Image Forgery and DeepFake Detection

Deep Learning for Image Generation 2023
Lisa Artmann, Lisa Golla & Benjamin Peters

Overview

- Introduction
- What are the different kinds of manipulations?
  - Types of Image Forgery
  - Types of DeepFakes
- How can they be detected?
- Datasets and Evaluation Metrics
- State of the Art Detection Approaches
  - Image forgery
  - DeepFakes
- Future Challenges

Introduction

Misinformation

Manipulation

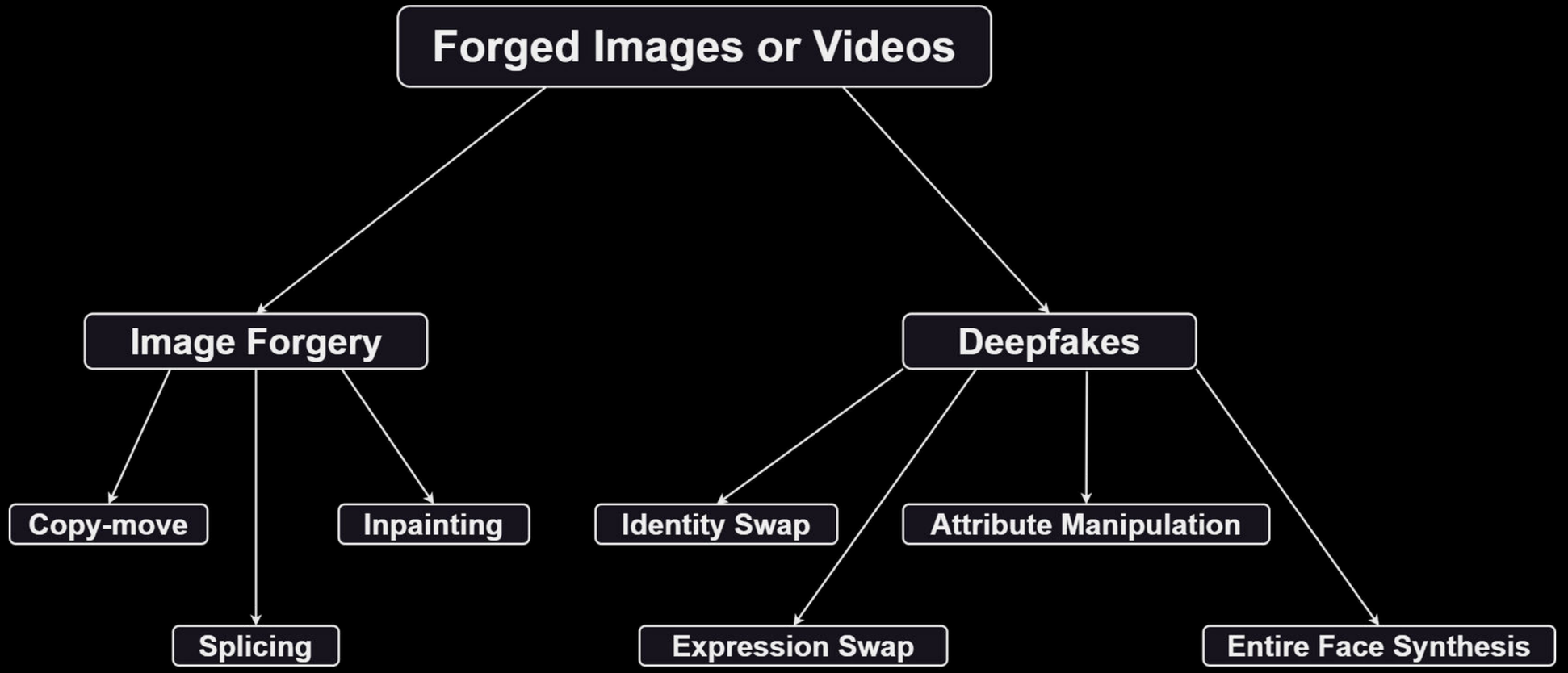
Credibility

Privacy

Altered version of reality

How to detect what is real? [5,6]

Types of image manipulation



Types of Image Forgery

Copy-move

Copies a part of the image and moves it to another place in the same image

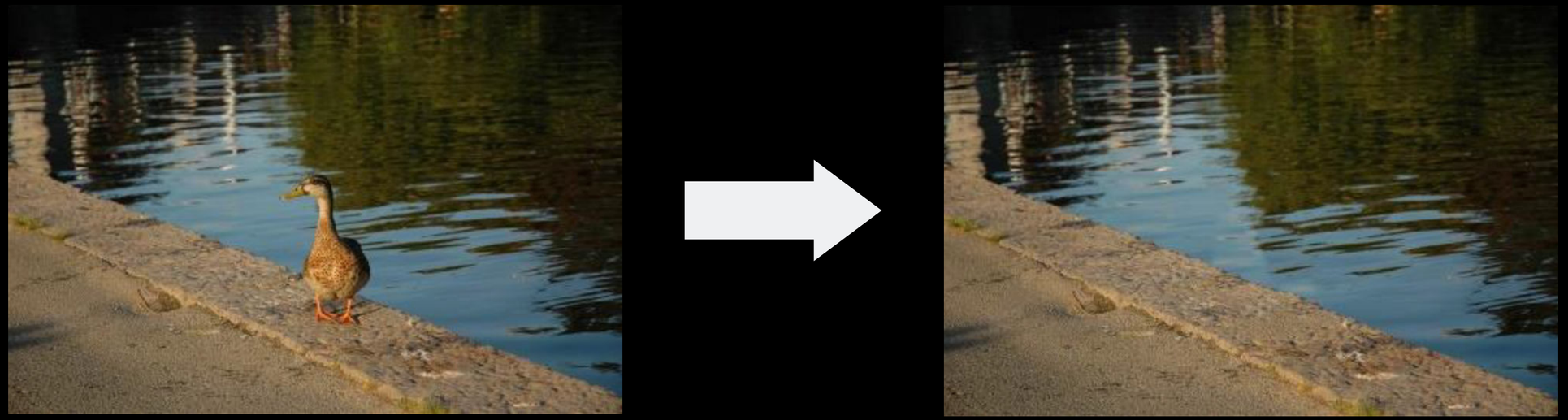
[7] [Zanardelli et al. (2022)]

Splicing

[3]

