagini generate con la BigGAN [10], si riscontra nelle un blurry delle alte frequenze.

E i metodi basati su latent diffusion e stable diffusion?

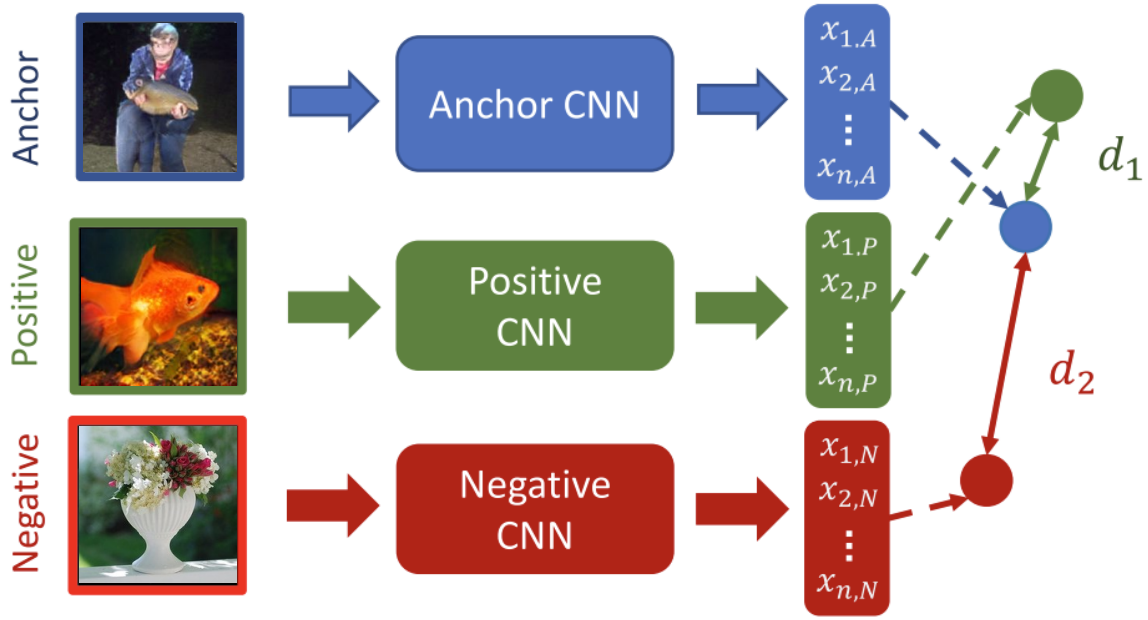


- Considerazioni simili valgono anche per questi nuovi strumenti, anche se nelle reti più evoluto come DALLE i pattern della fingerprint sono notevolmente più difficili da visualizzare *

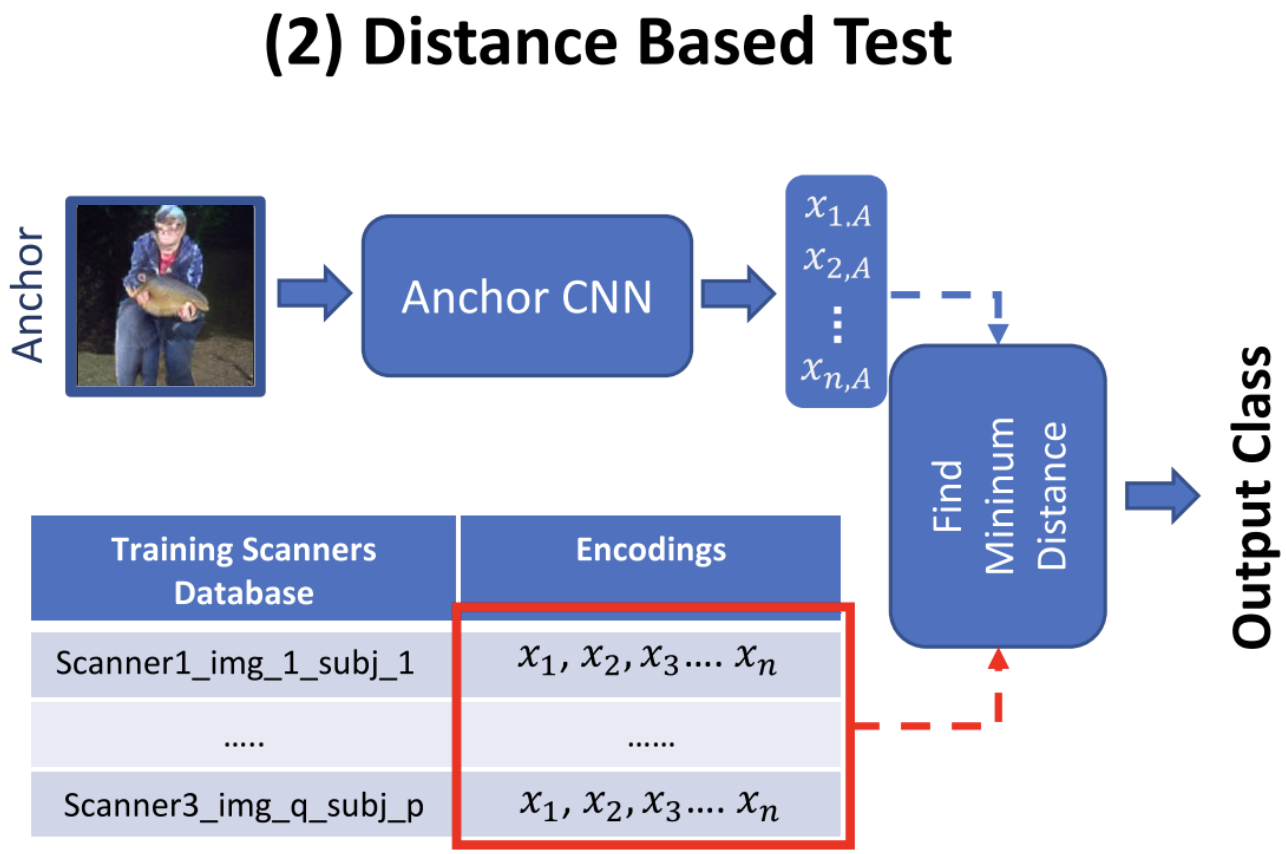
* Corvi, Riccardo, et al. "On the detection of synthetic images generated by diffusion models." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

