Video Forensics Beyond Deepfakes

Matthew C. Stamm
Multimedia & Information Security Lab
Drexel University
mstamm@drexel.edu

Video Forgeries

- Fake & manipulated video is important threat
- Deepfakes well studied
- Lots of other manipulations!
  - Splicing (greenscreen)
  - Video editing software
  - AI-based manipulations (inpainting)

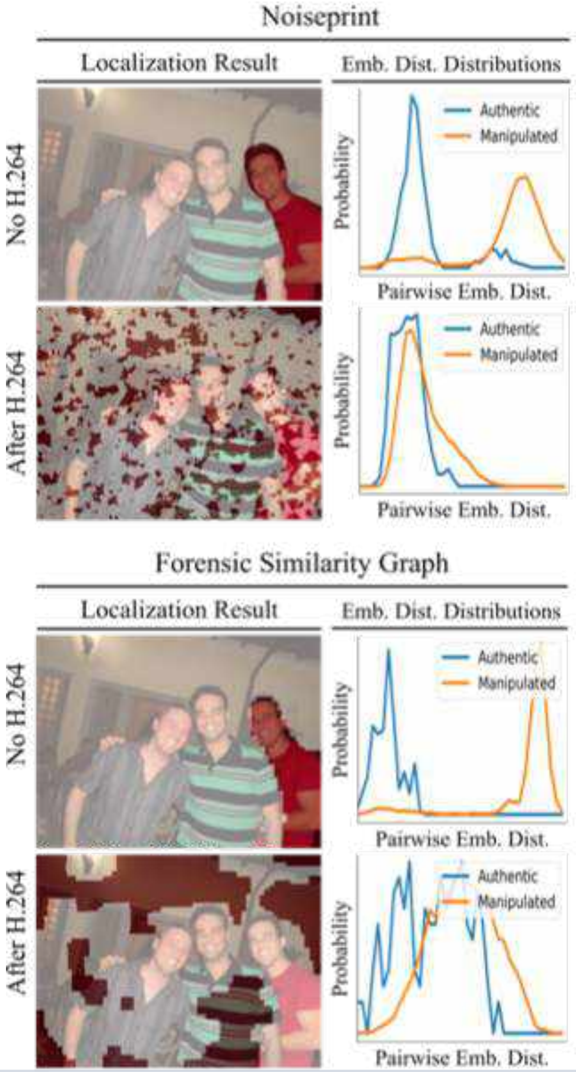
Forgery Detection & Localization

- Many manipulation detectors & localizers
- Strong performance on images
- Video is just a sequence of images, right?
- *When applied to video existing detectors all fail!*

Noiseprint Forensic Similarity Graph

T. D. Nguyen, S. Fang, and M. C. Stamm, “VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces,” to appear at WACV 2024, https://arxiv.org/abs/2211.15775

Effect of H.264 Encoding



- Why does this happen?
- Detectors/localizers search for inconsistent forensic traces
- H.264 encodes each macroblock differently within a frame
  - Introduces local variation into forensic traces
  - Decreases quality of traces nonuniformly
- Introduces unintended forensic inconsistencies

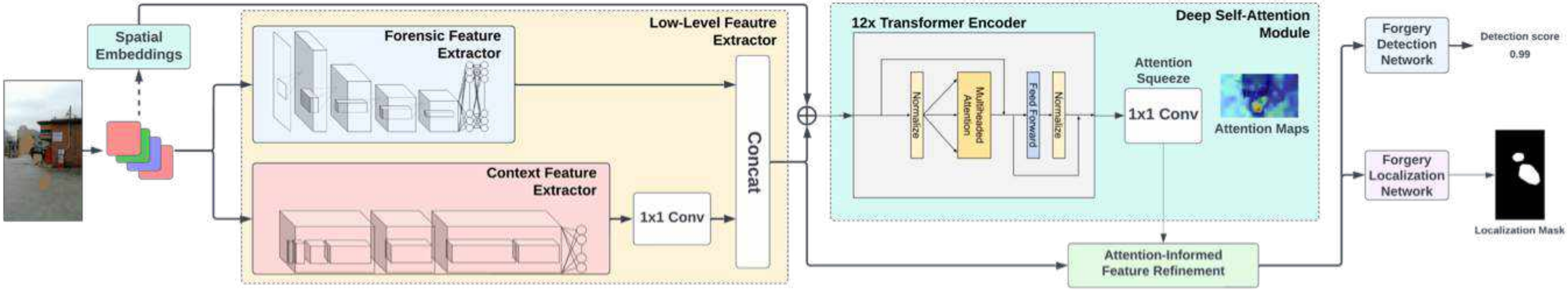
T. D. Nguyen, S. Fang, and M. C. Stamm, "VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces," to appear at WACV 2024, https://arxiv.org/abs/2211.15775

Video Forensics

How can we overcome this problem?

