Computer vision forensics and security: deepfake detection

Vision and Perception 2022-2023

Luca Maiano
maiano@diag.uniroma1.it

ALCOR Lab

Overview

1. What are deepfakes?
2. Why is deepfake recognition possible?
3. Current limitations
4. Toward more general approaches
5. Our research at the Alcor Lab

Which faces are fake?

A

B

C

D

Which faces are fake?

A

B

C

D

B and C

Let’s try again… which faces are fake?

A

B

C

D

Let’s try again… which faces are fake?

A

B

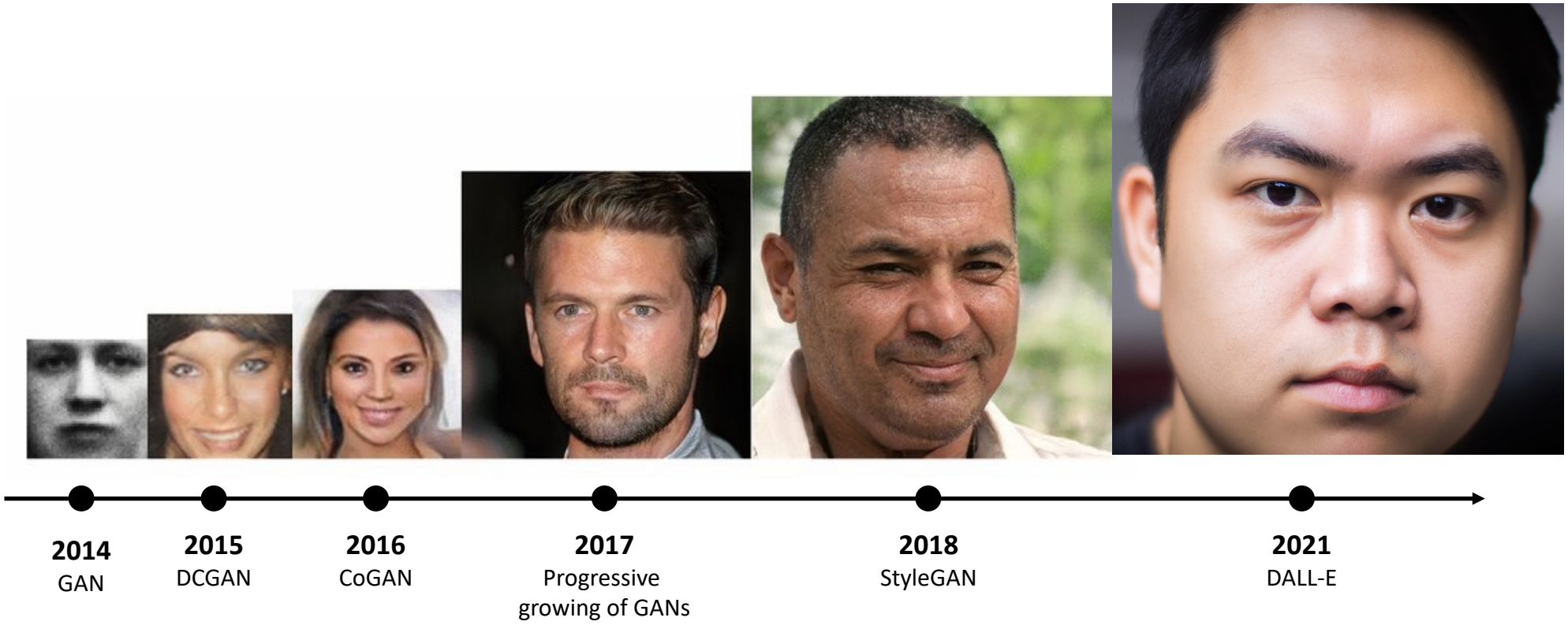
C

D

All of them!

https://www.youtube.com/watch?v=cQ54GDm1eL0&t=4s&pp=ygUOb2JhbWEgZGVlcGZha2U%3D

Deepfakes over years




https://www.midjourney.com/

What are deepfakes?

What are deepfakes?

- Face manipulations in images/videos
- Photoshopped images
- All manipulated images/videos
- Everything in computer graphics
- Any synthetic media: images, videos, text, audio…
- Anything where AI is used ("*Deep*" from Deep Learning)
- …

Wikipedia: "*Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness*."

A general definition here

Manipulated images and videos of faces

The “real” deepfake

reddit

The term was introduced in 2017 when a Reddit user of the same name posted doctored porn clips on the site

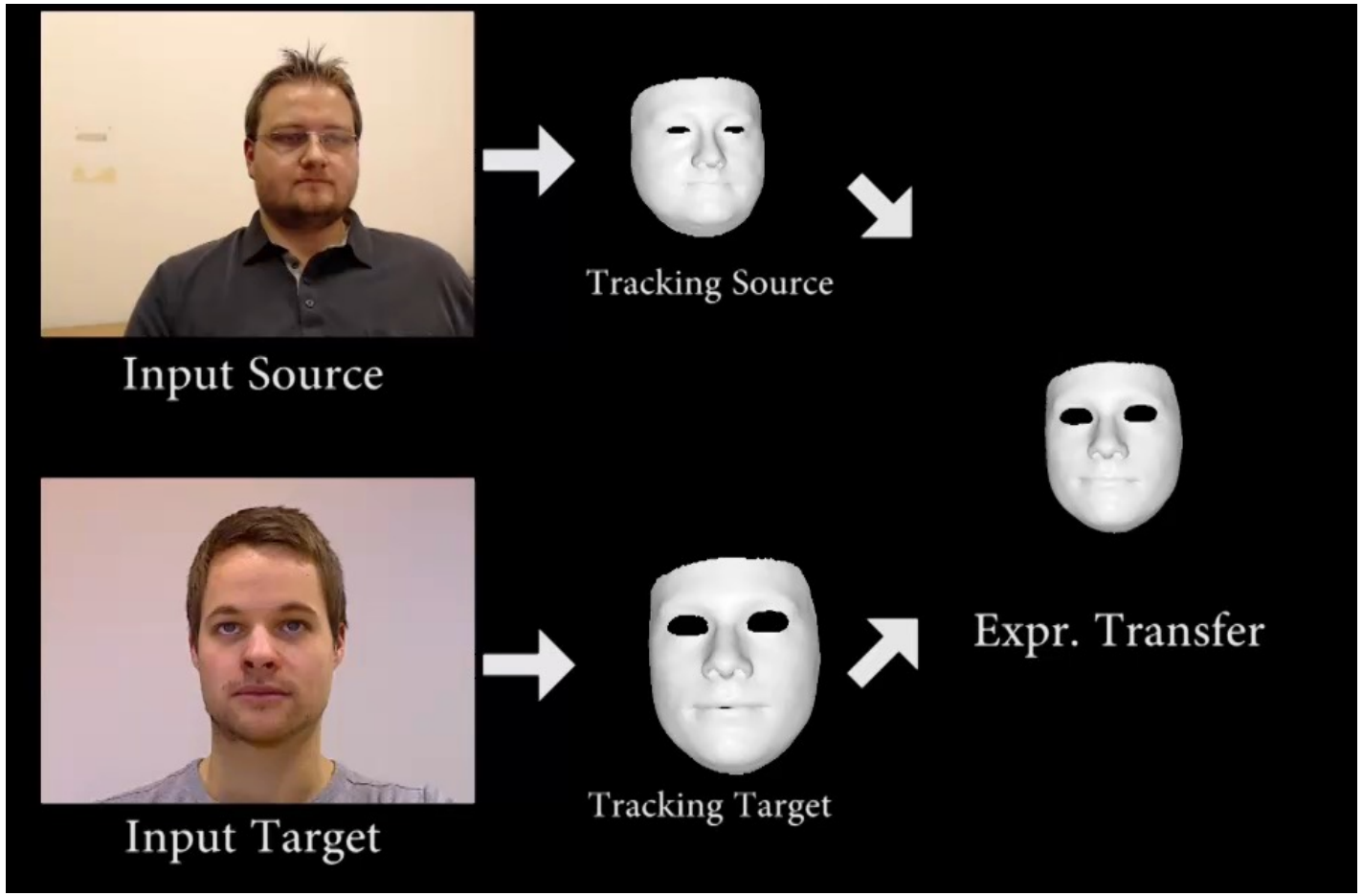
https://github.com/deepfakes/faceswap

Two ways to generate deepfakes

Automatic face manipulations can be split in two main categories:

- **facial reenactment** alters the facial expression preserving the identity
- **face-swapping** modifies the identity of a person preserving the facial expression

Reenactment: an example



Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. and Nießner, M., 2016. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2387-2395)

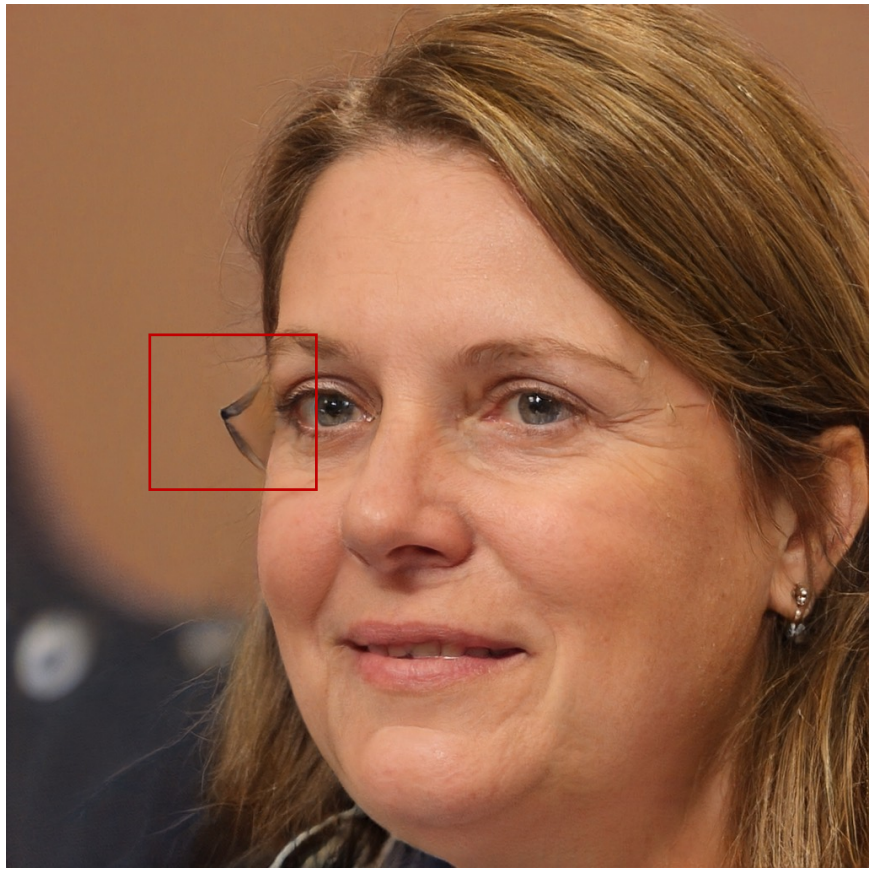
Why is deepfake recognition possible?

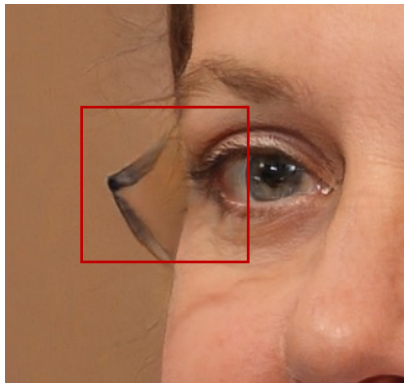
Why is deepfake recognition possible?

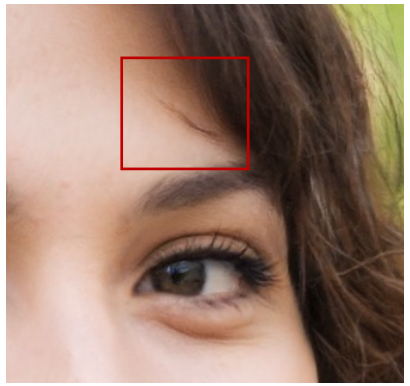
- Visual artifacts
- Semantic inconsistencies
- Identity-related inconsistencies
- Camera-related artifacts
- Model fingerprints

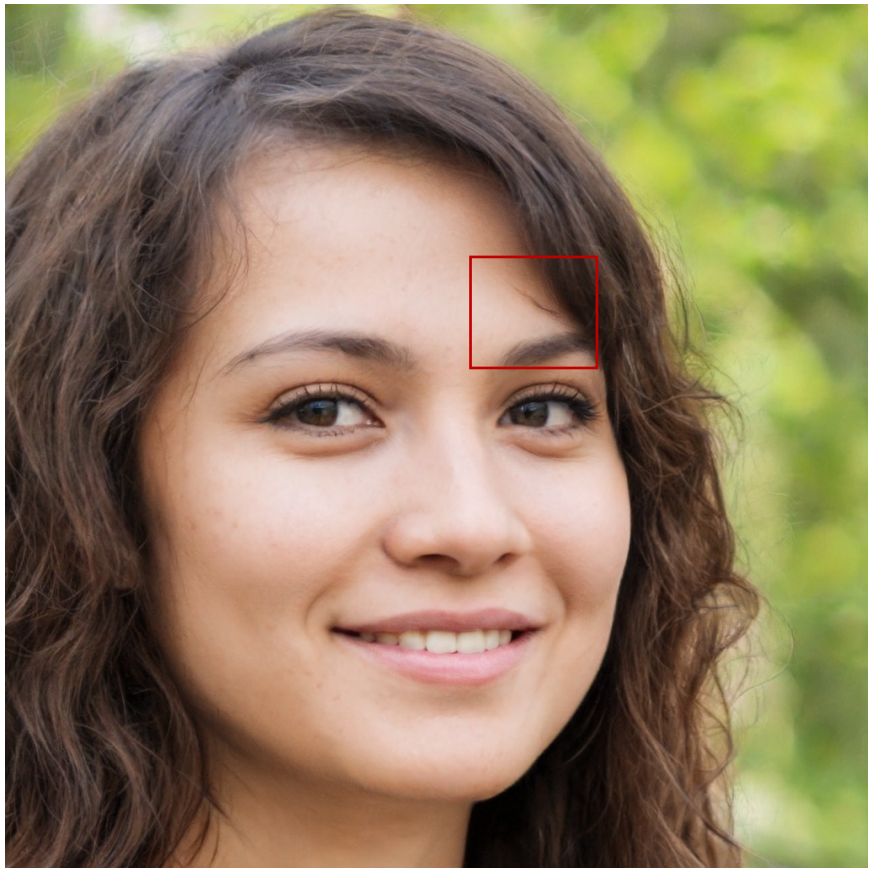
Visual Artifacts

Color anomalies, sharp boundaries, strange artifacts…









Semantic inconsistencies

Spatial inconsistencies in frames, semantic anomalies (e.g. different colour of the eyes), eye blinking absence, biological signal, symmetry inconsistencies