Siamese Neural Network (1)

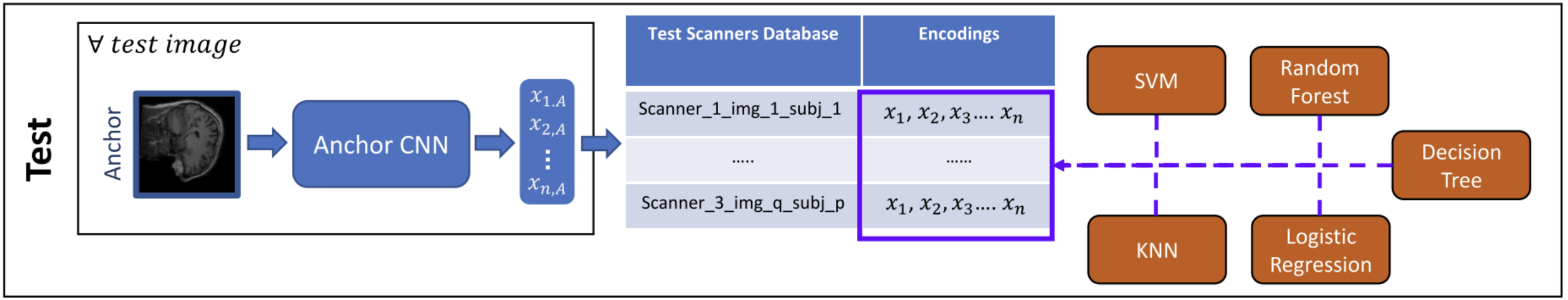
Think about the differences



Siamese Neural Network (2)



Siamese Neural Network (3)



Nuovi obiettivi in Detective V2

- Ci focalizzeremo su 3 nuovi punti fondamentali:

1. Strategie avanzate per la scelta delle triplette;
2. Strategie avanzate per la classificazione dei vettori di embeddings;
3. Nuovi modelli DL (ad esempio i Transformers) per la costruzione dell’architettura SNN.

Detectives: strumenti

- A tutti i gruppi che sceglieranno questo progetto, verrà fornito un Dataset di immagini Real e Fakes così strutturato:
    1. Immagini di Training
    2. Immagini di Test

- Tutti i gruppi si cimenteranno con le stesse immagini e i loro scripts verranno valutati in base a diverse metriche (Accuracy, Specificity, Sensitivity etc).

- Gli script dovranno essere sviluppati in liguaggio Python utilizzando il framework Pytorch.

Quali sono gli elementi per la valutazione progetto?

1. **Codice sorgente:** deve essere leggibile, commentato e ben organizzato;
2. **Lista dei requirements:** bisogna utilizzare conda per la gestione del progetto, come visto a lezione;
3. **Il progetto deve funzionare su qualsiasi PC:** il progetto deve avere un main eseguibile che eventualmente accetti parametri da riga di comando;
4. **Report del progetto:** Una documentazione esaustiva, redatta nello stile di un articolo scientifico, che descriva dettagliatamente i tentativi effettuati e i risultati conseguiti.
5. **Logica di sviluppo:** non si ottengono i risultati sperati…non è un problema! L'importante è aver lavorato bene, seguendo un filo logico di prove successive.

Grazie per l’attenzione

Per ulteriori informazioni: mpolsinelli@unisa.it