Image Forgery

Inpainting

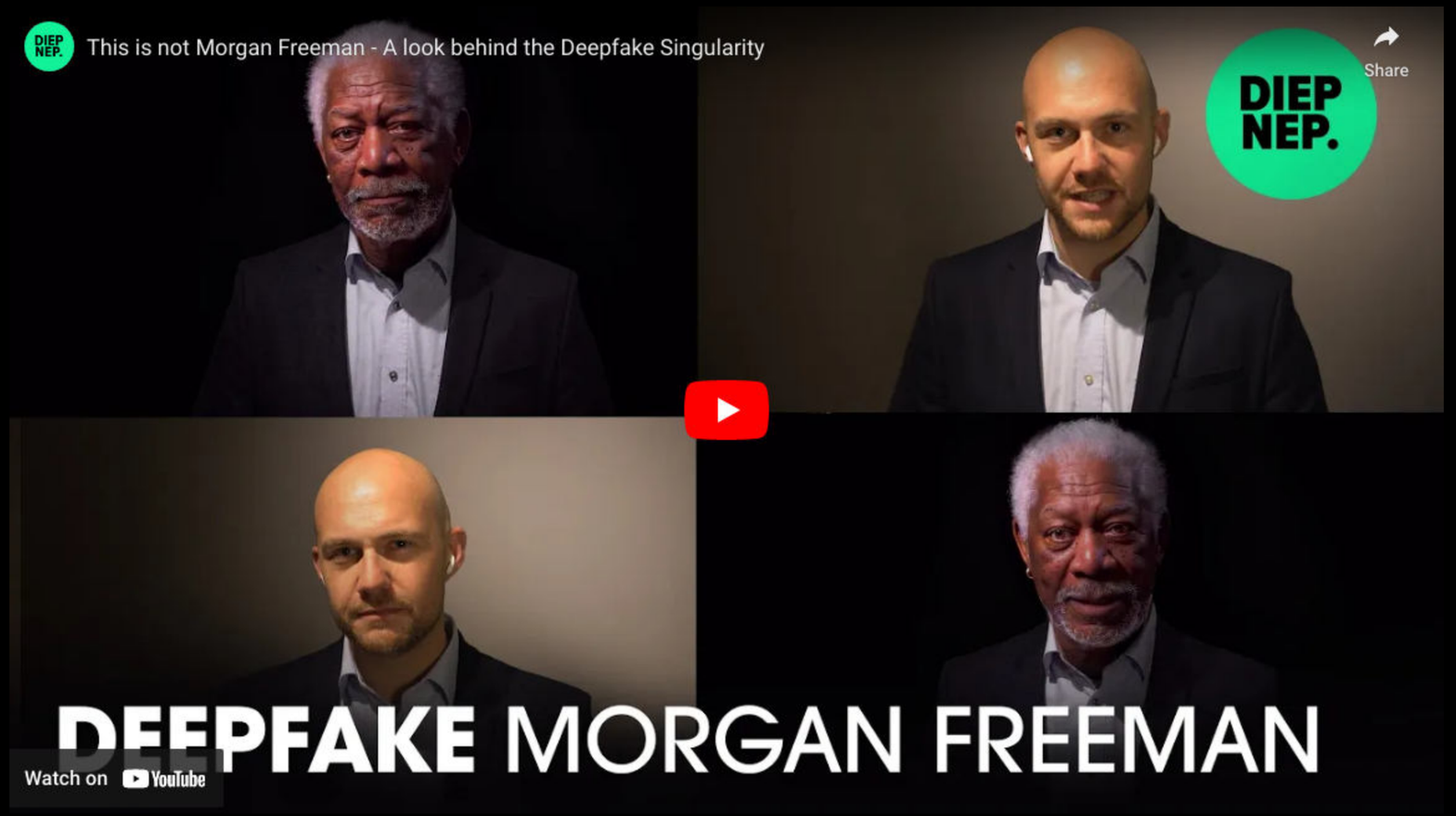


[4]

Types of DeepFakes

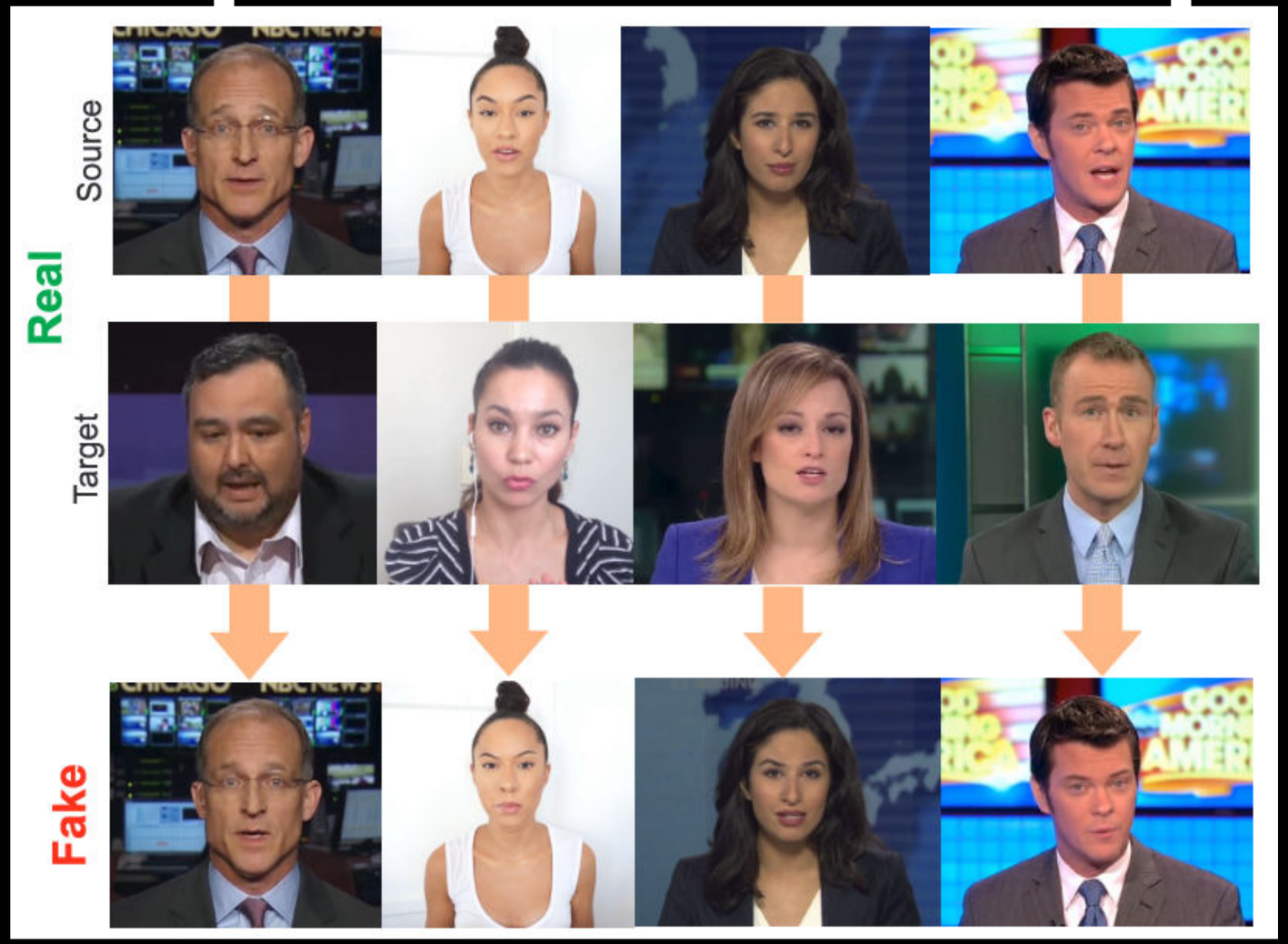
DeepFake

Identity Swap



DeepFake

Expression Swap



[Tolosana et al. (2020)]

DeepFake

Attribute Manipulation

[Tolosana et al. (2020)]

DeepFake

Face Synthesis

[Karras et al. (2020)]

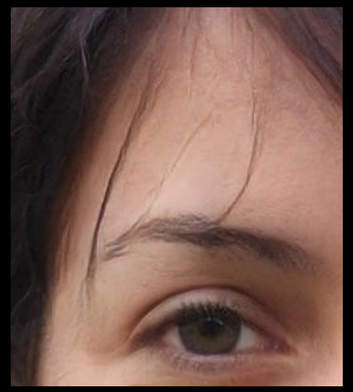
How can they be detected?

Can you decide which image is fake?

You are **incorrect**. The image on the right is the real one.

Play again.

... for humans?



Water sploches

Some GANS are producing water splotches between background and hair which is a clear indicator

Background problems

it can happen that the background is generated weirdly, as well as other people in the background

Asymmetries

Are the glasses symmetric? How about facial hair, facial expressions or earrings?

Teeth & Hair

Teeth & Hair are not easy to render, they might me assymetric, or weird looking

[8]

… for algorithms?

Image Forgery

Lighting-based detection
Geometry-based detection

DeepFake

Frequeny-based detection
Biological signal based detection

Spatial-based detection

[Juefei-Xu et al. (2022)]
[Zanardelli et al. (2022)]

Image Forgery detection

Lighting-based

- Find inconsistencies in lighting

Geometry-based

- Look for inconsistencies in the geometric properties of the 3D scene

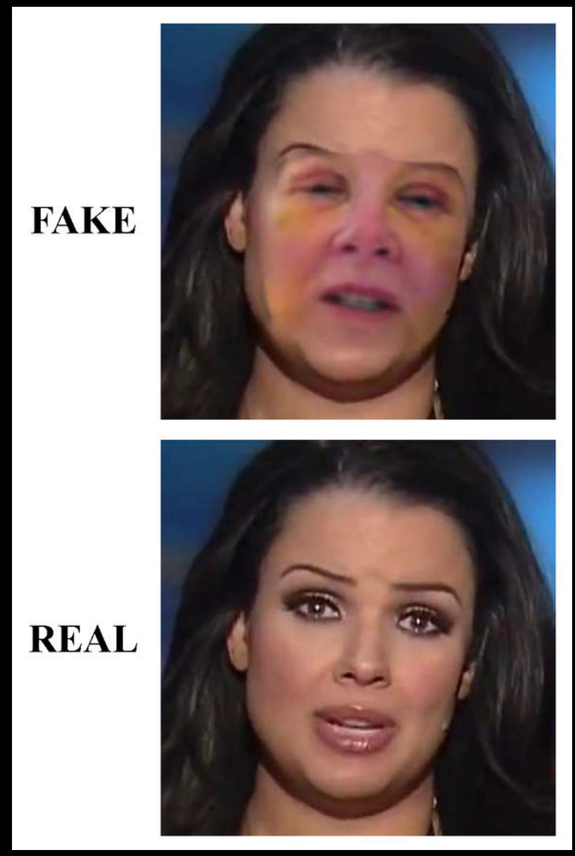
[Zanardelli et al. (2022)]