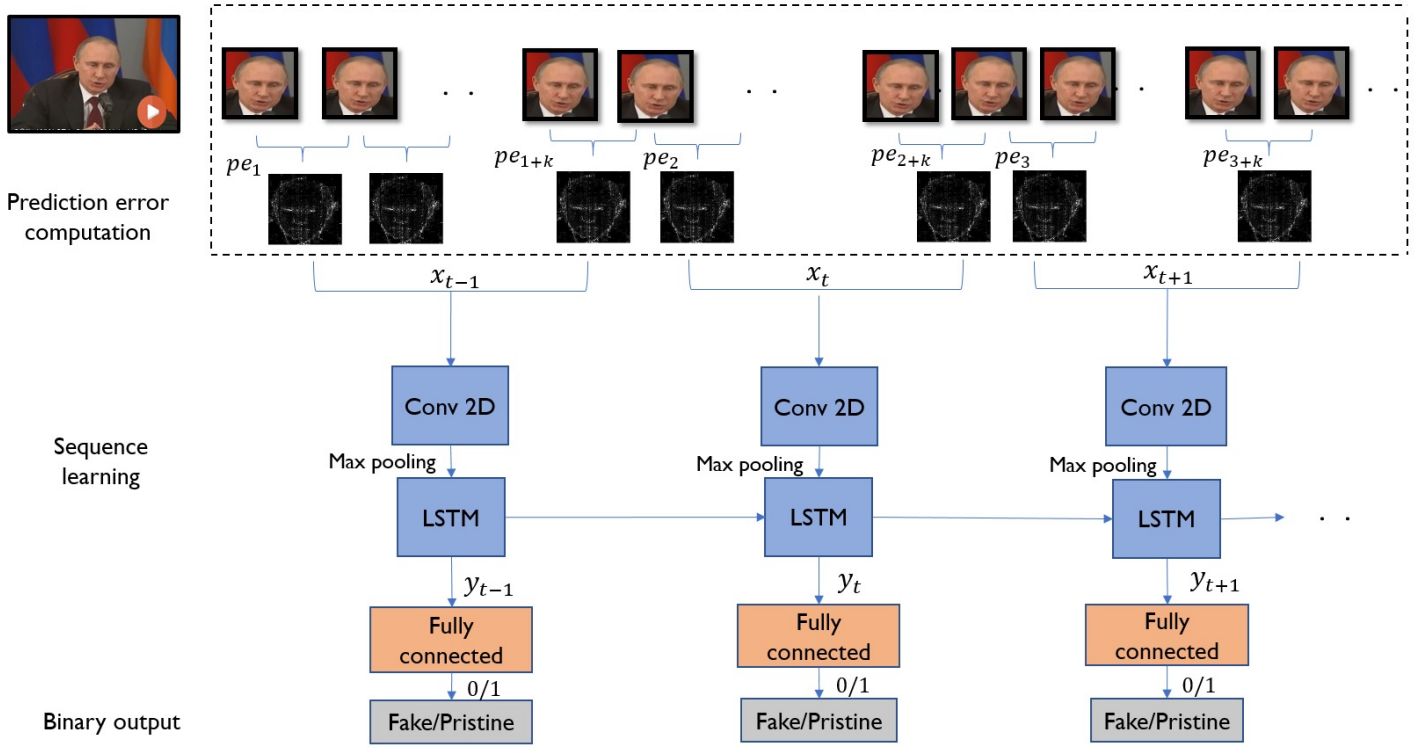
Semantic inconsistencies: time

Real

Deepfake

I. Amerini, R.Caldelli, “Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos”, 8th ACM Workshop on Information Hiding and Multimedia Security 2020

Semantic inconsistencies: time

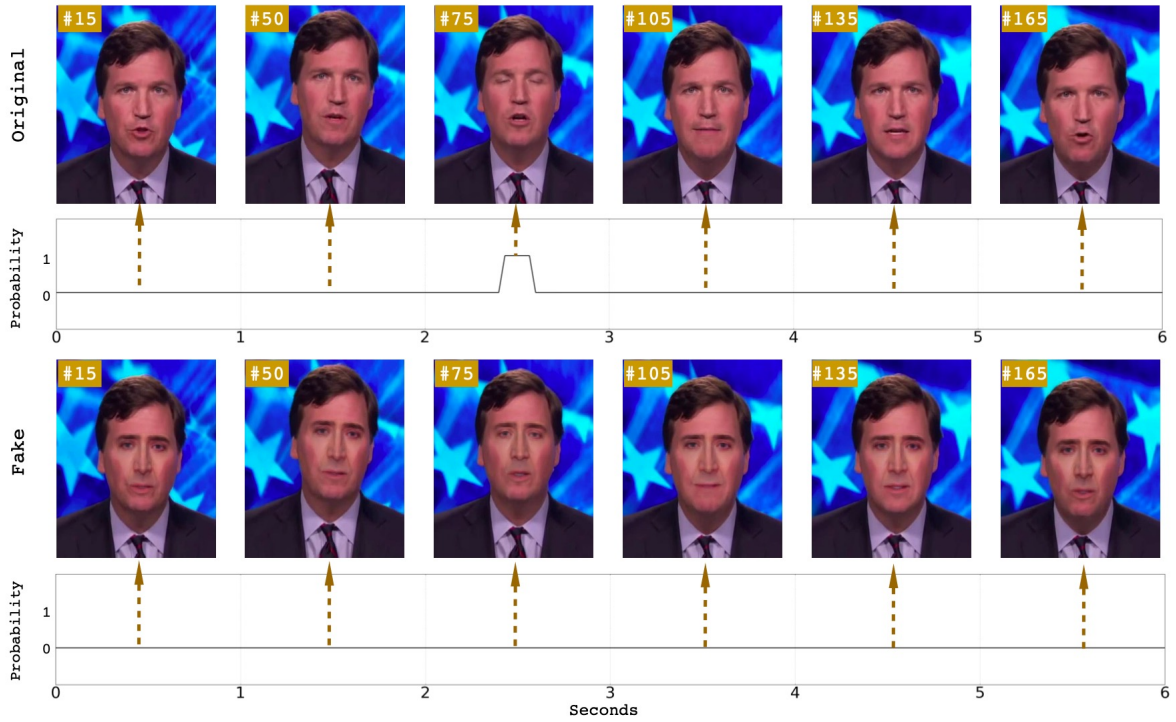


Compute the *Prediction Error* (PE) between the predicted and next frame

$$PE = MB_t(i,j) - MB_{t+1}(i + v_x, j + v_y)$$

I. Amerini, R.Caldelli, "Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos", 8th ACM Workshop on Information Hiding and Multimedia Security 2020

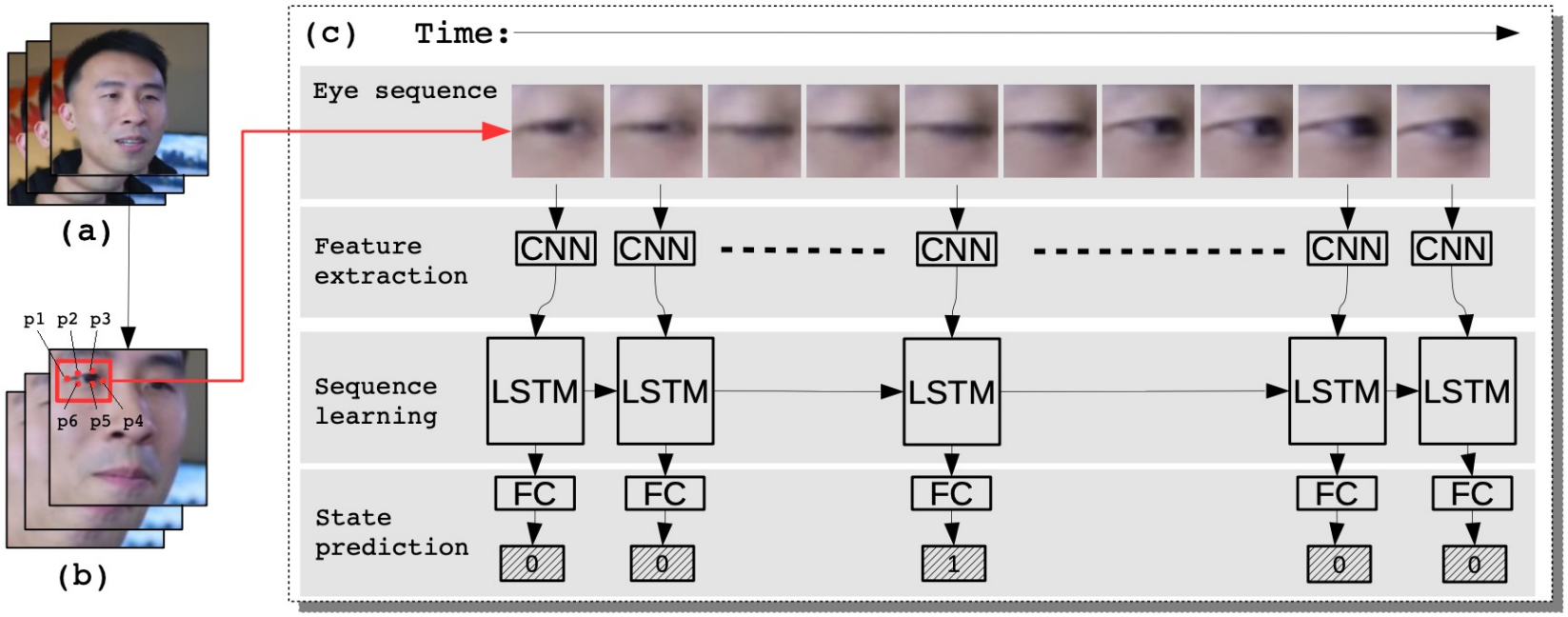
Semantic inconsistencies: eye blinking



In the original video (top), an eye blinking can be detected within 6 seconds, while in the fake video (bottom) such is not the case

Yuezun Li et al., "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking", WIFS 2018

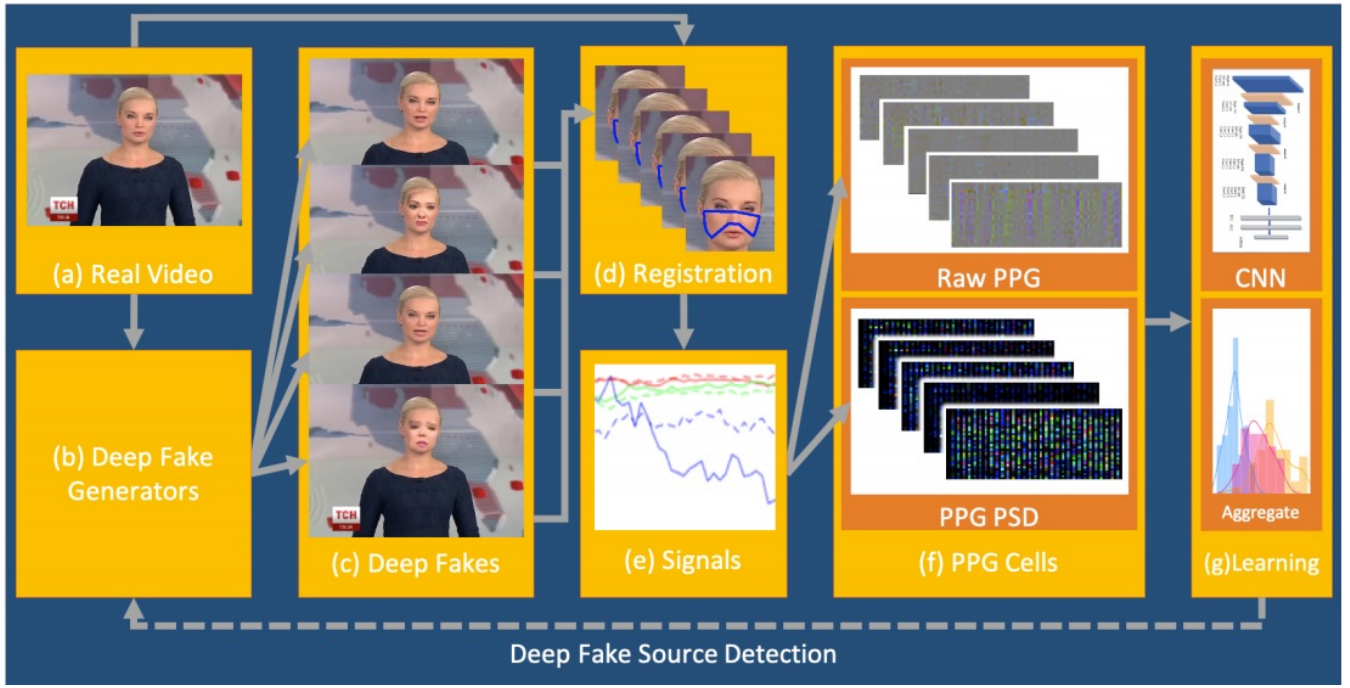
Semantic inconsistencies: eye blinking



The LSTM analyzes the features extracted by the CNN over time. This allows the model to identify inconsistencies in the eye-blinking sequence

Yuezun Li et al., "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking", WIFS 2018

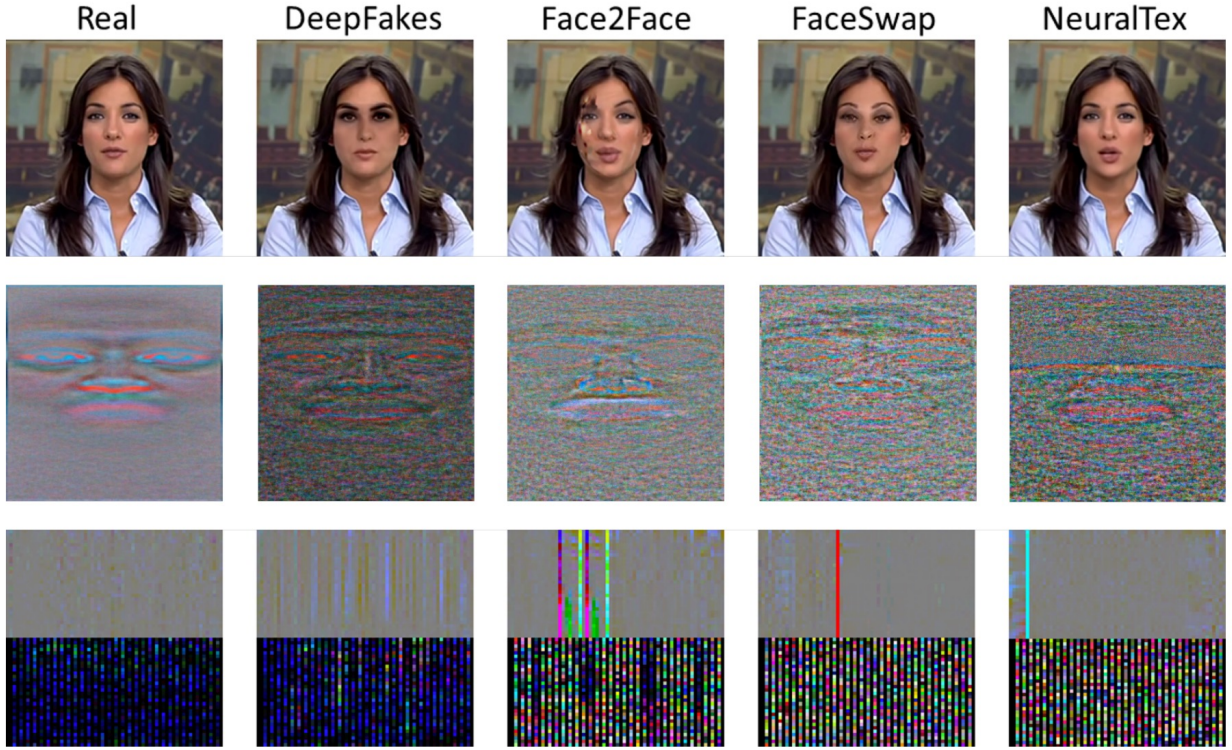
Semantic inconsistencies: heart variations



When blood moves through the veins, it changes the skin reflectance over time, due to the hemoglobin content in the blood. Photoplethysmography (PPG) signals can be extracted to recognize such changes by image processing techniques

Ciftci et al., "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals." IJCB 2020