- Use *context* and *self-attention* to account for variation in forensic traces

VideoFACT: **Video F**orensics using **A**ttention, **C**ontext, and **T**races



T. D. Nguyen, S. Fang, and M. C. Stamm, “VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces,” to appear at WACV 2024, https://arxiv.org/abs/2211.15775

Overcoming Video Challenges

Context: Exploit conditional information

- Distribution of forensic traces changes based on several factors
  - Coding parameters/strength, scene texture, illumination, etc.
- Use *context embeddings* to capture this information
- Network learns distribution of forensic traces conditioned on context

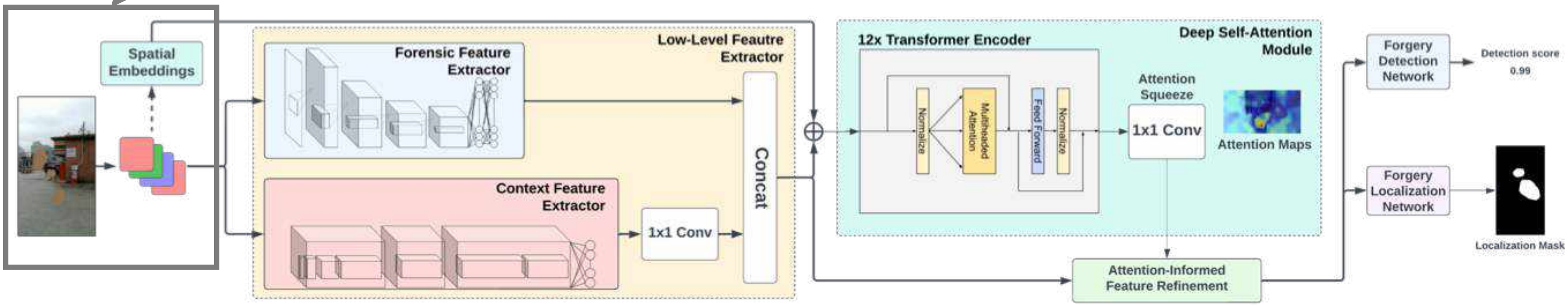
Self-Attention: Estimate quality & relative importance of info

- De-emphasize embeddings from regions with low-quality traces
- Emphasize embeddings from regions important for forensic decision making

High-Level Overview

- Video frame divided into 128 x 128 pixel analysis blocks
- Spatial position remembered for later use by transformer

Frame Pre-Processing



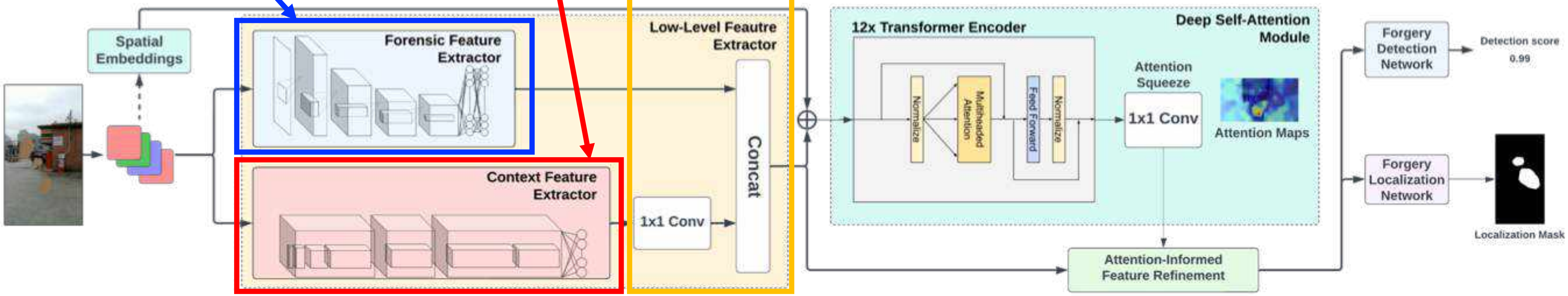
High-Level Overview

- Forensic & context embeddings extracted from each analysis block
- Concatenated to produce joint feature set