Spatial-based detection

- Observe visible and invisible artifacts in the spatial domain
- Mostly on pixel-level
  - Chrominance components
  - PRNU pattern
- Currently most popular technique for DeepFake detection

Criticism:

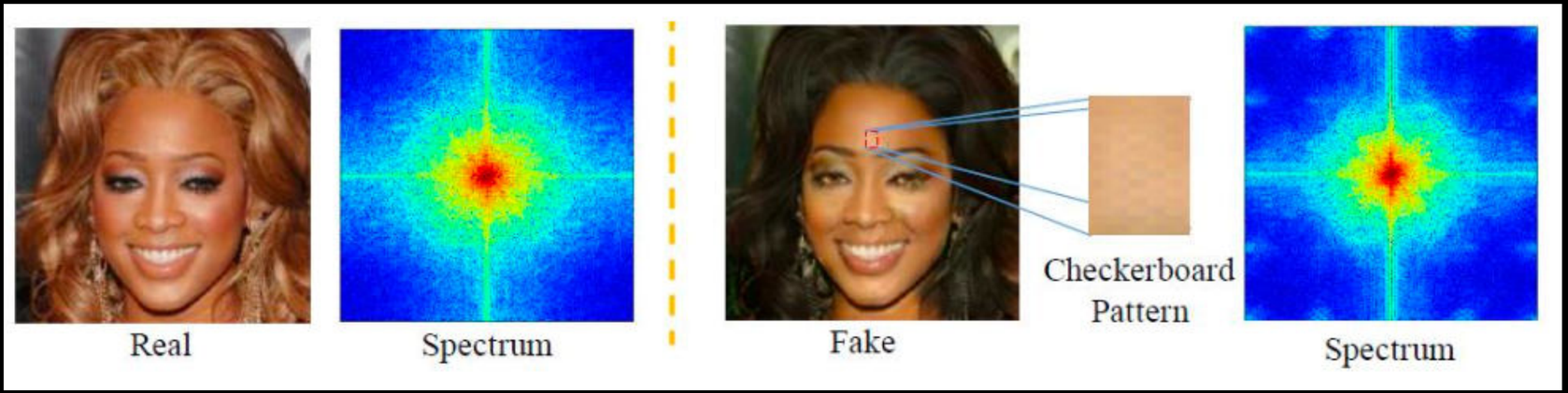
- Poor generalization against unknown generation techniques
- Low robustness to adversarial noise attacks and permutation attacks



[Juefei-Xu et al. (2022)]

Frequency-based detection

- Exploits...
  - artifacts on the frequency level introduced by GANs (GAN fingerprint)
  - difference between frequency domain features of real and fake faces
- Criticism:
  - Fingerprint can be destroyed by simple permutation attacks (blur or JPEG compression)



[Juefei-Xu et al. (2022)]

Biological signal-based detection

- Audio-visual inconsistency (e.g. lip-sync inconsistency)
  - Simple lib-sync is not enough for accurate DeepFake detection
- Visual inconsistency
  - Non-natural properties of synthesized faces (shape, facial features, ...)
  - Repetitive actions tend to disappear in DeepFakes (e.g. eye blinking)
  - Subtle biological characteristics tend to be corrupted in DeepFakes
    - E.g. heart rate (difficult to extract from a video)

- Criticism:
  - A lot of biological signals could be improved in future GANs
  - Many of them work only for videos

[Juefei-Xu et al. (2022)]

DeepFake videos

Clip or frame-by-frame?

Is the temporal dimension important for the detection approach used?

- Spatial-based and frequency-based:
  - In general clip approaches do **not** outperform frame-by-frame
- Biological signal-based:
  - Multiple coherent frames are often needed depending on the feature that's analyzed

[Juefei-Xu et al. (2022)]

Evolution of DeepFake Detection

2017.08

Spatial-based detection

Using DNN models to investigate pixel artifacts

2018.06

Biological signal-based detection

Inspecting visual, audio-visual and other biological signals

2019.03

Frequency-based detection

Investigating GAN-fingerprints and the image frequencies

2021.10

Vision Transformers

Introduction of Vision Transformers for DeepFake detection

[Juefei-Xu et al. (2022)]

Image Forgery

Datasets and Benchmarks

Dataset name	Number of original/forged images	Image size	Type(s) of forgery	Format
CASIA2	7491 / 5123	320 x 240 - 800 x 600	Copy-move & splicing	JPG, BMG & TIFF
MICC-F2000	1300 / 700	2048 x 1536	Copy-move	JPG
CoMoFoD	260 / 260	512 x 512 3000 x 2000	Copy-move	JPG & PNG
DVMM	933 / 912	128 x 128	Splicing	BMP (grayscale)

[Zanardelli et al. (2022)]

DeepFake

Datasets and Benchmarks

Dataset name	Number of original/fake videos	Video resolution	Real / deepfake video source	Format
FaceForensic++	1000 / 1000	480p, 720p & 1080p	YouTube / Manually crafted	MP4
DFDC	Full: ca. 20k / 124k Preview: 1131 / 4113	180p - 2160p	Volunteer actors / Manually crafted	MP4
Celeb-DF	590 / 5639	Various	YouTube / Manually crafted	MP4
WildDeepfake	3805 / 3509	Various	Internet / Internet	MP4

[Zanardelli et al. (2022)]

Evaluation Metrics

- Performance metrics are the same as for other binary classification tasks.
  - Terminology: True Positives (TP), False Positives (FP), True Negatives (TN) & False Negatives (FN); Total number of queries (T)

- Commonly used evaluation metrics:
  - Precision: TP / (TP + FP)
  - Recall: TP / (TP + FN)
  - F1 score: 2*Precision*Recall / (Precision + Recall)
  - Accuracy: (TP + TN) / T
  - AUC (Area Under the ROC Curve)

[Zanardelli et al. (2022)]

State of the art approaches

What are we looking for?

Critical factors for image forgery and DeepFake detection methods:

1.) Generalization to unseen synthesized forgery techniques

2.) Robustness against variety of adversarial attacks and image / video transformations

3.) Explainable detection results

[Juefei-Xu et al. (2022)]

Forensic Similarity for Digital Images

- Paper published by Owen Mayer & Matthew C. Stamm (2019)
- Digital image forensics approach called **forensic similarity**
- **Goal**: determine whether or not two image patches contain the same forensic trace
  - Source camera model, processing history etc.
- **Improvements on prior work**:
  - Model does not require prior knowledge of forensic trace in order to make a similarity decision
  - Model is not limited to specific forensic traces
    - Focuses on forensic consistency of an image

[Mayer & Stamm (2019)]

Image Forgery

Forensic Similarity for Digital Images

- Employs a two-part deep learning system
  - **Feature extracto**r: $f : \mathbb{X} \to \mathbb{R}^N$
  - **Similarity network:** $S : \mathbb{R}^N \times \mathbb{R}^N \to [0, 1]$
- 94% overall accuracy when comparing image patches from 25 different camera models
  - **BUT**: very low accuracy for certain pairs of camera models
- **Practical applications**:
  - Forgery detection and localization
  - Database consistency verification

[Mayer & Stamm (2019)]