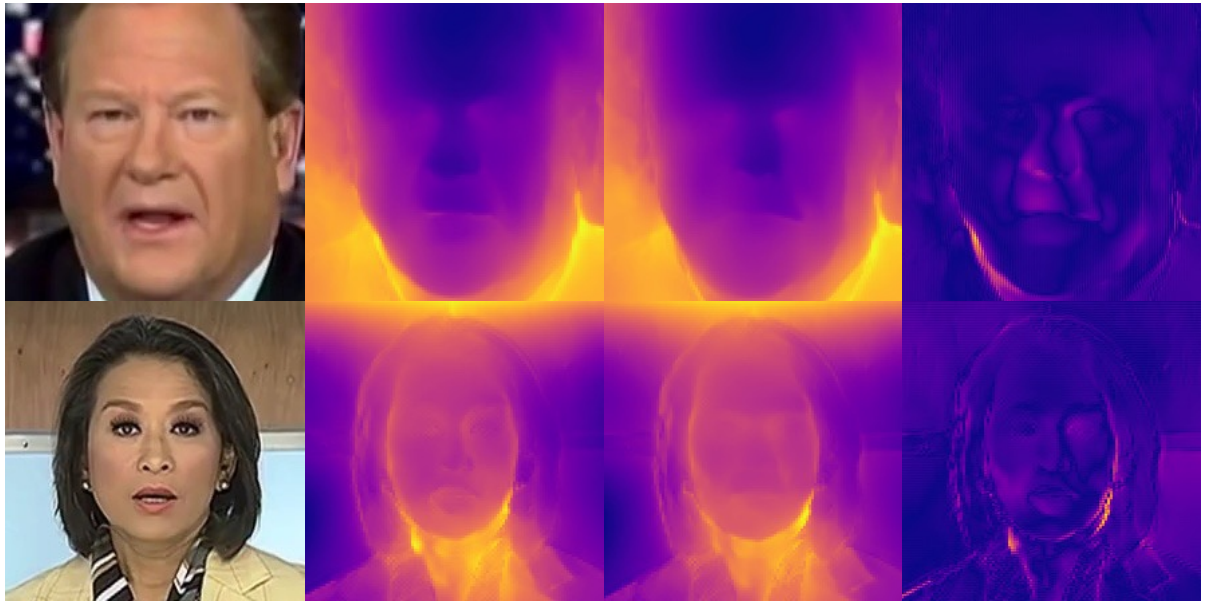
Semantic inconsistencies: heart variations



Example frames per $\omega = 64$ window (top), and their PPG cells (bottom) consisting of raw PPG and PPG PSD, of a real video (left) and its fakes versions per generative model (rest). Middle row represents an approximation to the accumulated residuals over all videos, which correlates with the colors in the PPG spectra

Ciftci et al., "How do the hearts of deep fakes beat? Deep fake source detection via interpreting residuals with biological signals." IJCB 2020.

Semantic inconsistencies: depth

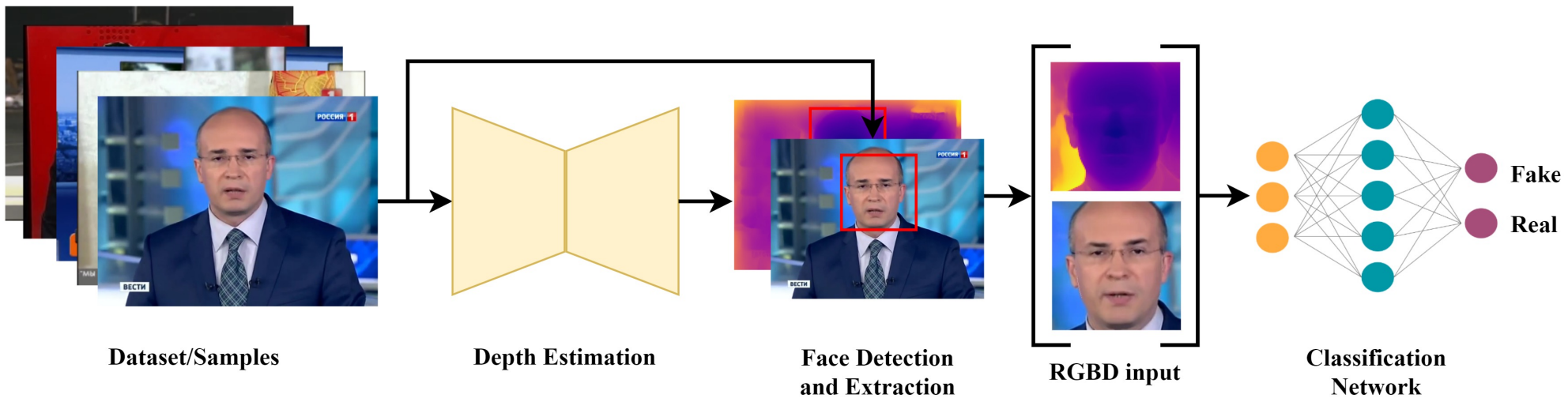


Deepfake face | Real-face's depth | Fake-face's depth | Difference in depths

Some example inconsistencies introduced in the depth map of manipulated faces. Deepfake faces tend to have less details than the original ones

Maiano L., Papa L., Vocaj K., Amerini I., *DepthFake: a depth-based strategy for detecting Deepfake videos*, ICPR Workshop on Artificial Intelligence for Multimedia Forensics and Disinformation Detection, 2022

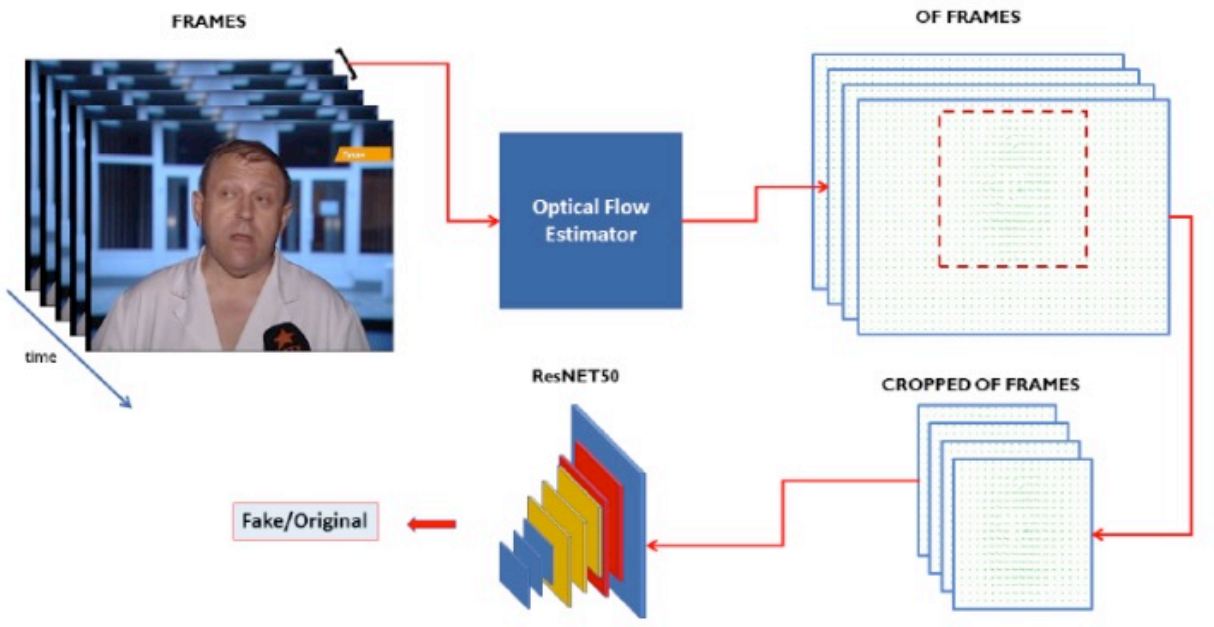
Semantic inconsistencies: depth



In the fist step, we estimate the depth for each frame. Then, we extract the face and crop the frame and depth map around the face. In the last step, we train a classifier on RGBD input features

Maiano L., Papa L., Vocaj K., Amerini I., *DepthFake: a depth-based strategy for detecting Deepfake videos*, ICPR Workshop on Artificial Intelligence for Multimedia Forensics and Disinformation Detection, 2022

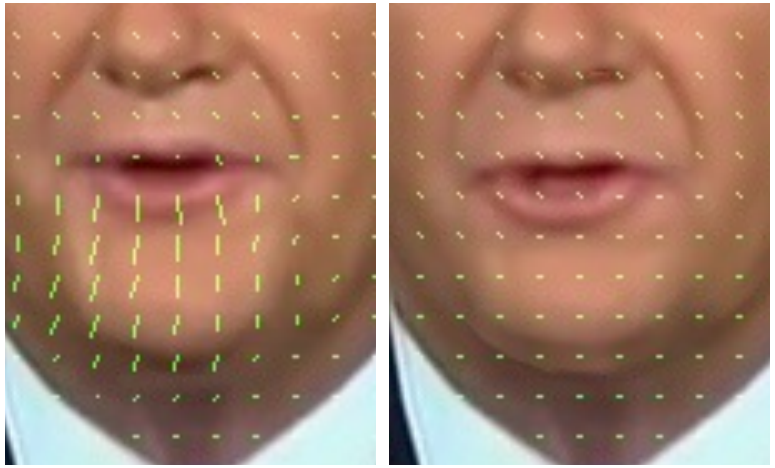
Semantic inconsistencies: optical flow



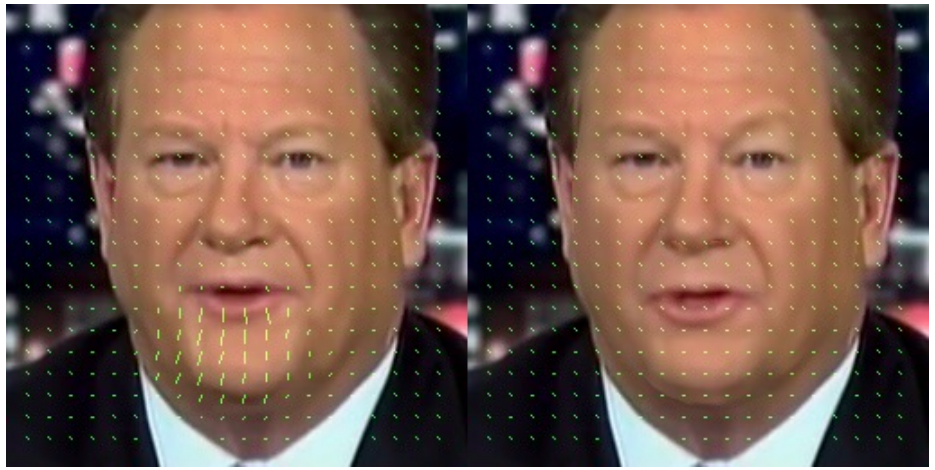
- Optical flow fields are used as input for a semi-trainable neural network
- Fine tune the last convolutional and dense layers on a deepfake dataset

Amerini et Al, "Deepfake Video Detection through Optical Flow based CNN", Human Behaviour and Understanding Workshop, ICCV 2019
R. Caldelli, L. Galteri, I. Amerini, A. Del Bimbo, "Optical Flow based CNN for detection of unlearnt deepfake manipulations, Pattern Recognition Letter 2020.

Semantic inconsistencies: optical flow



Real Deepfake



Real Deepfake

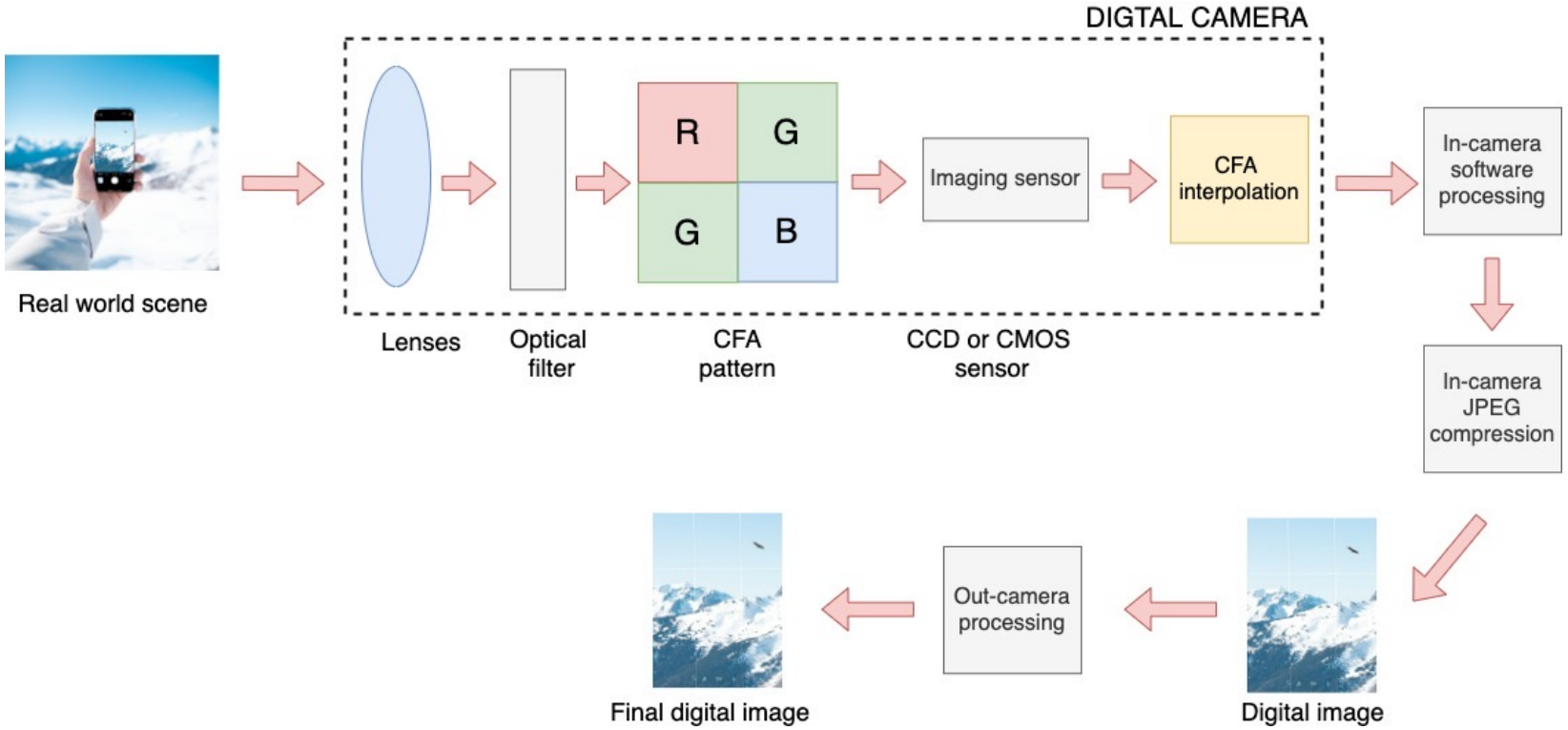
Looking at the motion vectors around the mouth, we can observe a different distribution of the optical flows. Deepfakes tend to be smoother

Amerini et Al, “Deepfake Video Detection through Optical Flow based CNN”, Human Behaviour and Understanding Workshop, ICCV 2019
R. Caldelli, L. Galteri, I. Amerini, A. Del Bimbo, “Optical Flow based CNN for detection of unlearnt deepfake manipulations, Pattern Recognition Letter 2020