Image Forgery

Forensic Similarity - Architecture

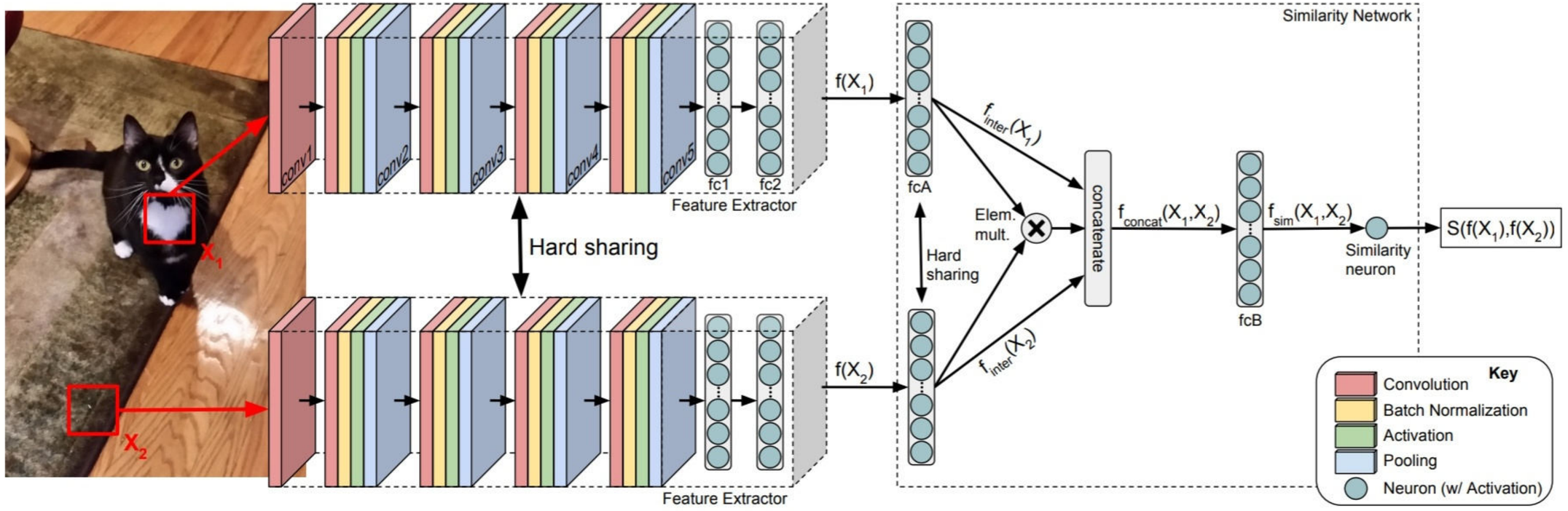


Fig. 3. The neural network architecture of the proposed forensic similarity system. The system is composed of a pair of CNN-based feature extractors, in a hard sharing (Siamese) configuration, which feed low-dimensional, high-level forensic feature vectors to the similarity network. The similarity network is a neural network that maps feature vectors from two image patches to a similarity score indicating whether they contain the same or different forensic traces.

Mayer & Stamm (2019): 1335

Forensic Similarity - Detection Process

Feature extractor:

- MISLnet CNN architecture
  - *Siamese* configuration (hard-sharing)
  - 5 convolutional blocks & 2 fully connected layers
- Trained on 2 million image patches from 50 camera models

- Neuron activations from the last fully connected layer represent **deep features**
  - Encode high-level information about about an image patch *X*
  - Representations are able to generalize to other forensic tasks

[Mayer & Stamm (2019)]

Forensic Similarity - Detection Process

Similarity network:

- Consists of 3 layers of neurons
  - Hierarchical mapping of two input feature vectors to an output score indicating the forensic similarity

1.) *Siamese* fully connected layers with 2048 neurons

- Map feature vectors to an intermediate feature vector space via an artificial neuron function:

$$f_{k,inter}(X) = \phi\left(\sum_{i=1}^{N} w_{k,i} f_i(X) + b_k\right)$$

[Mayer & Stamm (2019)]

Forensic Similarity - Detection Process

Similarity network:

2.) Fully connected layer with 64 neurons

- Input is given via concatenation of intermediate feature vectors:

$$f_{concat}(X_1, X_2) = \begin{bmatrix} f_{inter}(X_1) \\ f_{inter}(X_2) \\ f_{inter}(X_1) \odot f_{inter}(X_2) \end{bmatrix}$$

- Maps to a similarity feature space $f_{sim}(X_1, X_2)$
  - Encodes information about the relative forensic information between the image patches

[Mayer & Stamm (2019)]

Image Forgery

Forensic Similarity - Detection Process

Similarity network:

3.) *Similarity neuron*

- Maps similarity vector to a single score via threshold η:

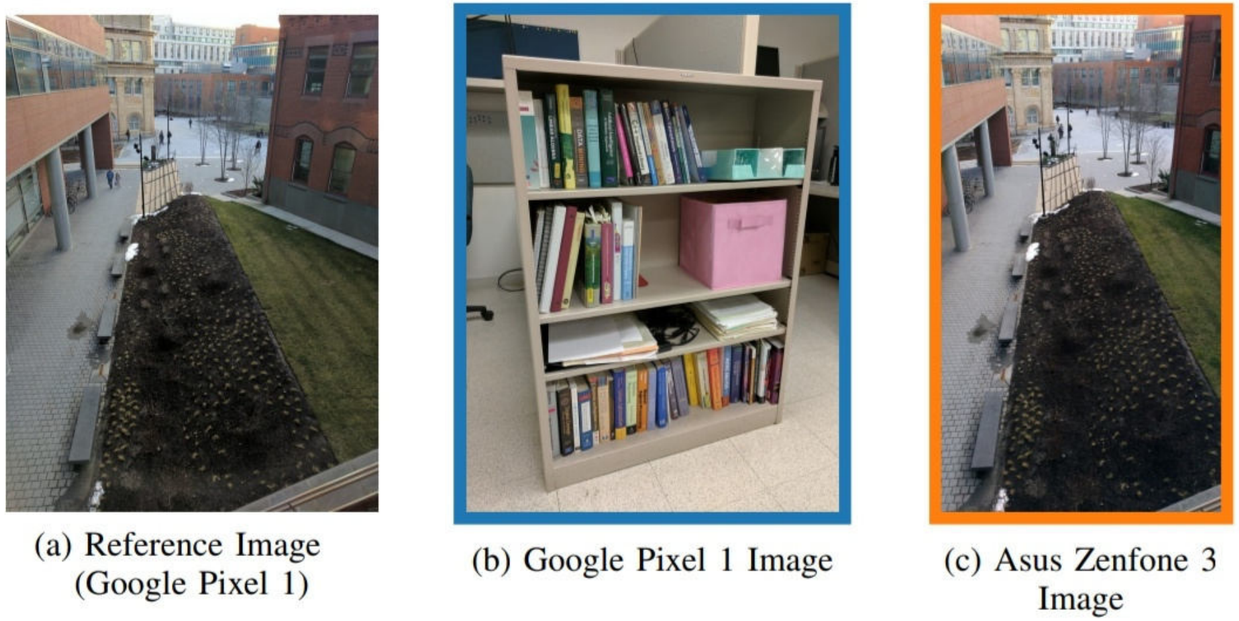
$$C(X_1, X_2) = \begin{cases} 0 & \text{if } S(f(X_1), f(X_2)) \leq \eta \\ 1 & \text{if } S(f(X_1), f(X_2)) > \eta \end{cases}$$

- Similarity network is trained with data from 30 camera models
  - Also allows backpropagation through feature extractor layers

[Mayer & Stamm (2019)]

Image Forgery

Forensic Similarity - Examples



(a) Reference Image (Google Pixel 1)

(b) Google Pixel 1 Image

(c) Asus Zenfone 3 Image

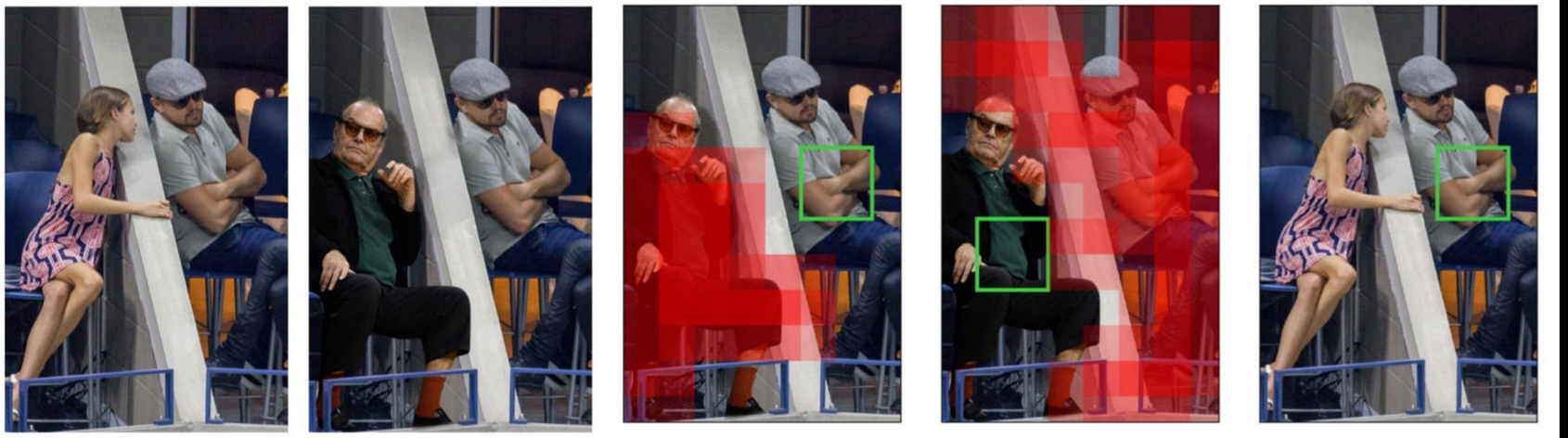
(d) Forensic Similarity Distribution

Fig. 2. Camera-model based forensic similarity scores measured between random pairings of patches from the reference image (a), captured by a Google Pixel 1 smartphone rear-facing camera, and from (b) a different Google Pixel 1 image or (c) a Asus Zenfone 3 image of the same scene as the reference image. The histogram (d) shows the distribution of these forensic similarity scores. Forensic similarity between patches from the two Google Pixel images is high, even though the two images depict different scenes and content, and were captured in different lighting environments. The forensic similarity between patches from the Google Pixel and Asus Zenfone cameras is low, even though the two images depict similar content.

Mayer & Stamm (2019): 1334

Image Forgery

Forensic Similarity - Examples



(a) Original (b) Spliced (c) Host reference (d) Spliced reference (e) Host ref, original image

Mayer & Stamm (2019): 1344

(a) Spliced Image (b) Host reference, 256x256 (c) Host reference, 128x128

Fig. 11. Splicing detection and localization example. The green box outlines a reference patch. Patches, spanning the image with 50% overlap, that are detected as forensically different from the reference patch are highlighted in red.

DeepFake

Most common Architectures

Images and Videos

- 2D CNN
- Vision Transformer
- SVM, KNN, Random Forest, MLP
- Pre-trained models for image classification (e.g.ResNet)

Just Videos

- RNN, LSTM
- 3D CNN

Multi-attentional DeepFake Detection

What is the idea behind the model?

The model uses a spatial-based detection with an multi-attentional approach.

Why this model?

- Spatial-based detection is widely used in state of the art approaches
- Good performance on benchmark datasets

[Zhao et al. (2021)]

About the model

- Trained on FaceForensic ++ dataset
- Uses Multi-Attentional approach
  - Creates attention maps to make the model attend to different local parts
  - Fine-grained classification

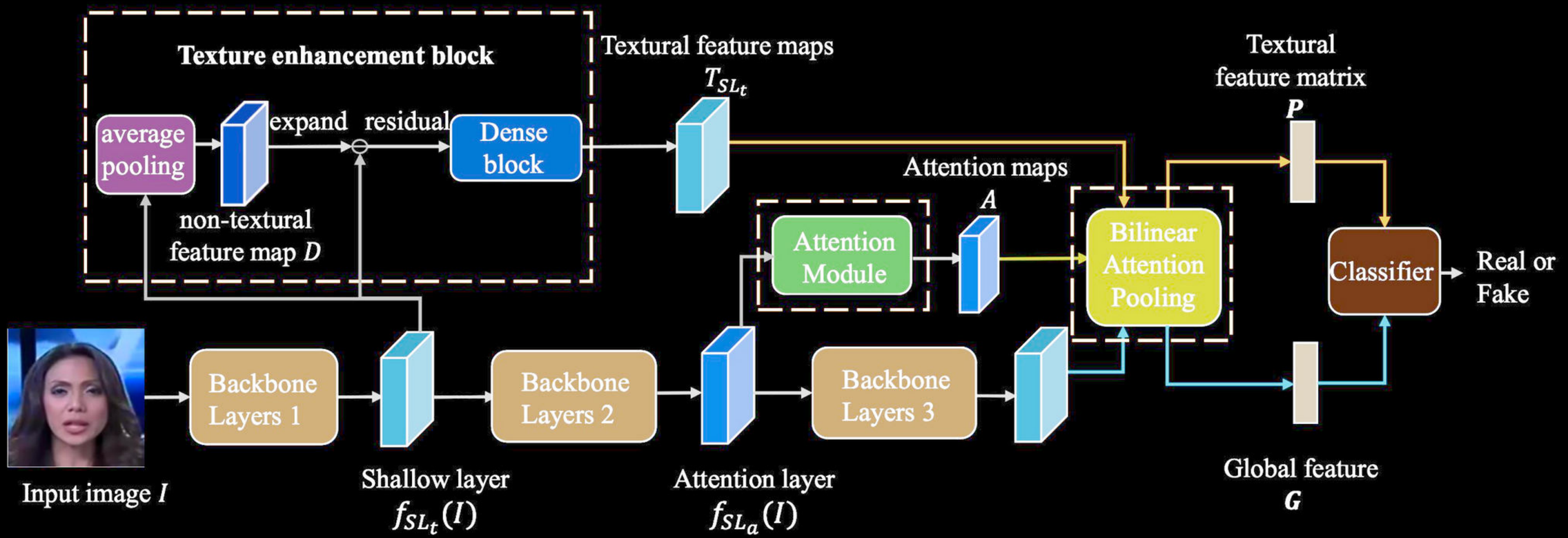
Performance

- 97.6% accuracy on FaceForensic ++
- 67.4% accuracy on Celeb-DF (Cross-dataset evaluation)

[Zhao et al. (2021)]

DeepFake (1)

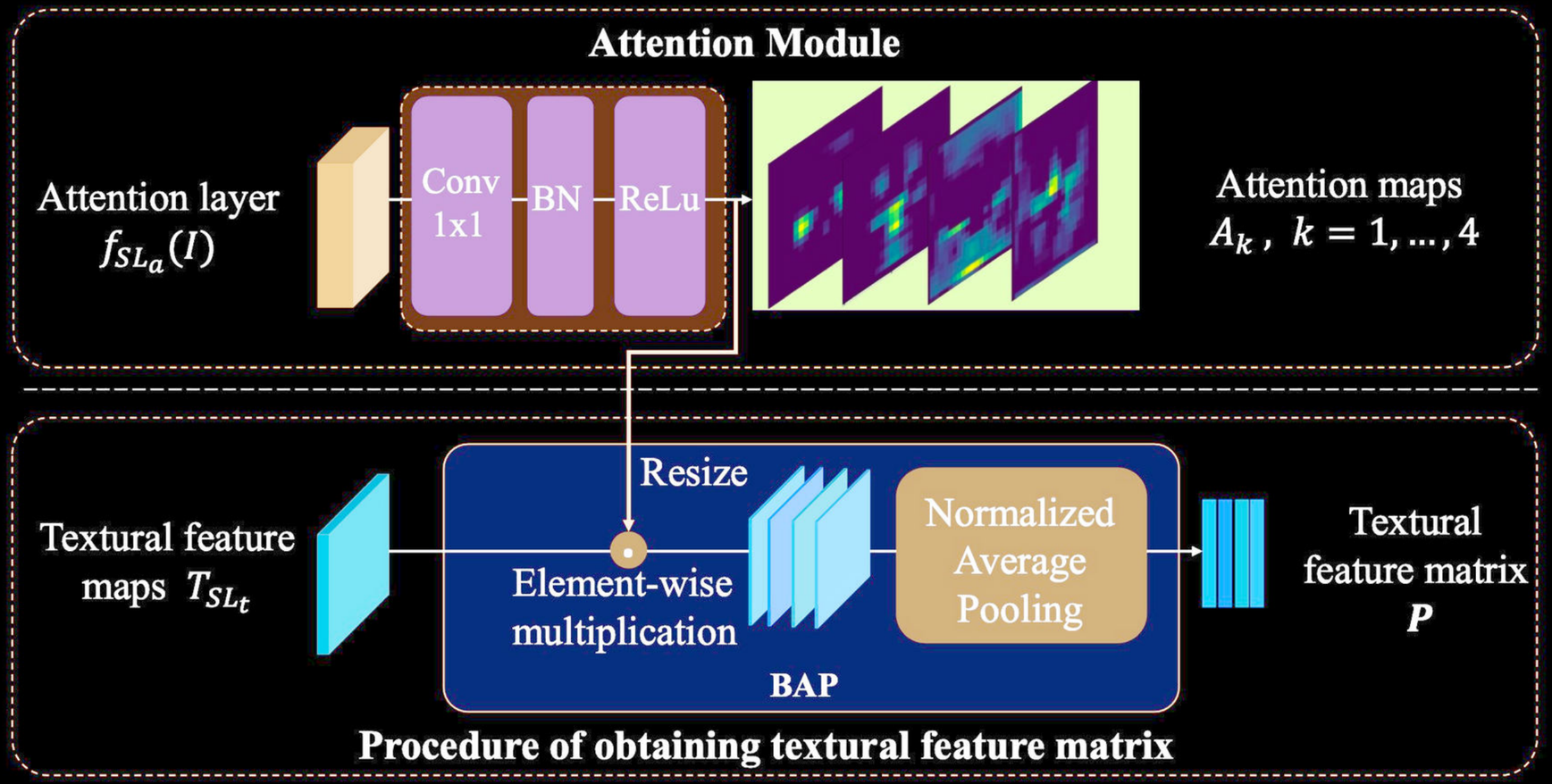
Overview



[Zhao et al. (2021)]