Semantic inconsistencies

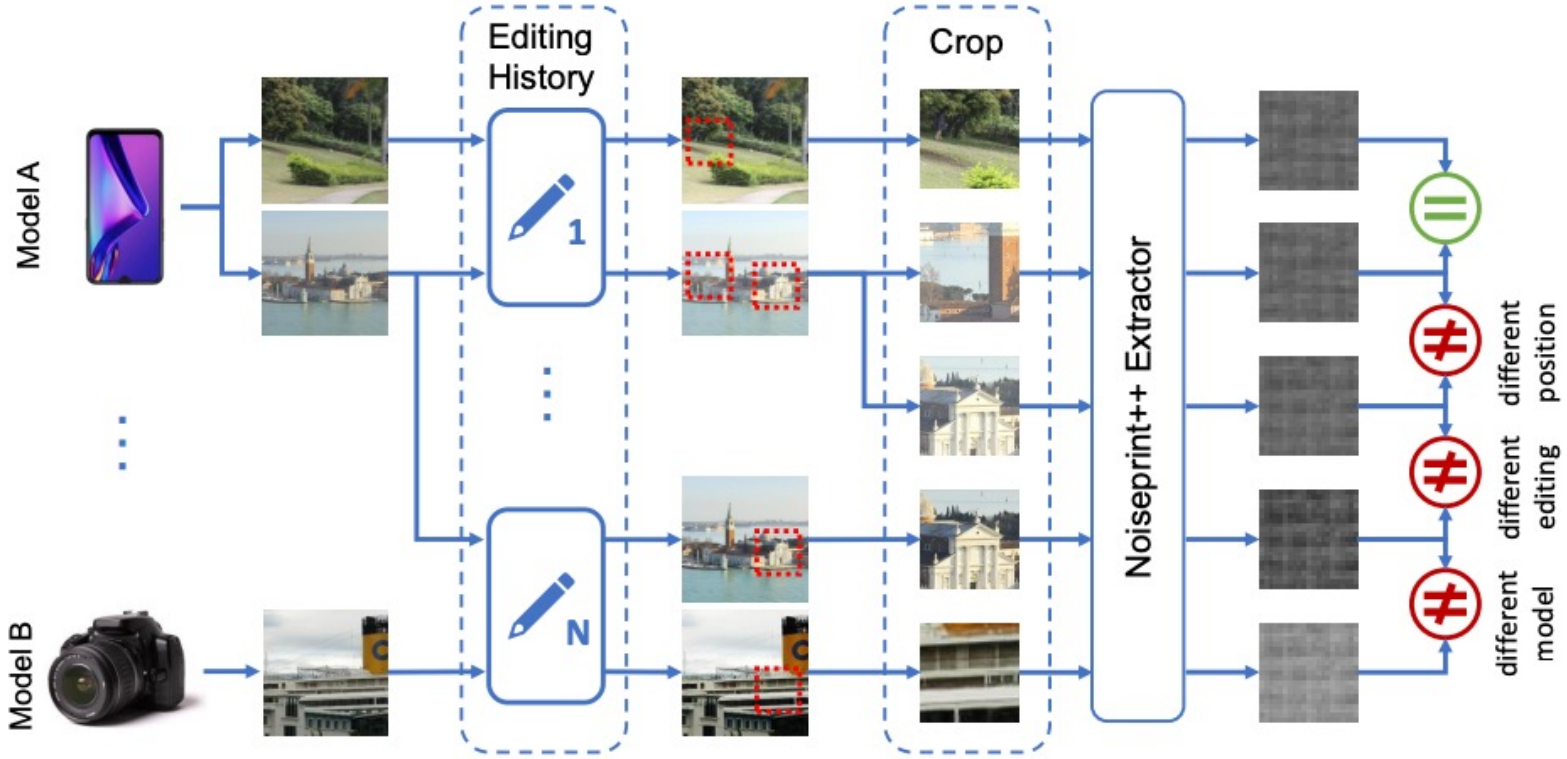
- Corneal specular highlights [Hu 2020]
- Warping artifacts [Li 2019]
- Head pose inconsistencies [Yang 2019a]
- Landmark locations [Yang 2019b]
- Visual artifacts [Matern 2019]
- Color cues [McCloskey 2018, Li 2018, Tondi 2020, Farid 2023a]
- Visual quality metrics [Korshunov 2018]
- Texture features [Bonomi 2020]
- Perspective inconsistency [Farid 2023b]

Camera-related artifacts



Fakes generated with CGI and deep learning have much in common, since they both lack the characteristic features that are typical of human faces acquired by real cameras

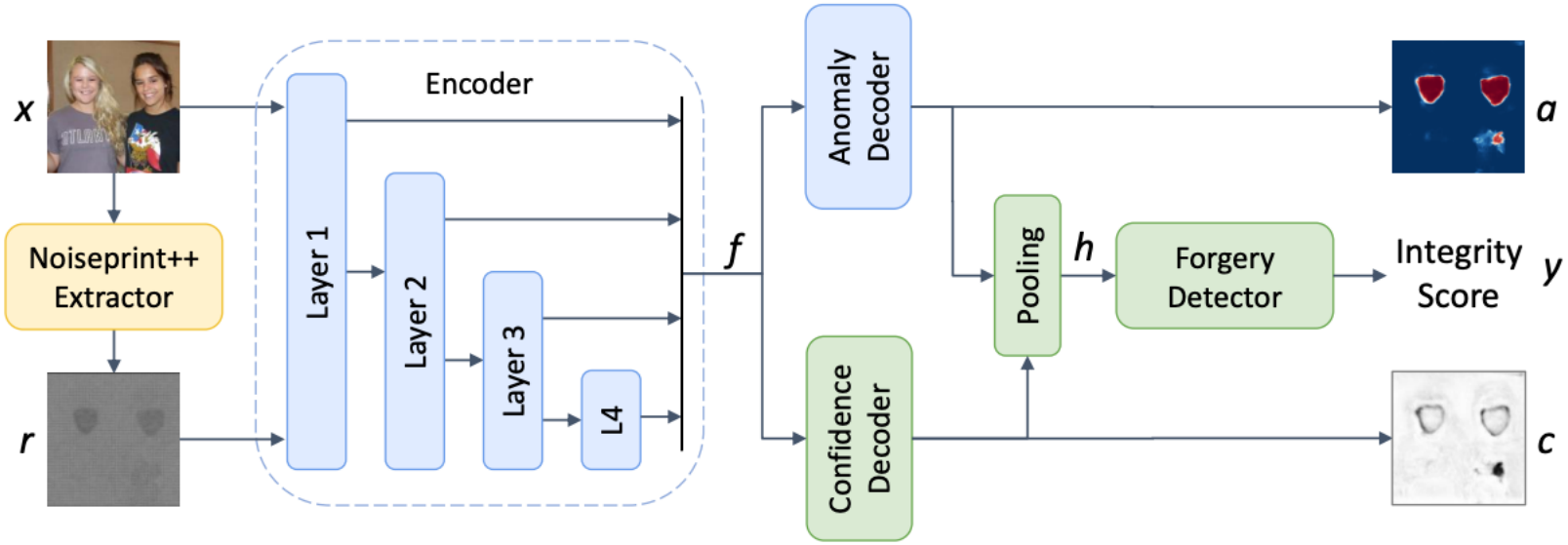
Camera-related artifacts: TruFor



- Different crops are extracted from real images taken from many different cameras
- During training, the distance between the outputs is minimized for patches coming from the same camera model, same position, and same editing history

Guillaro, F., Cozzolino, D., Sud, A., Dufour, N. and Verdoliva, L., 2022. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. arXiv preprint arXiv:2212.10957

Camera-related artifacts: TruFor



- The encoder uses both the RGB input and Noiseprint++ for jointly computing the features that will be used by the anomaly decoder and the confidence decoder for pixel-level forgery localization and confidence estimation, respectively
- The forgery detector exploits the localization map and the confidence map to make the image-level decision

Guillaro, F., Cozzolino, D., Sud, A., Dufour, N. and Verdoliva, L., 2022. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. arXiv preprint arXiv:2212.10957

Camera-related artifacts

- Lugstein, F., Baier, S., Bachinger, G. and Uhl, A., 2021, June. PRNU-based deepfake detection. In Proceedings of the 2021 ACM workshop on information hiding and multimedia security
- Amerini, I., Conti, M., Giacomazzi, P. and Pajola, L., 2022, July. PRaNA: PRNU-based Technique to Tell Real and Deepfake Videos Apart. In 2022 International Joint Conference on Neural Networks (IJCNN)
- Zhang, L., Qiao, T., Xu, M., Zheng, N. and Xie, S., 2022. Unsupervised learning-based framework for deepfake video detection. IEEE Transactions on Multimedia

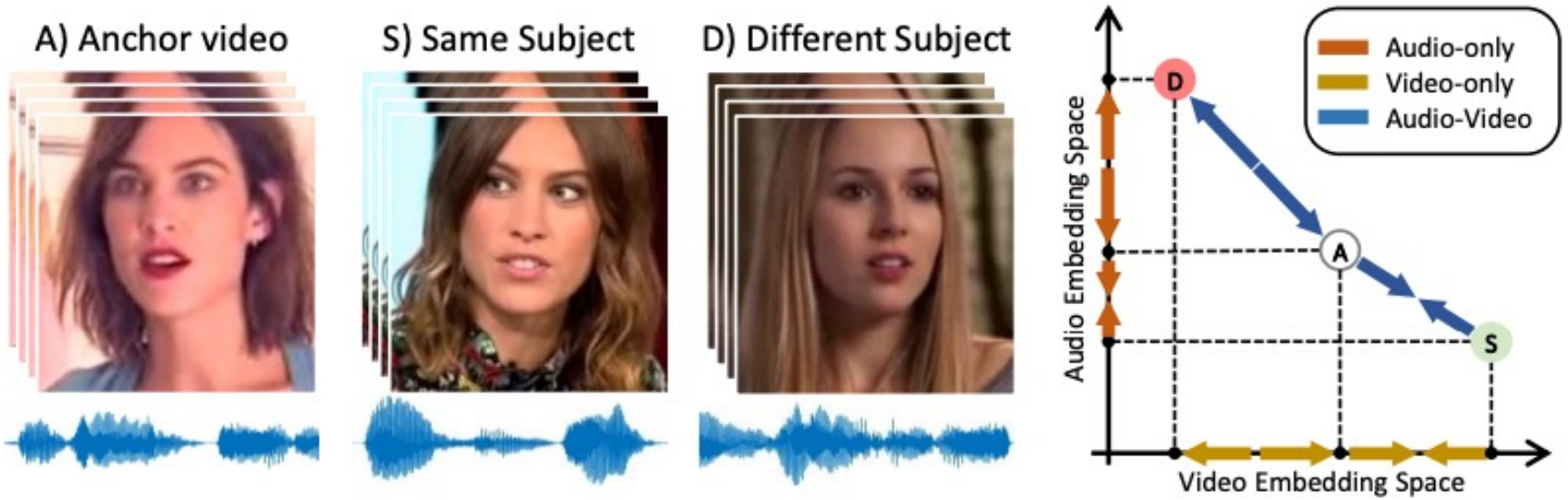
Identity-related inconsistencies

Source/Target

Deepfake

Specific facial expressions and characteristics are not well preserved

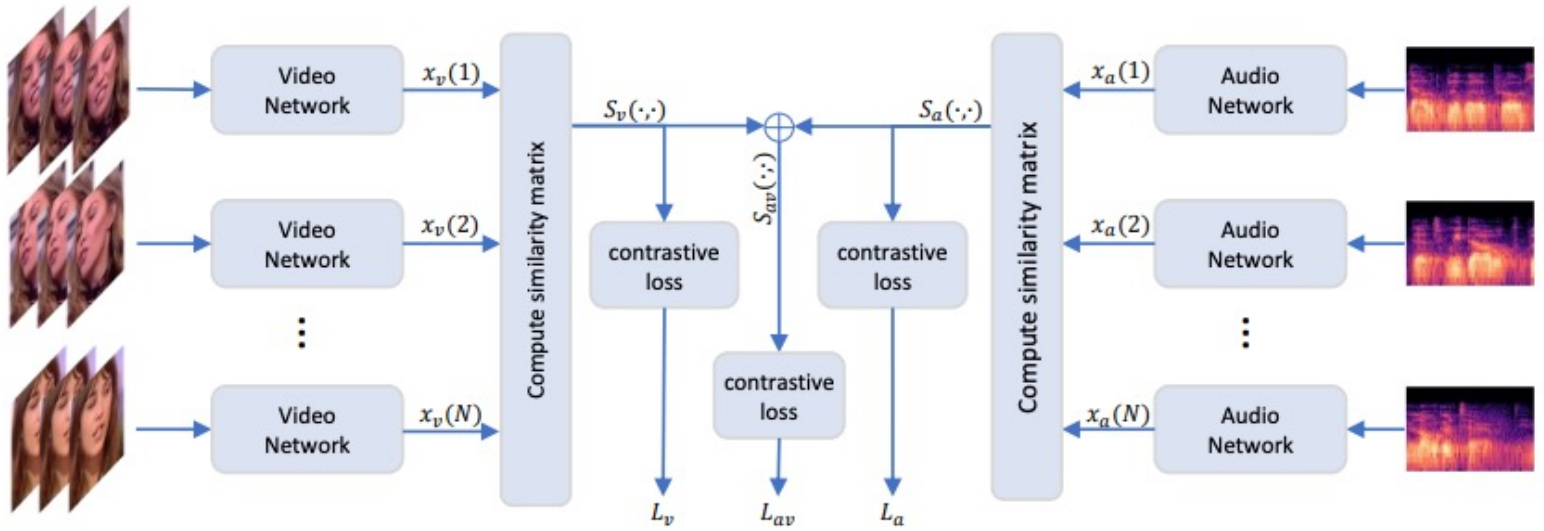
Identity-related inconsistencies



Given a set of real identities, a contrastive learning method encourages embedded vectors of a reference video (A) to be close to embedded vectors of the same subject (S) but far from those of different subjects (D)

Cozzolino, D., Nießner, M. and Verdoliva, L., 2022. Audio-visual person-of-interest deepfake detection. *arXiv preprint arXiv:2204.03083*.

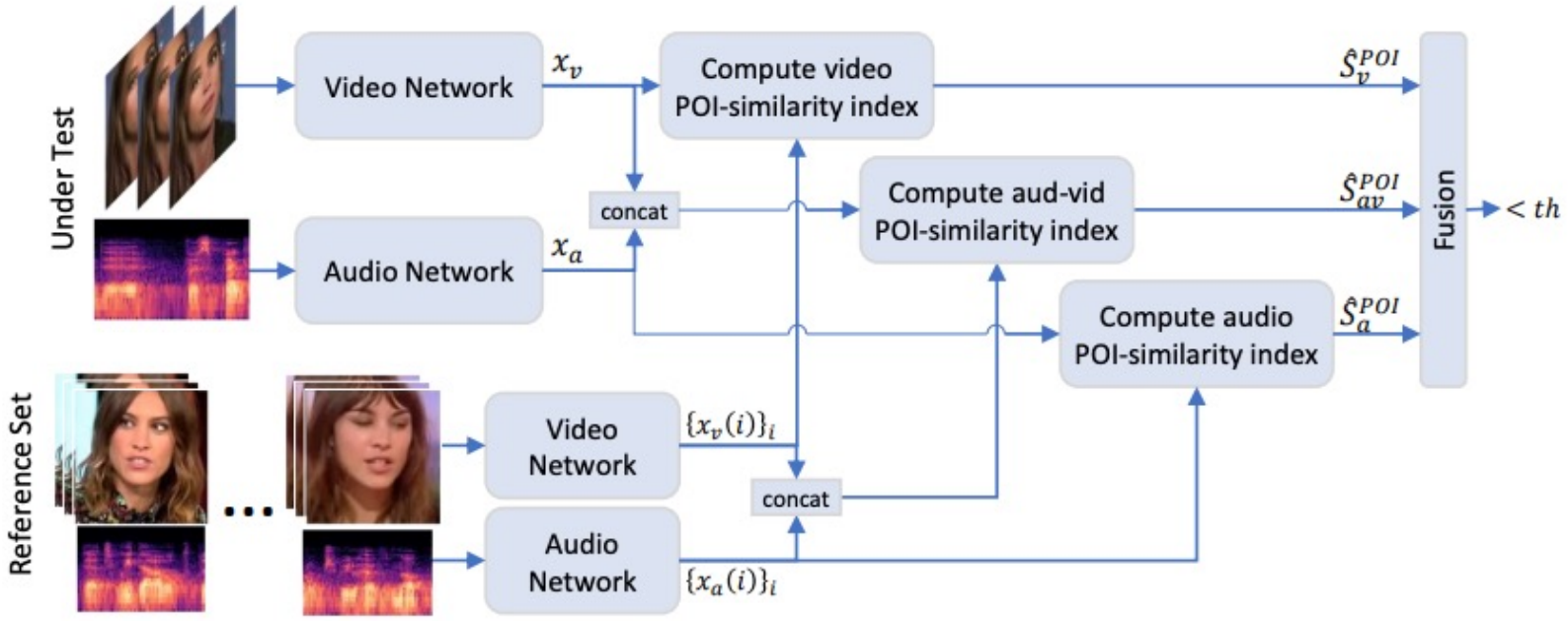
Identity-related inconsistencies



At each training iteration, we analyze N video-segments and extract the embedded vectors from the audio and video signals. All the embedded vectors are compared computing **three N × N matrices of similarity measures** for only-video, only-audio and audio-video, respectively. For each matrix, a **contrastive loss** is evaluated to push closer the embedded vectors of the same individual but move farther from those of different individuals

Cozzolino, D., Nießner, M. and Verdoliva, L., 2022. Audio-visual person-of-interest deepfake detection. *arXiv preprint arXiv:2204.03083*.

Identity-related inconsistencies



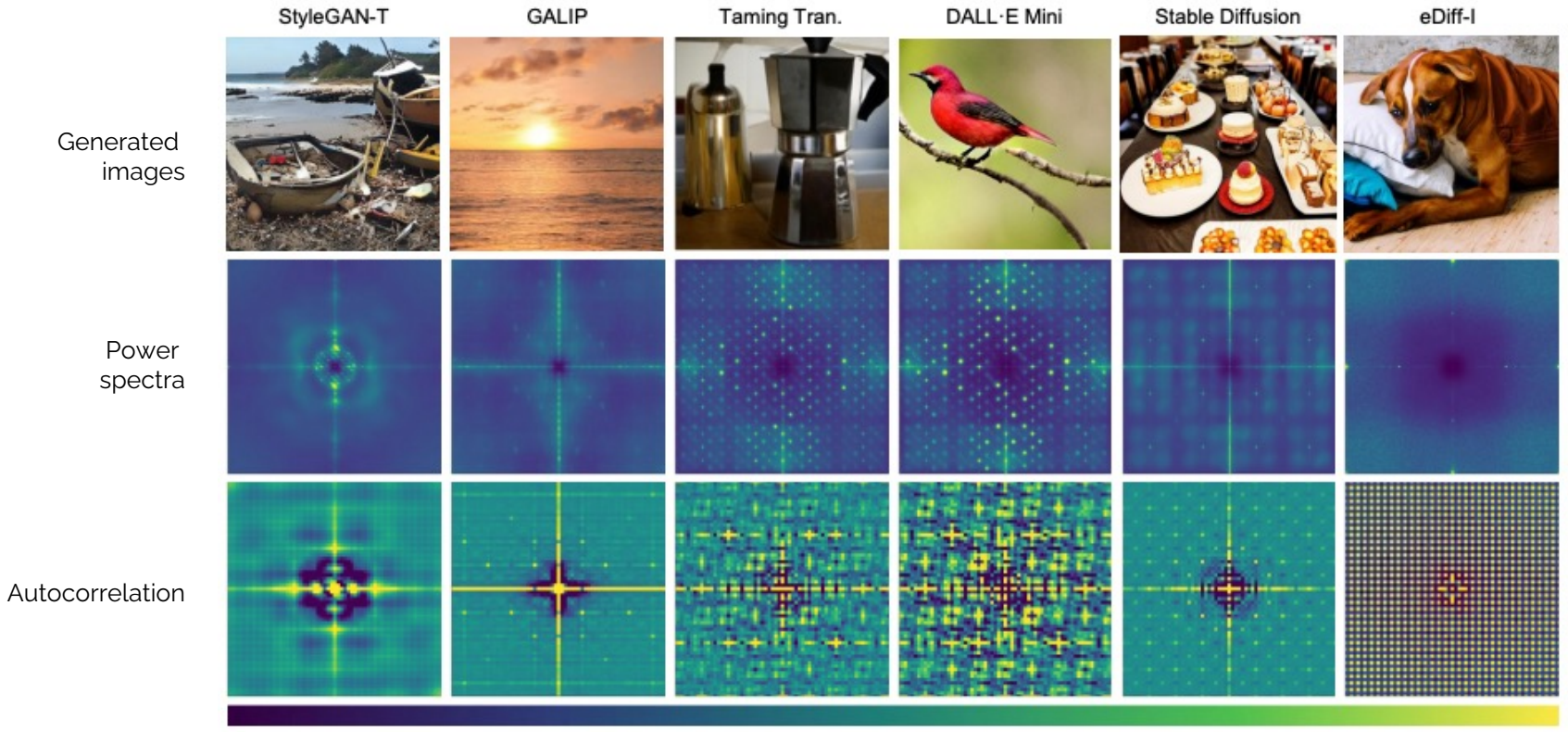
We extract from the audio and video segments the embedded vectors and compare them with those extracted from a set of pristine videos of person of interest by means of the POI similarity indices (audio-only, video-only, audio-video). Finally, a fusion POI similarity index is computed

Cozzolino, D., Nießner, M. and Verdoliva, L., 2022. Audio-visual person-of-interest deepfake detection. *arXiv preprint arXiv:2204.03083*.

Identity-related inconsistencies

- Cozzolino, D., Nießner, M. and Verdoliva, L., 2022. Audio-visual person-of-interest deepfake detection. *arXiv preprint arXiv:2204.03083*
- Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F. and Guo, B., 2022. Protecting celebrities from deepfake with identity consistency transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9468-9478)
- Liu, B., Liu, B., Ding, M., Zhu, T. and Yu, X., 2023. TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4691-4700)
- Nirkin, Y., Wolf, L., Keller, Y. and Hassner, T., 2021. Deepfake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), pp.6111-6121

Model fingerprints



Each model leaves peculiar traces in the generated images, which can be regarded as a sort of artificial fingerprint and used for forensic analyses, much like a real camera is characterized by its PRNU pattern. They are typically visible in the **frequency domain as spectral peaks in the power spectra** (middle), or in the **spatial domain as anomalous regular patterns** in the autocorrelation (bottom). It is easy to observe that models based on the same architecture, like Taming Transformers and DALL· E Mini, give rise to similar artifacts

Model fingerprints

For frequency domain analyses, we start from the Fourier transform of the $M \times N$ image

$$X_i(k,l) = \mathcal{F}[x_i(m,n)] = \sum_{m=1}^{M}\sum_{n=1}^{N} x_i(m,n)e^{-j2\pi\left(\frac{k}{M}m+\frac{l}{N}n\right)}$$

and obtain the **source power spectrum**, again, by averaging all individual power spectra

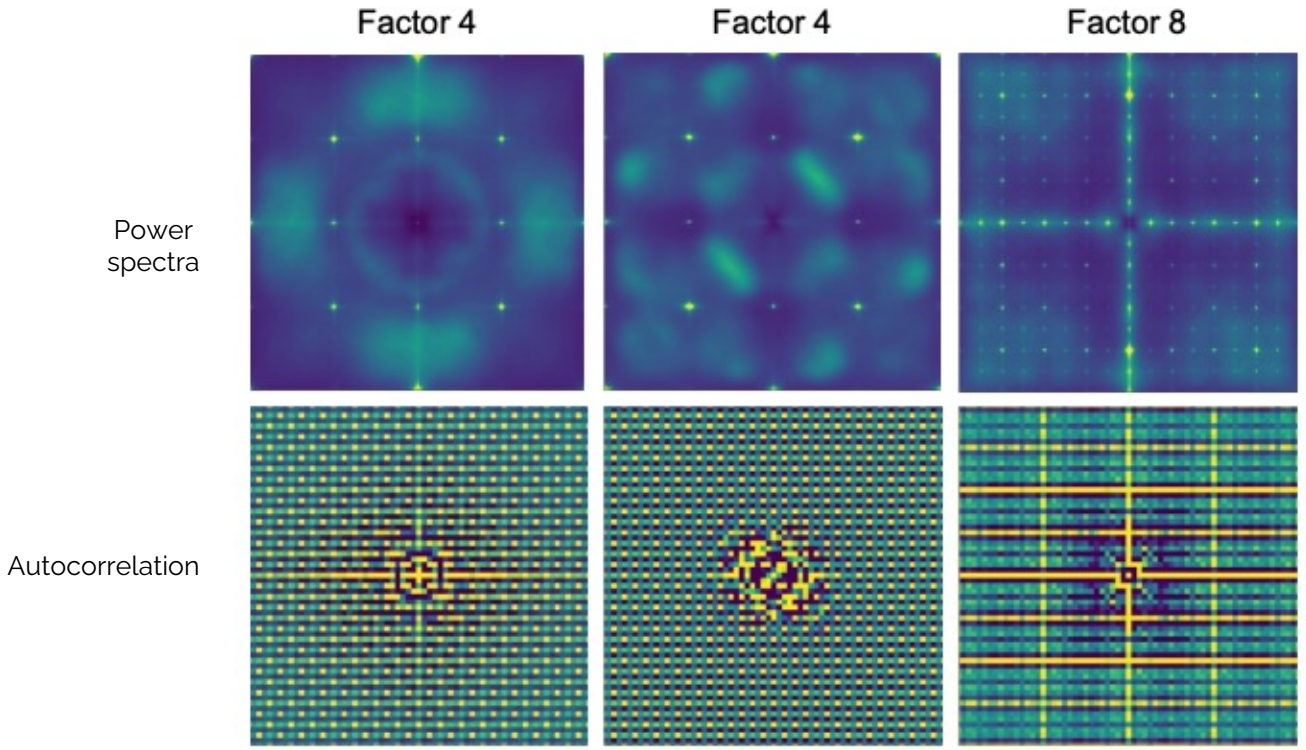
$$S_x(k,l) = \frac{1}{I}\sum_{m=1}^{M} S_{x_i}(k,l) = \frac{1}{I}\sum_{i=1}^{I}|X_i(k,l)|^2$$

The power spectrum accounts for the fraction of the total image power concentrated at a given (horizontal,vertical) frequency pair $(\frac{k}{M}, \frac{l}{N})$

Often, these traces are only visible in the **noise residuals**, obtained by removing the high-level semantic content from the image by means of a denoising filter $\mathfrak{D}$

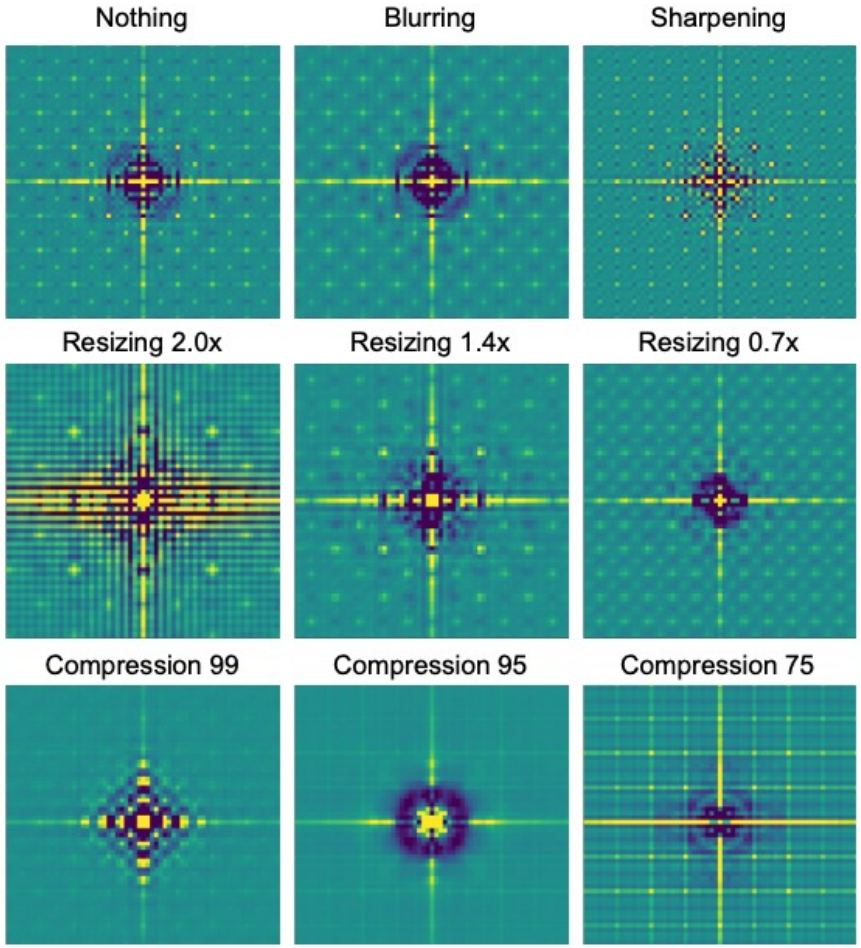
$$r_x(m,n) = x_i(k,l) - \mathfrak{D}(x_i(m,n);\sigma)$$

Model fingerprints: upsampling factor



Power spectra (top) and autocorrelation functions (bottom) of noise residuals of three slightly different latent diffusion models. The upsampling factor between latent space and pixel space is 4 for the first two architectures and 8 for the third one. This single parameter appears to be **responsible for the positioning of the peaks in the power spectra and the periodicity of the regular patterns in the autocorrelations**

Model fingerprints: post-processing



Effect of post-processing operations on the autocorrelation of noise residuals. We consider Stable Diffusion images and apply different operations to them to see how the autocorrelation changes. **In some cases, post-processing artifacts completely hide the original generation traces** (top-left)

Corvi, R., Cozzolino, D., Poggi, G., Nagano, K. and Verdoliva, L., 2023. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. arXiv preprint arXiv:2304.06408
Papa L., Corvitto L., Faiella L., Maiano L., Amerini I., On the use of Stable Diffusion for creating realistic faces: from generation to detection, 11th International Workshop on Biometrics and Forensics - IWBF2023

Model fingerprints

- Marra et al. “Do GANs leave artificial fingerprints”, IEEE MIPR 2019
- Yu et al., “Attributing Fake Images to GANs: Learning and Analyzing GAN fingerprints”, ICCV 2019
- Zhang et al., ”Detecting and simulating artifacts in GAN fake images”, IEEE WIFS 2019
- Frank et al., “Leveraging Frequency Analysis for Deep Fake Image Recognition”, IEEE CVPR 2019
- Papa et al. “On the use of Stable Diffusion for creating realistic faces: from generation to detection”,11th International Workshop on Biometrics and Forensics - IWBF2023