DeepFake (1)

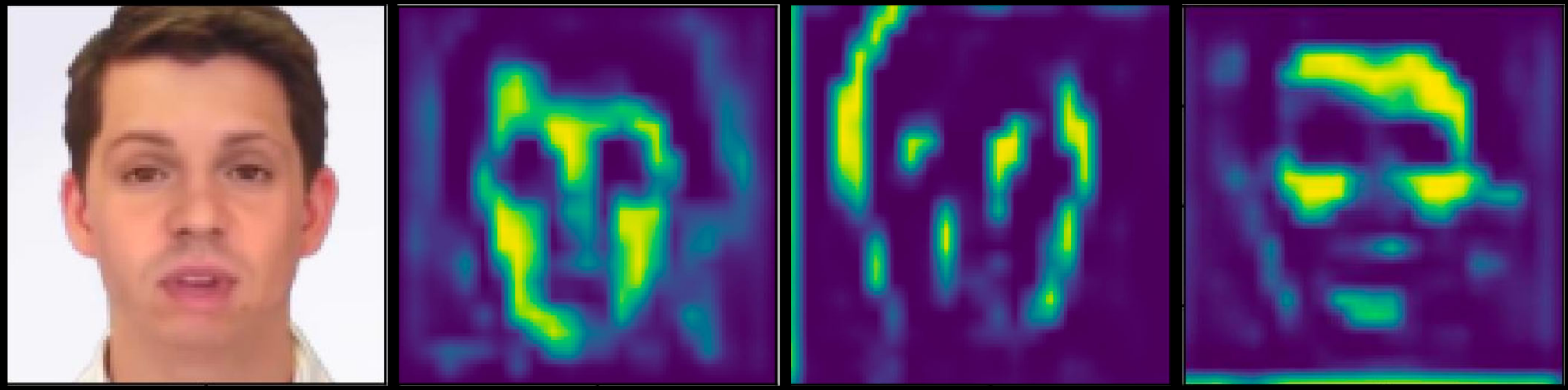
Attention Module



[Zhao et al. (2021)]

DeepFake (1)

Attention Maps



[Zhao et al. (2021)]

DeepRhythm for DeepFake Detection

What method is the model using?

- biological signal-based detection method

Why this model?

- In the near future, the DeepFake could be realistic where the spatial and frequency based detection methods could hardly exhibit noticeable and detectable artifacts by human eyes and machines.

→ the biological signals would be a more effective solution for fighting against DeepFake that could be deployed in the real world.

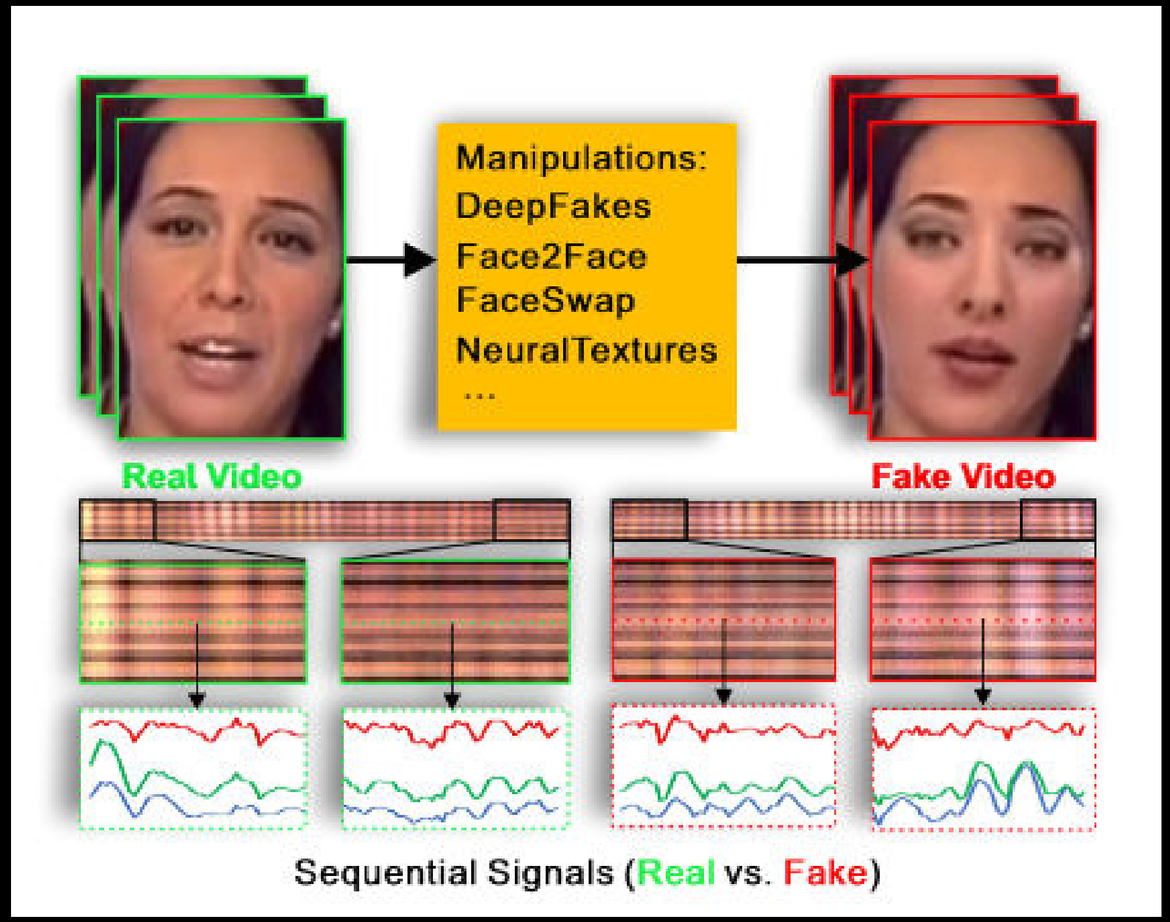
[Qui et al. (2020)]

DeepFake (2)

DeepRythm

Idea

- **Remote visual photoplethysmography (PPG)** is made possible by monitoring the minuscule periodic changes of skin color due to blood pumping through the face
- **Dual-spatial-temporal attention** to adapt to dynamically changing face and fake types
- Manipulations easily diminish the sequential signals representing remote heartbeat rhythms

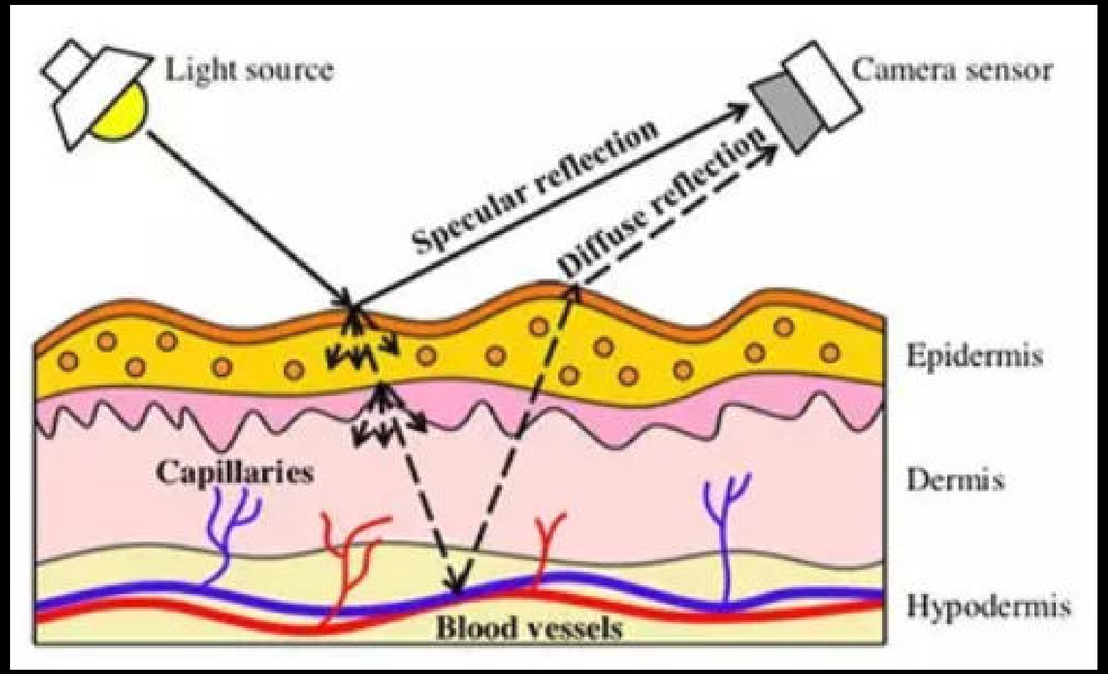


[Qui et al. (2020)]