Deepfake detection datasets

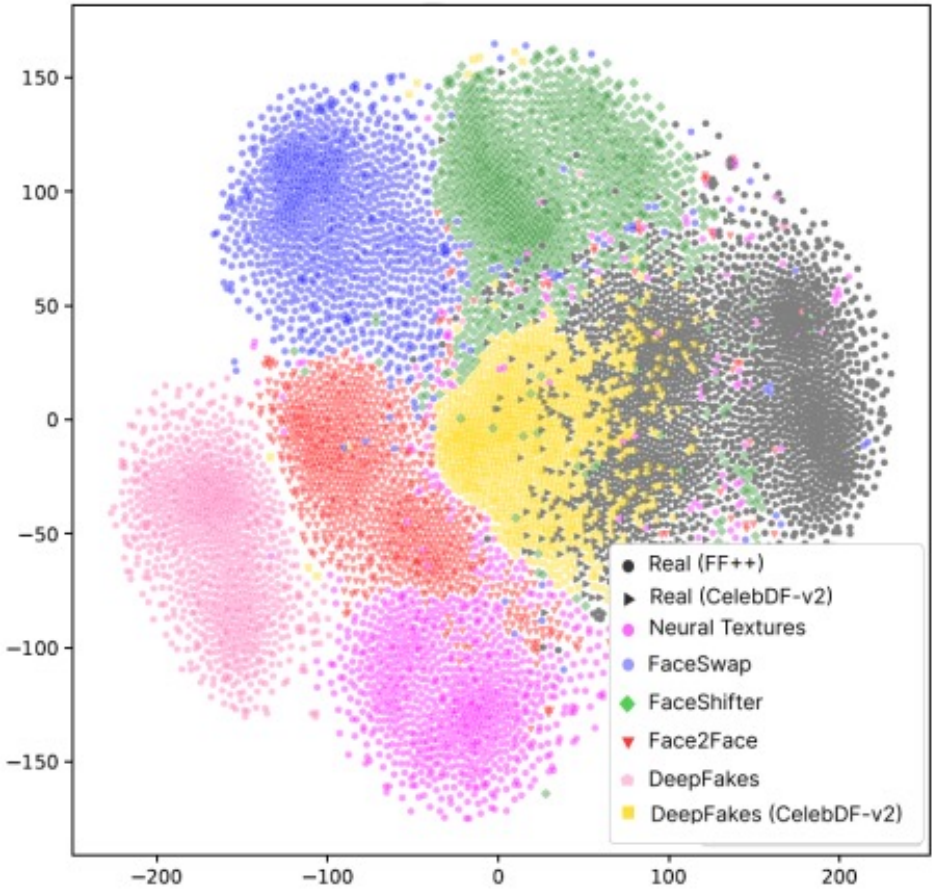
Dataset	Real Source	Deepfake Source	Continual
Deepfake-TIMIT [40]	VidTIMIT dataset [66]	Known Deepfake tech	✗
UADFW [79]	EBV dataset [44]	Known Deepfake tech	✗
FaceForensics++ [64]	YouTube	Known Deepfake tech	✗
Celab-DF v2 [47]	YouTube	Known Deepfake tech	✗
DFDC [19]	Actors	Known Deepfake tech	✗
WildDeepfake [88]	Internet	Unknown Deepfake tech	✗
WhichFaceReal [3]	Internet	Unknown Deepfake tech	✗
CNNfake [75]	Multi-Datasets	Known Deepfake tech	✗
GANfake [53]	Multi-Datasets	Known Deepfake tech	✓
CoReD [36]	Multi-Datasets	Known Deepfake tech	✓
CDDB (ours)	**Multi-Datasets&Internet**	**Known&unknown tech**	✓

Faceforensics++, Celeb-DF v2 and DFDC are the most used for deepfake video detection; however, new datasets are continuously introduced

Li, C., Huang, Z., Paudel, D.P., Wang, Y., Shahbazi, M., Hong, X. and Van Gool, L., 2023. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1339-1349)

Current limitations

Data shift and lack of generalization



Le, B., Tariq, S., Abuadbba, A., Moore, K. and Woo, S., 2023. Why Do Deepfake Detectors Fail?

- Data shift refers to changes in the statistical properties of the data distribution used to train the detection model compared to the distribution of data the model encounters in deployment
- The image on the left shows the intermediate representations of several deepfake datasets extracted through a pre-trained self-supervised learning model
- State-of-the-art methods suffer from a problem of overfitting on the training data and the lack of generalization across different datasets and generative models

Deepfake Detection Challenge (DFDC)

In 2019, Meta AI (i.e., Facebook) partnered with other industry leaders and academic experts to accelerate the development of new ways to detect deepfake videos

Two key ingredients boost the robustness of the winning detectors:

1. **Data augmentations** like dropping portions of faces (either randomly, using landmarks, or using attention-based networks) or blending a real face and an AI-generated one and then using the blending coefficient as a target label
2. **Ensabling** can also be a solution to increase generalization. Fusing different networks has been shown to boost performance

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. and Ferrer, C.C., 2020. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397

A good recipe for a robust detector

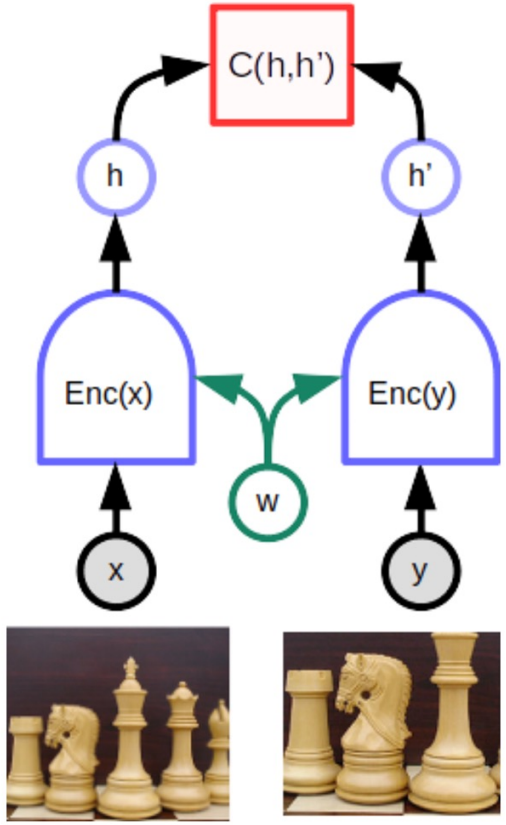
Here are some best practices to train a robust detector:

- **Fine-tune models** pre-trained on ImageNet or very large datasets
- Perform **strong augmentation** by including Gaussian noise adding transformations, cut-outs, and brightness and contrast changes
- **Do not perform downsampling** in the first layer
- **Avoid image resizing**

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

Toward more general approaches

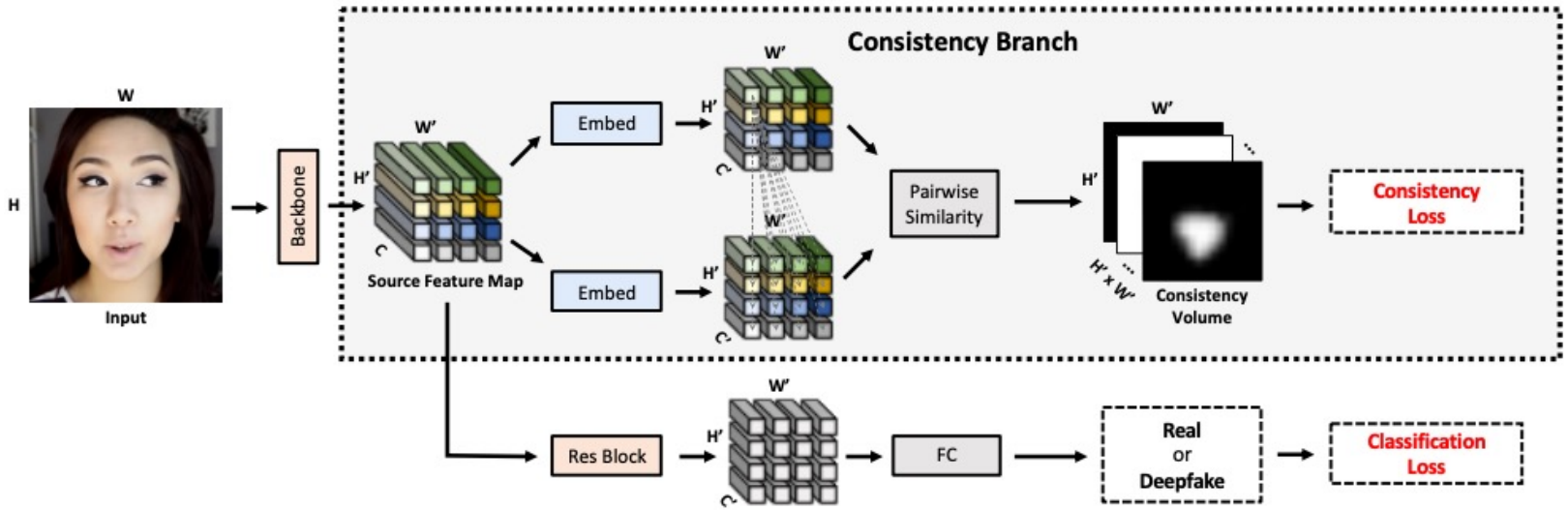
Self-supervised learning (SSL)



- In computer vision, models learn how to predict **masked patches of an image** or representation. Other SSL objectives encourage **two views** of the same image, formed by say adding color or cropping, to be mapped to **similar representations**
- While traditional supervised learning methods are trained on a specific task often known a priori based on the available labeled data, **SSL learns generic representations useful across many tasks**
- There's also evidence SSL models can learn representations that are **more robust to adversarial examples**, label corruption, and input perturbations—and are more fair— compared to their supervised counterparts [Hendrycks 2019, Goyal 2022]

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y. and Schwarzschild, A., 2023. A Cookbook of Self-Supervised Learning. *arXiv preprint arXiv:2304.12210*

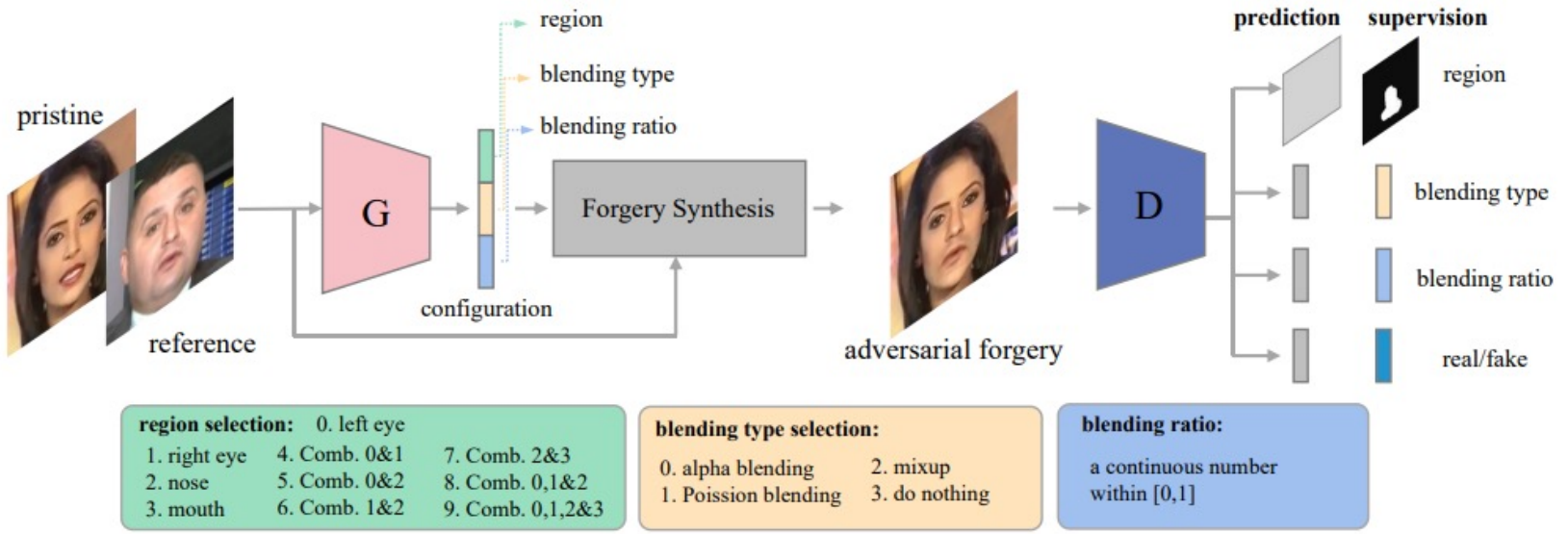
SSL for self-consistency



- Given a pre-processed video frame X, we first feed it into the backbone and extract the source feature F of size $H' \times W' \times C$ from an intermediate convolution layer
- For each patch in the source feature map, we compare it against all the rest to measure their feature similarities and obtain a 2D consistency map of size $H' \times W'$ of consistency scores in the range of $[0, 1]$
- A classification branch is applied after the source feature map and predicts the binary score for deepfake detection

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y. and Xia, W., 2021. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15023-15033)

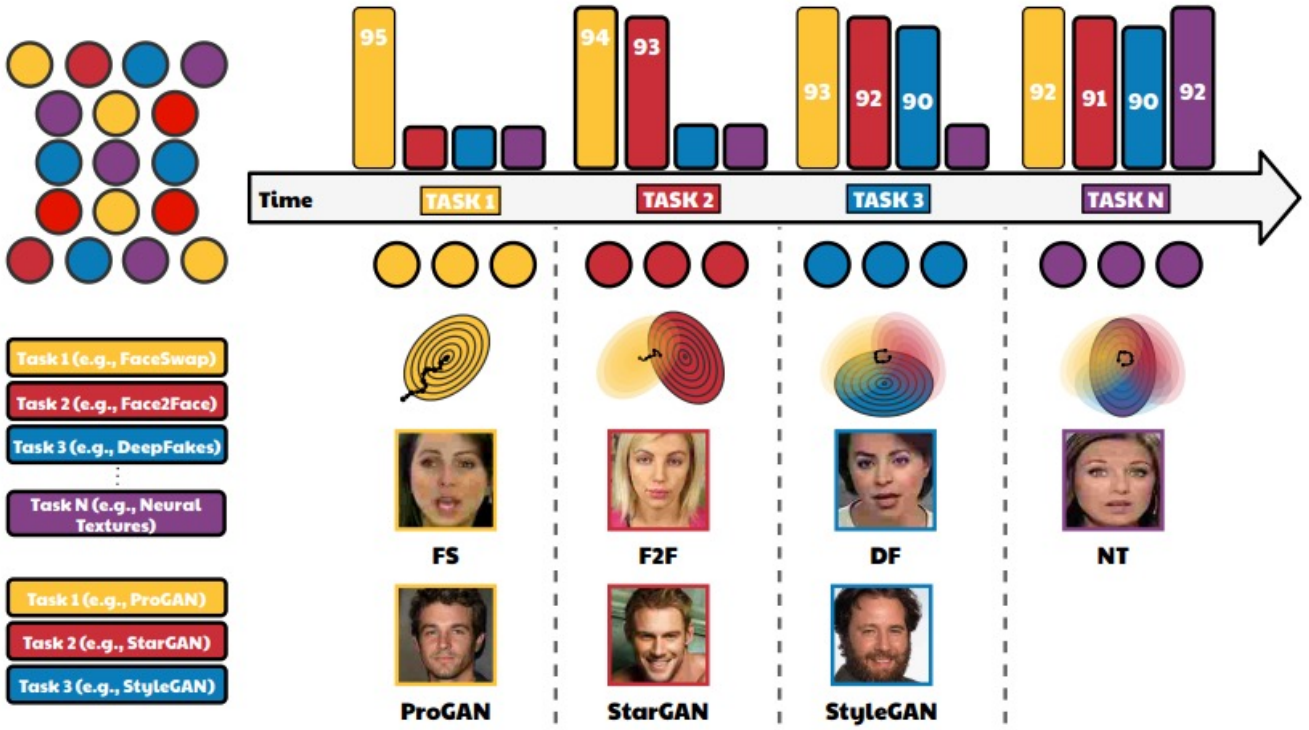
SSL with adversarial examples



- The synthesizer network (i.e. generator) outputs three forgery configurations that are further used to synthesize a new forgery, and these forgery configurations are also used as labels to guide the detector network (i.e. discriminator)
- We train the generator and discriminator in an adversarial manner
- We don't need fake images at training time. The fake examples are dynamically generated at training time