Remote PPG

Variance of red, green, and blue light reflection changes from the skin, as the contrast between specular reflection and diffused reflection.

- Specular reflection is the pure light reflection from the skin.
- Diffused reflection is the reflection that remains from the absorption and scattering in skin tissue, which varies by blood volume changes.



[Qui et al. (2020)]

DeepFake (2)

[Qui et al. (2020)]

$$V = \{I_i\}_{i=1}^{T}$$

MMSTR – motion-magnified spatial-temporal representation

- Highlight the heart rhythm signals and output a motion-magnified spatial-temporal map (MMSTmap)

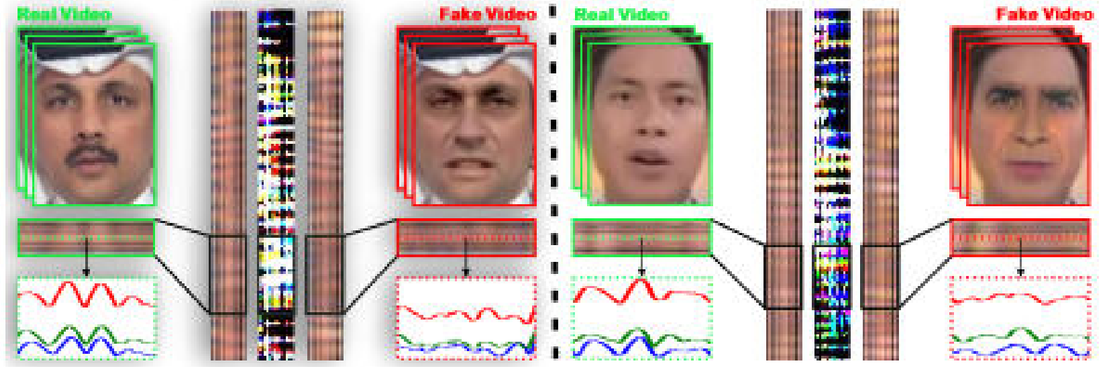


Figure 3: Two real-fake video pairs, their MMST maps and the colorful difference maps between real and fake.

→ characterize the sequential signals of face videos

$$X = mmstr(V) \epsilon \mathbb{R}^{T \times N \times C}$$

DeepFake (2)

various interference, e.g., head movement, illumination variation, and sensor noises, may corrupt the MMST map

$$y = \phi(A \odot X)$$

$$A \in \mathbb{R}^{T \times N}$$

Provides different weights to different positions of X and is known as an attention mechanism.

assign different weights to different positions of the MMST map before further performing the fake detection

$$y = \phi((t \cdot s^T) \odot X)$$

Decomposing A into 2 parts: **spatial and temporal attention** because it is hard to get a proper A

$$s \in \mathbb{R}^{N \times 1}$$

$$t \in \mathbb{R}^{T \times 1}$$

Key problem is how to generate t and s to adapt to dynamically changing faces and various fake types

[Qui et al. (2020)]

DeepFake (2)

Dual-Spatial-Temporal Attentional Network

Jointly considering prior & adaptive spatial attention and frame & block temporal attention

Dual spatial attention

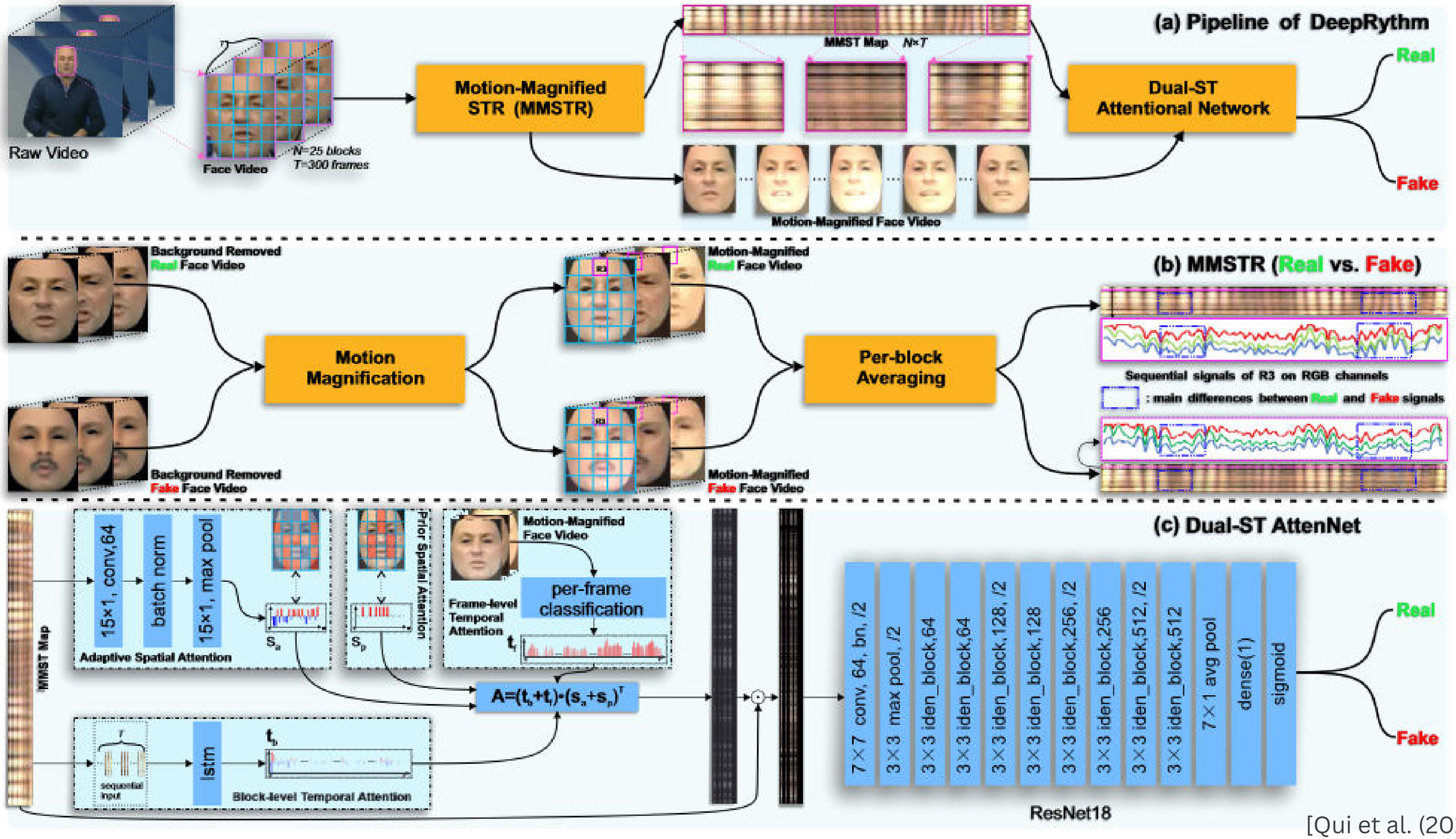
$$s = s_a + s_p$$

Realize accurate DeepFake detection through the MMST map and its spatial and temporal attentions

Dual temporal attention

$$t = t_b + t_f$$

[Qui et al. (2020)]