Chen, L., Zhang, Y., Song, Y., Liu, L. and Wang, J., 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18710-18719)

Continual learning (CL)



- We consider deepfakes as a **stream** of examples that **appear over time**, and the early appeared deepfakes cannot be fully accessed due to the streaming nature of data, privacy concerns, or storage constraints
- At each learning session, standard neural networks typically forget most of the knowledge related to previously learned deepfake detection tasks (i.e., **catastrophic forgetting**)

Li, C., Huang, Z., Paudel, D.P., Wang, Y., Shahbazi, M., Hong, X. and Van Gool, L., 2023. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1339-1349)

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T., 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, *44*(7), pp.3366-3385

CL: forgetting in machines

Catastrophic interference, also known as catastrophic forgetting, is the tendency of an artificial neural networks to **completely and abruptly forget previously learned information** upon learning new information (mostly due to Gradient Descent)

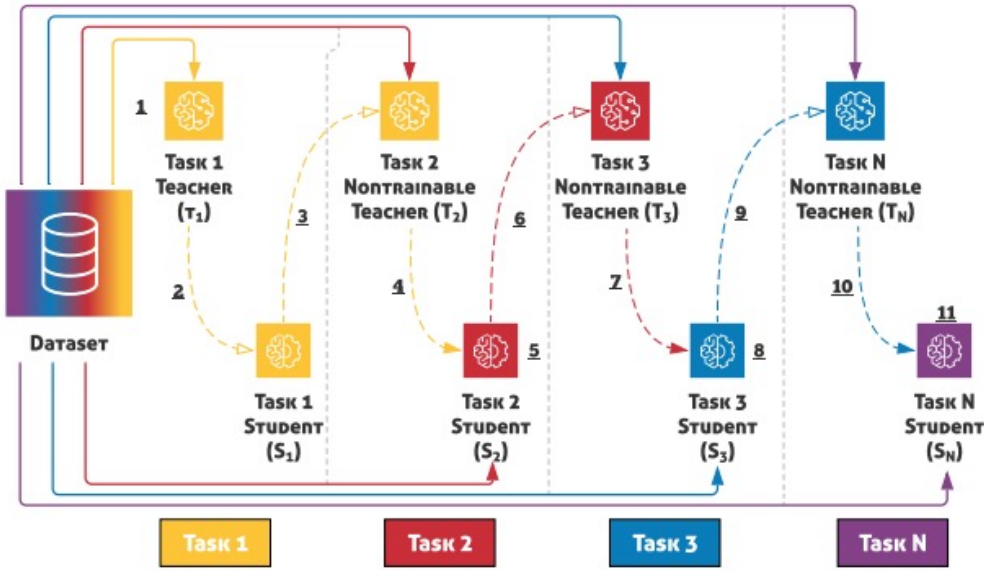
The objective of a continual learning algorithm is to minimize the loss $\mathcal{L}_s$ over the entire stream of data S:

$$\mathcal{L}_s = \frac{1}{\sum_{i=1}^{n} |\mathfrak{D}_{test}^i|} \sum_{i=1}^{n} \mathcal{L}_{exp}\left(f_n^{CL}, f_{test}^i\right)$$

$$\mathcal{L}_{exp}\left(f_n^{CL}, f_{test}^i\right) = \sum_{j=1}^{|\mathfrak{D}_{test}^i|} \mathcal{L}\left(f_n^{CL}(x_j^i), y_j^i\right)$$

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G. and Tuytelaars, T., 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, *44*(7), pp.3366-3385

CL with knowledge distillation



We have a teacher (T_i) and a student network (S_i). We train the student with the following objective:

$$\mathcal{L}_{CoReD} = \alpha\mathcal{L}_{stud} + \beta\mathcal{L}_{distill} + \gamma\mathcal{L}_{repr}$$

Student loss:

$$\mathcal{L}_{stud} = -t_1 \log\Big(\sigma\big(S(x_i, y_i)\big)\Big) - (1-t)\log\Big(1-\sigma\big(S(x_i, y_i)\big)\Big)$$

Distillation loss:

$$\mathcal{L}_{distill} = \sum_{j=1}^{i} \mathcal{L}\big(T(x_i), S(x_i)\big)$$

$$= \sum_{j=1}^{i} \sigma_d\big(T(x_i, y_i), S(x_i)\big) \log \sigma_d\big(S(x_i, \hat{y}_i)\big)$$

$$\sigma_d(s, T)_i = \frac{e^{\frac{s_i}{\tau}}}{\sum_j^C e^{s_j}}$$

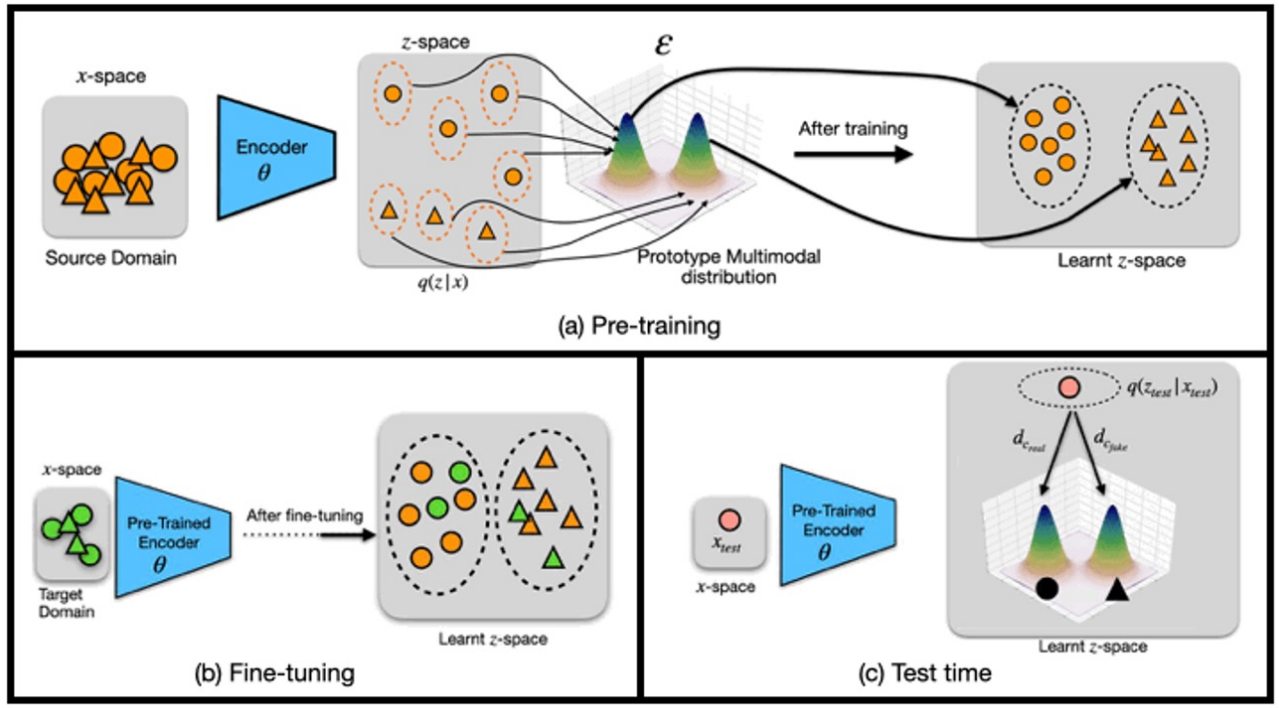
The representation loss $\mathcal{L}_{repr}$ is an additional (optional) term that forces the learned representations of real and fake images to stay consistent over time

Kim, M., Tariq, S. and Woo, S.S., 2021, October. Cored: Generalizing fake media detection with continual representation using distillation. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 337-346)

Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Model classes as distributions

Mixture models learn a multi-modal distribution, with modes corresponding to classes from the source domain. This helps, for example, with few-shot learning



a) **Pre-training**: Θ encodes samples from the source domain into a latent distribution $q(z_i|z_i)$. ε_c then maps class labels c to the encoded distributions of the prototype multi-modal distribution ε

b) **Fine-tuning**: the pre-trained encoder Θ is used to map the few-shot samples from the target dataset with the same prototype multi-modal distribution ε, which learns a common subspace between samples across domains

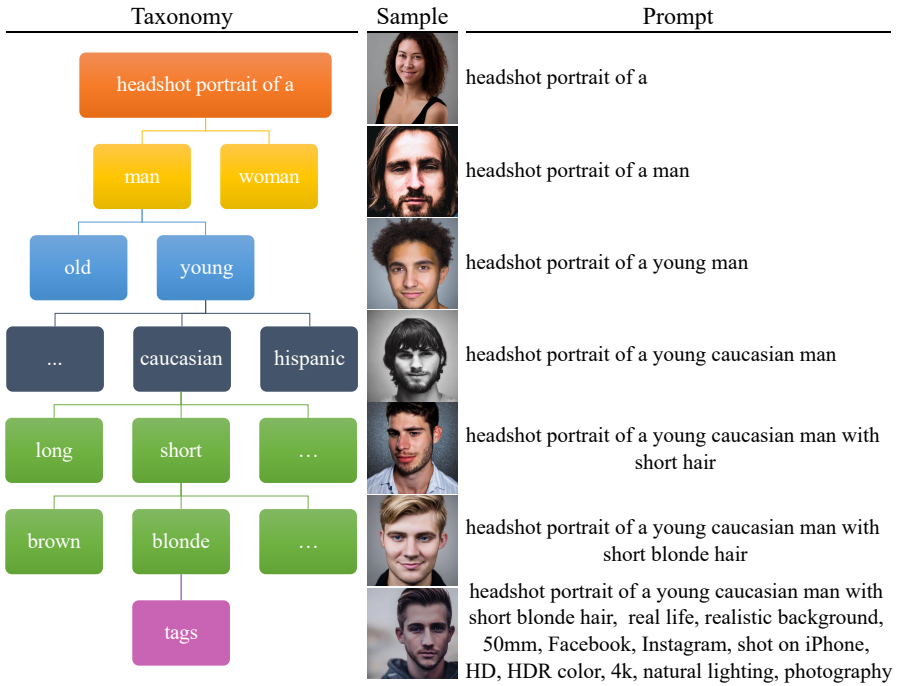
c) **Test-time**: a test sample x_{test} is encoded into the latent distribution $q(z_{test}|z_{test})$ by using the pre-trained encoder Θ. We then compute the distance of the latent code with respect to all components of the distribution, and assign a class label based on the component it is closest to

Aneja, S. and Nießner, M., 2020. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint arXiv:2006.11863*

Our research at the Alcor Lab

Stable diffusion: generating realistic faces

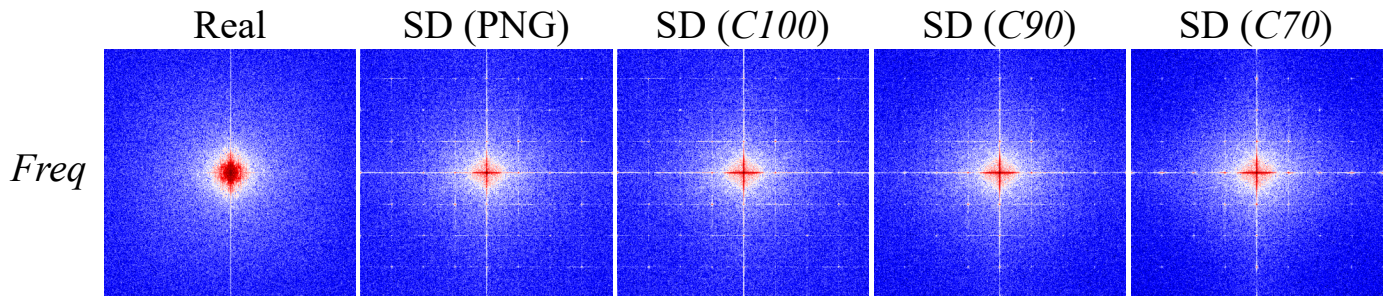
(a) (b) (c)



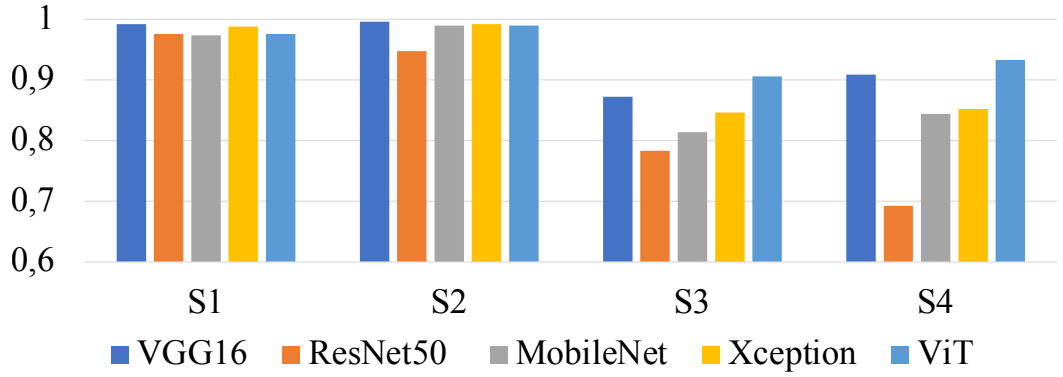
We focus our attention on the generation of realistic human faces (Figure c). These images must not seem fake (Figure a) and must be set in a realistic context that recalls a photo captured with a smartphone

Papa L., Corvitto L., Faiella L., Maiano L., Amerini I., On the use of Stable Diffusion for creating realistic faces: from generation to detection, 11th International Workshop on Biometrics and Forensics - IWBF2023

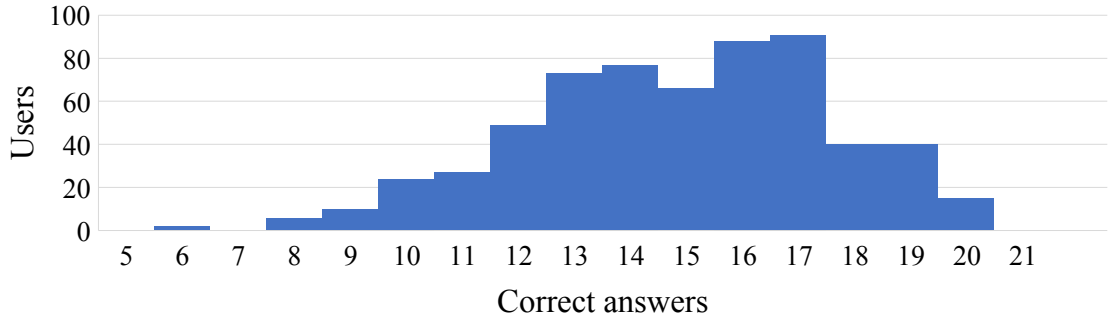
Stable diffusion: detecting realistic faces