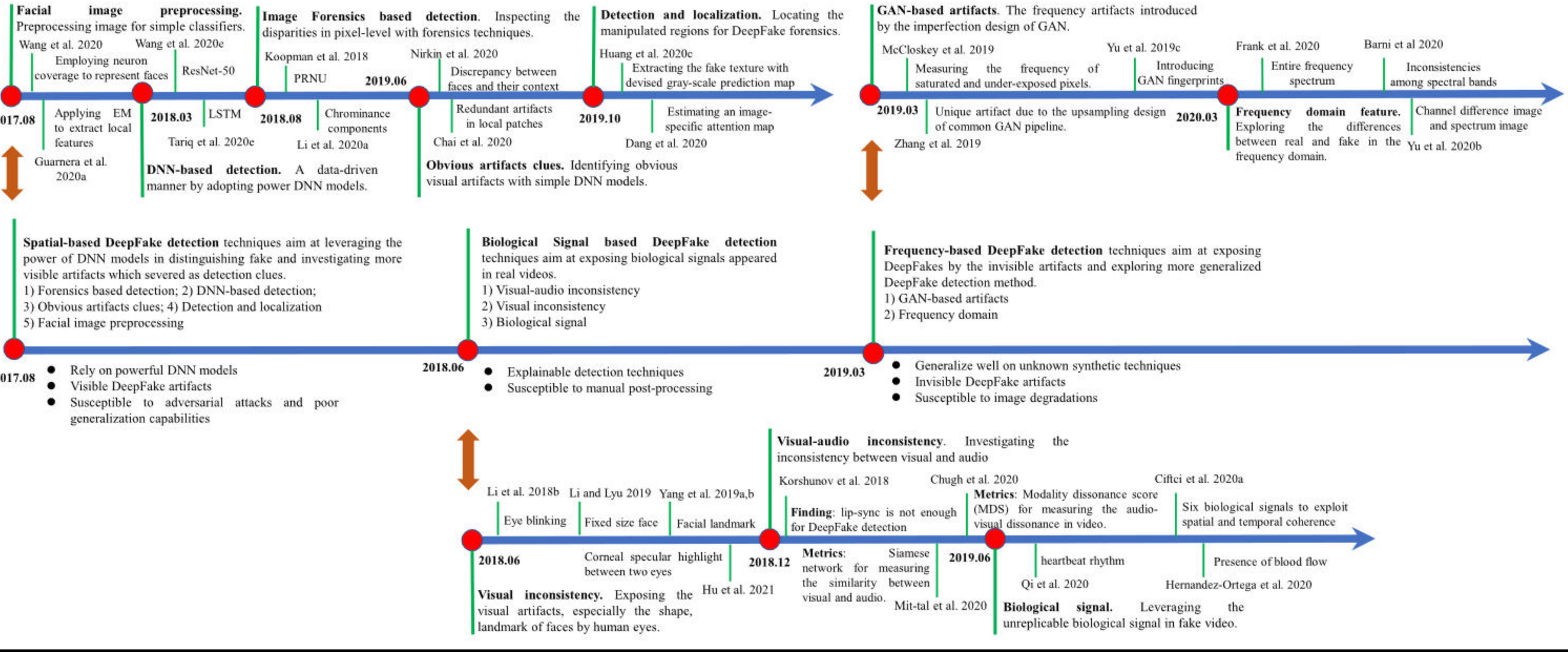
[Qui et al. (2020)]

What do you think? Will it ever be possible to create DeepFakes that can mimick the heartrate of humans properly?

DeepFake

Detection technique	Description	Benefits	Limitations
Detection based on visual artifacts.	Closed observation of deepfakes may reveal minor differences and inconsistencies between background and foreground.	Visible artifacts methods may detect inconsistency in the blending boundary of a modified object, such as image resolution.	These visible artifacts are rapidly diminishing as deepfake algorithms advance, demanding to exploit more intrinsic properties.
Detection based on GAN fingerprints.	Synthesized faces generated by GANs often have GAN generated fingerprints.	By the use of deep features, fingerprints identified in the GAN generated fake images.	Because of different versions of GANS, no universal fingerprint metric can be adopted.
Detection based on biological signals, such as eye blinking.	There can be abnormalities in the eye blinking frequency in deepfakes.	Synthetic biological signals are easier to detect, such as eye blinking and heartbeat.	Advanced versions of deepfakes face generation very precisely model the biological signals, making it harder to detect.
Detection based on adjacent video frames' continuity.	Deepfakes may have flickering, jittering, and different face positions due to the discontinuity among adjacent video frames.	Temporal consistency-based detection methods can recognize discontinuity in adjacent video frames.	Poor performance on low-quality videos as continuity between adjacent frames is affected by video compression.
Detection based on face emotions.	Alignment of facial emotions is improper on swapped faces in deepfakes.	Siamese network-based architecture can detect non-alignment of facial emotions by facial and audio features extraction.	This technique fails if the video has no emotions.
Detection based on out of lip-synced videos.	A deepfake video with synthesized audio may have out-of-sync lips.	The difference between visemes (mouth shapes) and spoken phenomes (utterances) is used for out-of-sync lips detection.	Improved GANs can now generate proper lip-synced deepfakes. Also, any out-of-sync video does not need to be deepfake.
Multimodal detection technique.	Deepfakes created by swapping the face and audio, same as Type IV in Figure 1, are known as multimodal deepfakes.	Multimodal detection techniques first detect swapped faces, then use lip-syncing techniques to identify manipulated speech.	Accuracy suffers as detection techniques extract audio from different datasets instead of the same deepfake video.

[Salmon et al. (2023)]



[Juefei-Xu et al. (2022)]

Future Challenges

- "Arms race" between forgery generation and detection methods
  - Competition between *adversaries* (generation) and *defenders* (detection)
- Improving the generalization & robustness of detection methods
  - Adversarial noise attacks, compression, low resolution etc.
- Diversification of existing DeepFake detection datasets
  - Use of multiple faces
  - Subcategorization of age, gender, ethnicity to prevent biases
- Creation of more fine-grained evaluation metrics
- Lack of commonly recognized evaluation datasets and baseline methods
- Incorporation of other modalities to counter malicious forgery attempts

[Zanardelli et al. (2022)]
[Juefei-Xu et al. (2022)]

Resources

- Zanardelli, M., Guerrini, F., Leonardi, R., & Adami, N. (2022). Image forgery detection: a survey of recent deep-learning approaches. Multimedia Tools and Applications, 1-46.
- Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. International Journal of Computer Vision, 130(7), 1678-1734.
- Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1), 1-41.
- Mayer, O., & Stamm, M. C. (2019). Forensic similarity for digital images. IEEE Transactions on Information Forensics and Security, 15, 1331-1346.
- Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional Deepfake Detection. https://doi.org/10.48550/ARXIV.2103.02406
- Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, Lei., Feng, W. (2020).DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. doi: https://doi.org/10.48550/arXiv.2006.07634
- Masood, M., Nawaz, M., Malik, K., Javed, A., Irtaza, A. (2021). "DeepFakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward". doi: https://doi.org/10.48550/arXiv.2103.00484
- Salman, S., Shamsi, J., Qureshi, R. (2023). "DeepFake Generation and Detection: Issues, Challenges, and Solutions". doi: 10.1109/MITP.2022.3230353
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. doi: https://doi.org/10.48550/arXiv.2001.00179
- Karras,T., Laine, S., Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. doi: https://doi.org/10.48550/arXiv.2001.00179
- Juefei-Xu, F., Wang, R., Huang, Y., Guo4, Q., Ma, L., Liu, Y. (2022)."Countering Malicious DeepFakes: Survey, Battleground, and Horizon" doi: https://doi.org/10.48550/arXiv.2103.00218

Image Resources

1. https://www.mdpi.com/2073-8994/14/12/2691
2. https://www.thetimes.co.uk/article/falling-faker-makers-online-videos-kim-kardashian-p5bjzlwsd
3. https://www.mdpi.com/2076-3417/13/3/1272
4. https://www.mdpi.com/electronics/electronics-09-00858/article_deploy/html/images/electronics-09-00858-g001.png
5. https://www.br.de/radio/bayern2/sendungen/zuendfunk/fail-of-the-week-ki-papst-problem-100.html
6. https://www.researchgate.net/figure/A-simple-example-of-copy-move-forgery-a-Original-image-b-Copy-move-image_fig1_334552767
7. https://link.springer.com/article/10.1007/s11042-022-13797-w
8. https://www.whichfaceisreal.com/learn.html