(c) Frequency analysis

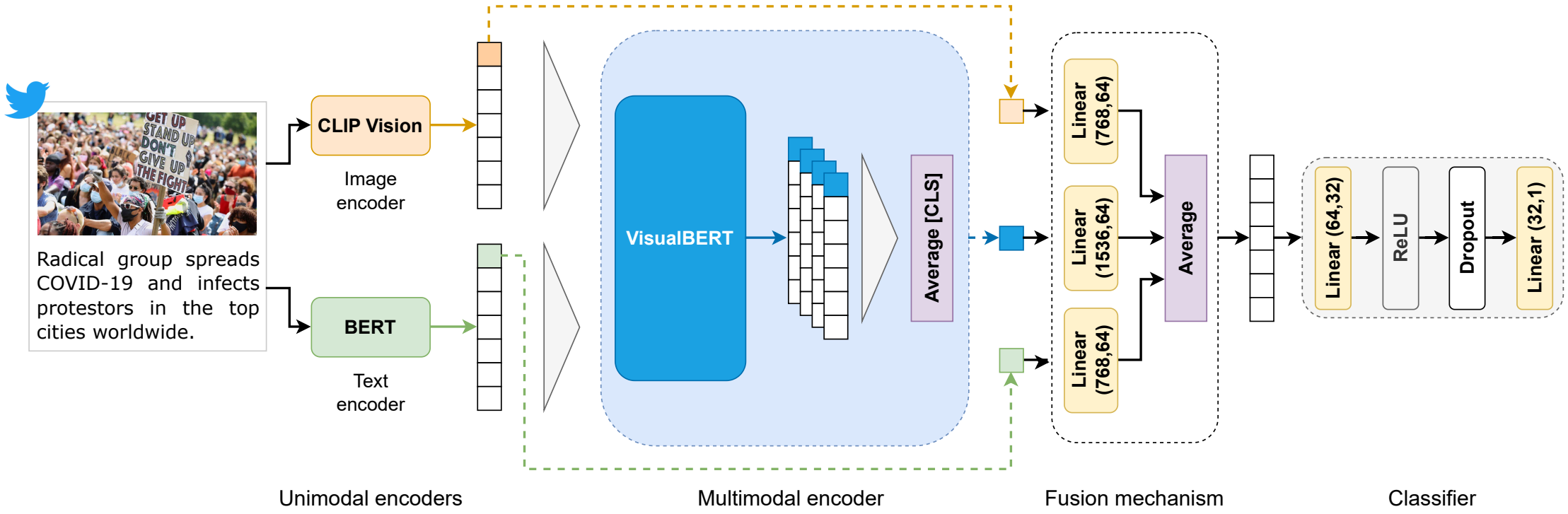


(b) Different detectors trained on five classes: real, Stable Diffusion, StyleGAN, StyleGAN2, and StyleGAN3



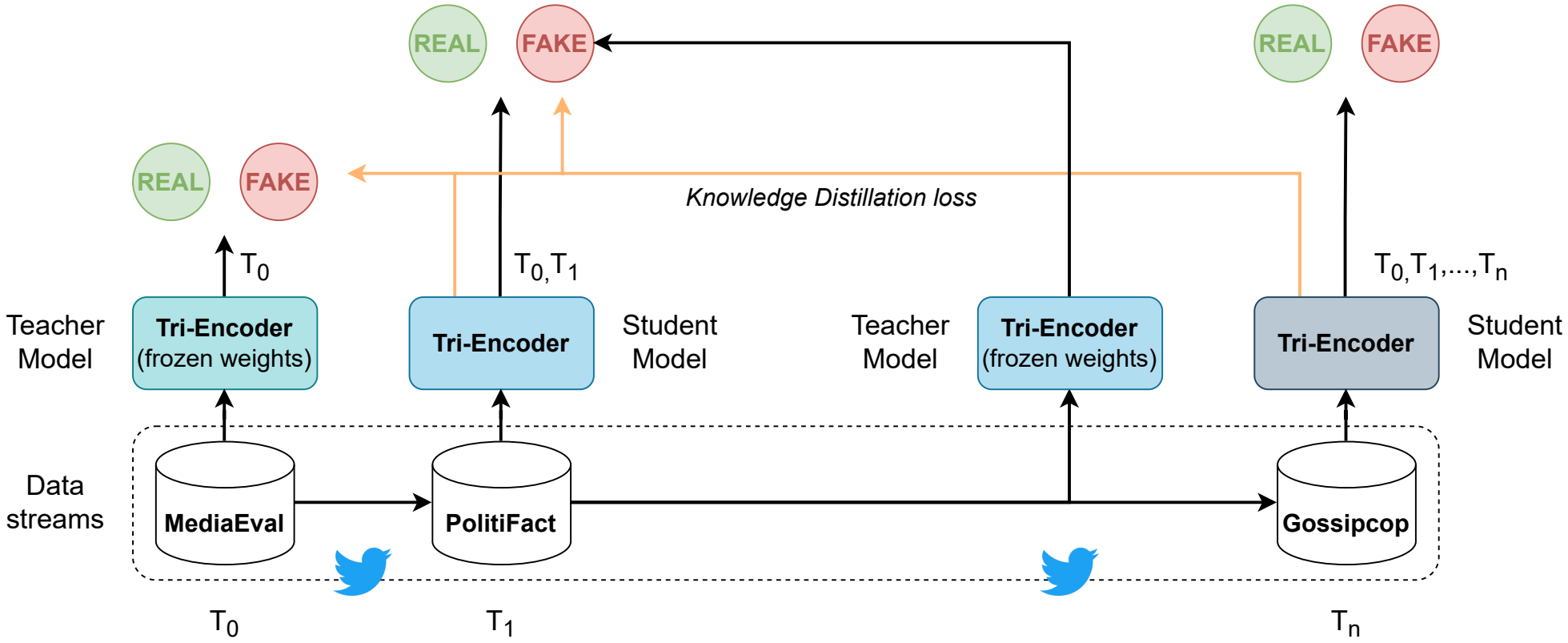
(c) Human performances

Multimodal fake news detection with continual learning



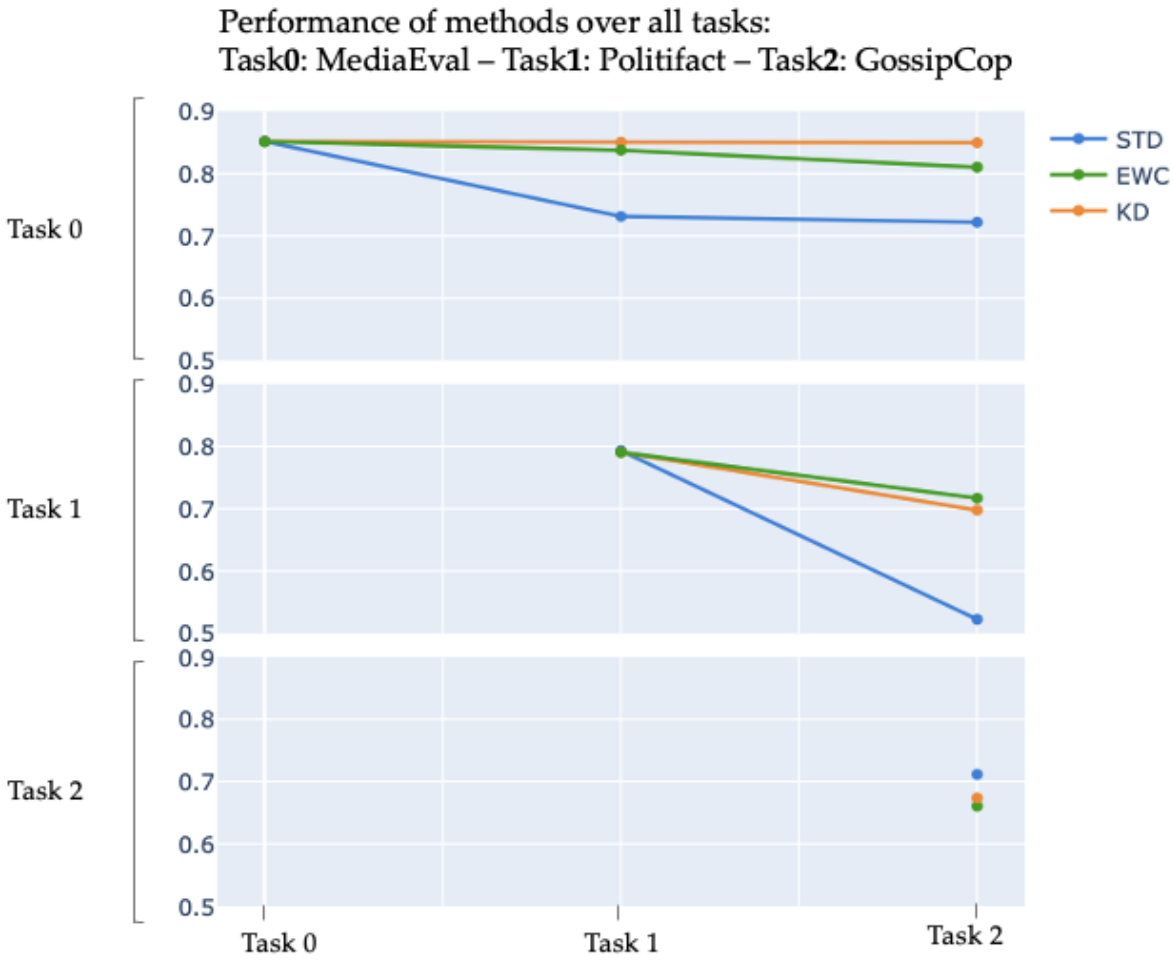
Maiano L., et al., Multimodal fake news detection through continuous learning: a content-based strategy, ongoing work (2023)

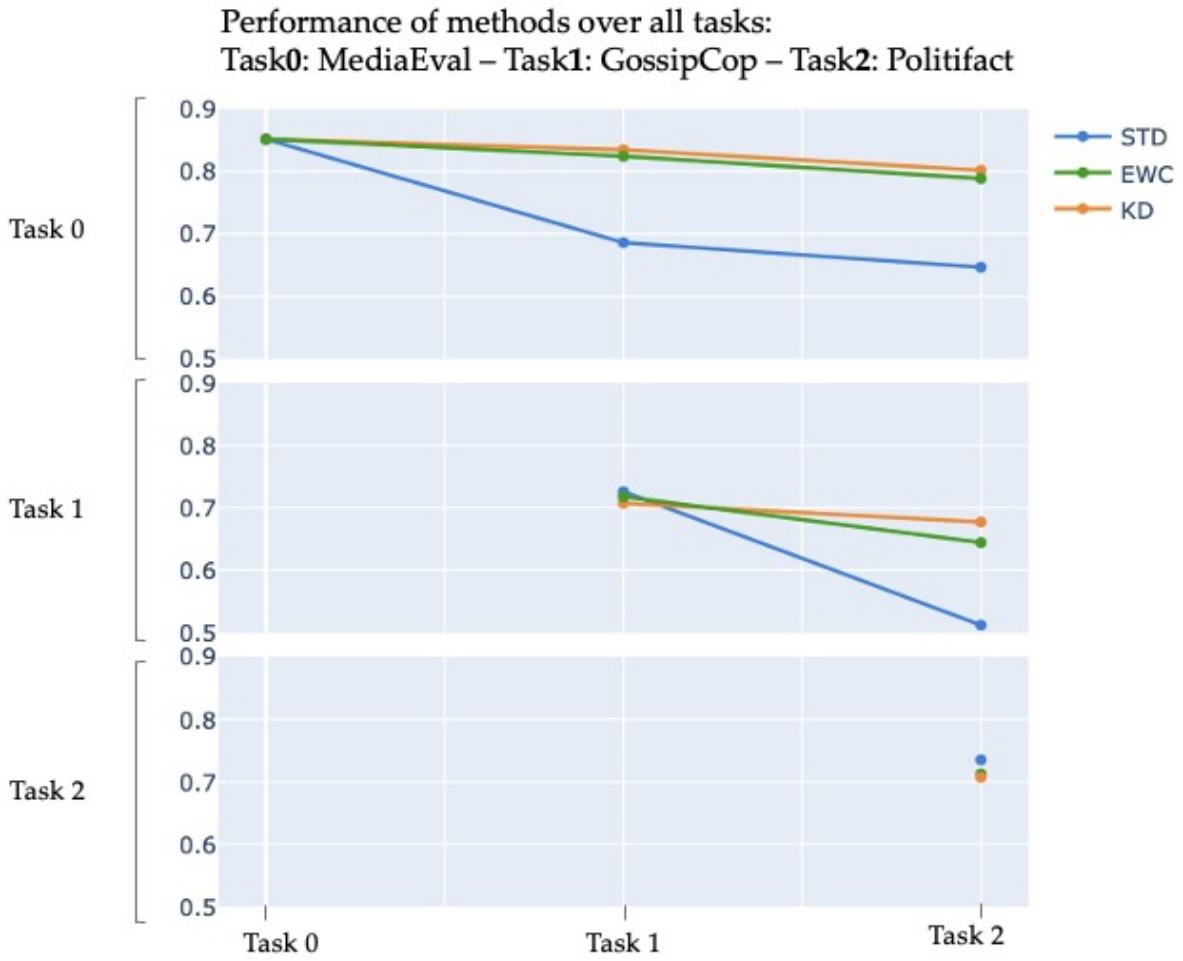
Multimodal fake news detection with continual learning



Maiano L., et al., Multimodal fake news detection through continuous learning: a content-based strategy, ongoing work (2023)

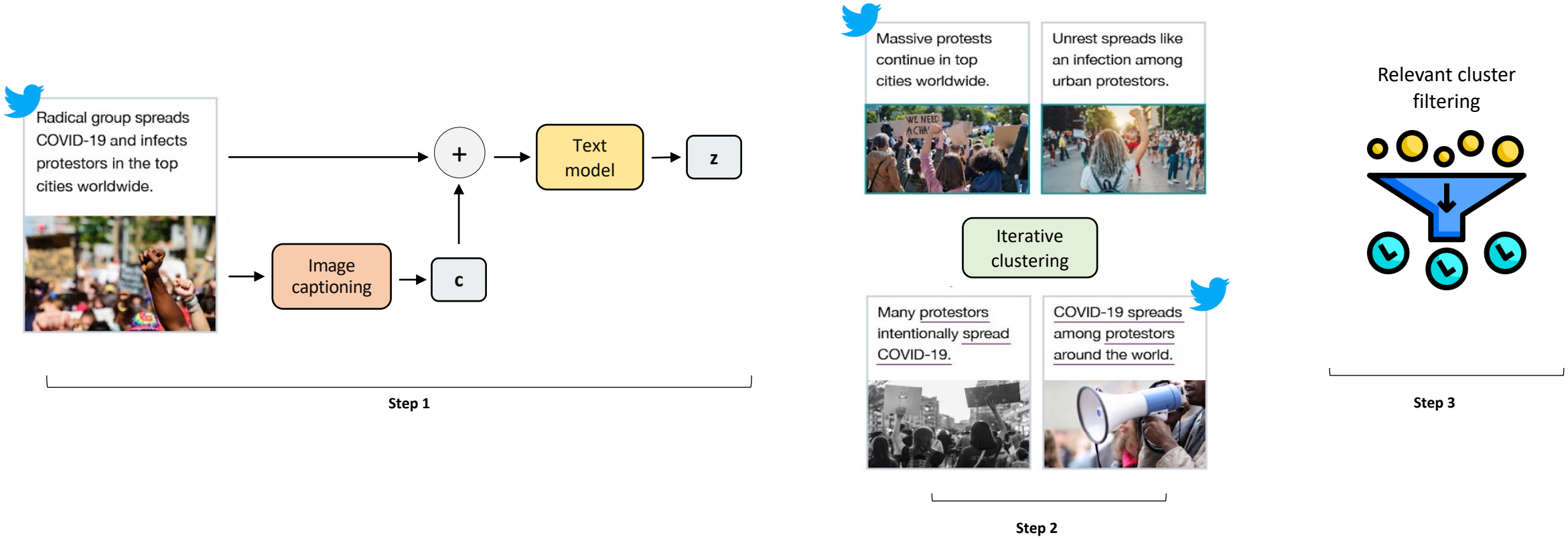
Multimodal fake news detection with continual learning





Maiano L., et al., Multimodal fake news detection through continuous learning: a content-based strategy, ongoing work (2023)

Multimodal topic filtering



Ongoing work for multimodal topic filtering. Step 1 projects text and images in a multimodal embedding. Step 2 clusters the topics iteratively. Step 3 filters relevant topics and removes outliers applying reinforcement learning with human feedback

Master thesis topics

- Continual learning
- Multimodal learning
- Self supervised learning
- Explainable AI
- Adversarial machine learning

Thank you!

QUESTIONS?