Forensic Embedding Extraction

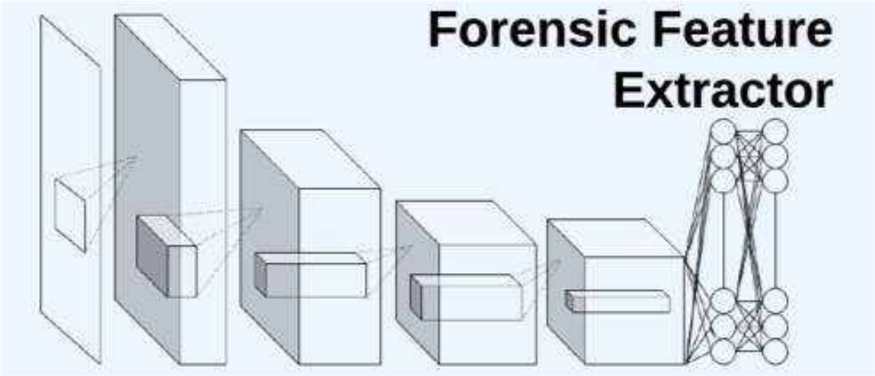
Context Embedding Extraction

Concatenate for Joint Feature Embeddings



Forensic Embeddings

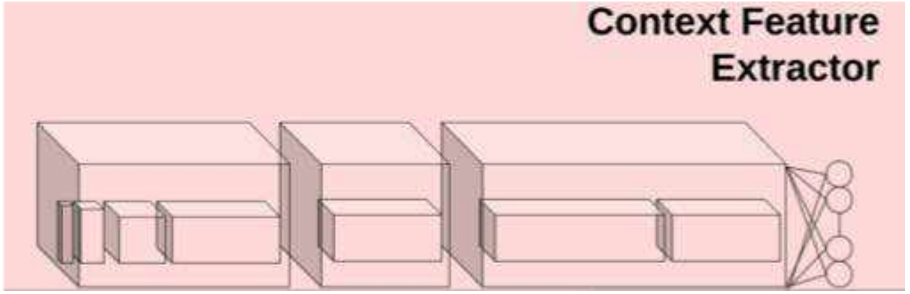
- Use MISLnet to extract forensic embeddings
- Pretrained to perform camera model identification
  - Prior work shows this learns transferrable generic forensic embeddings [1]
  - Ablation study shows this is important
- Weights frozen while context embeddings are initially learned



[1] O. Mayer, B. Bayar, and M. C. Stamm. "Learning unified deep-features for multiple forensic tasks." In ACM IH&MMSec, pp. 79-84. 2018.

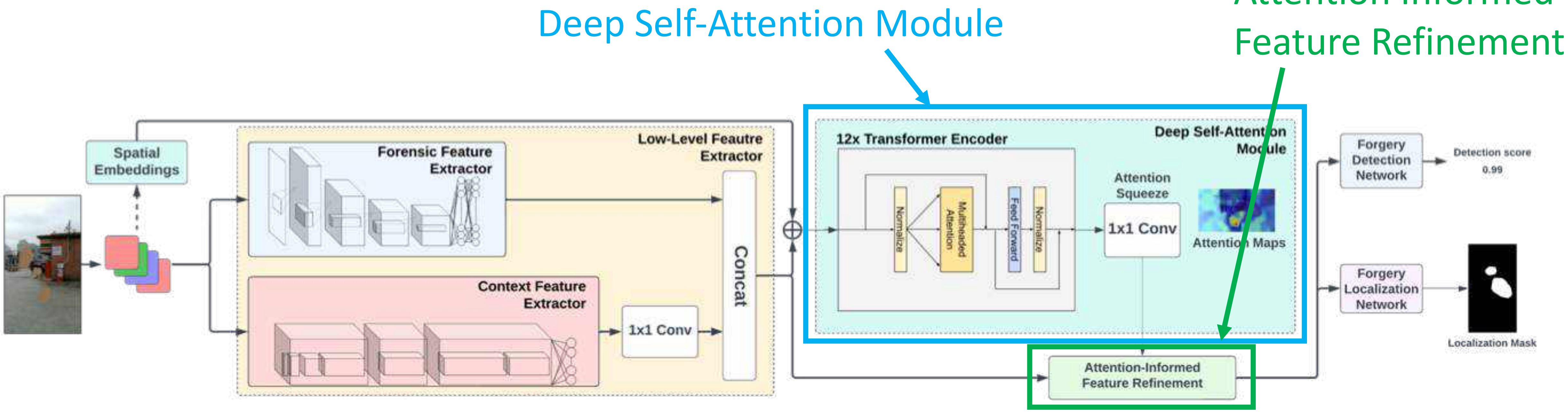
Context Embeddings

- Use separate CNN to learn context embeddings
  - Xception modified to use only a single middle flow module
  - Avoid overfitting to abstract scene representations
- Followed by a 1×1 layer to reduce dimensionality
- Trained while context feature extractor is frozen



High-Level Overview

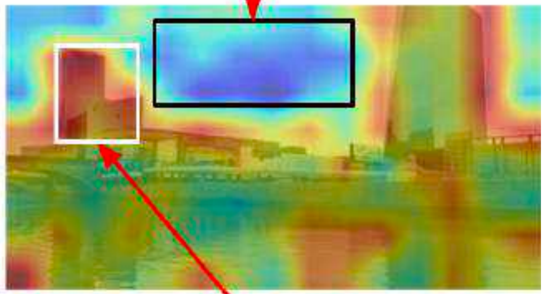
- Deep self-attention mechanism uses transformer to examine sequence of embeddings
  - Spatial position embeddings also used
  - Produces set of spatial attention maps
- Joint features weighted & combined using spatial attention maps



T. D. Nguyen, S. Fang, and M. C. Stamm, "VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces," to appear at WACV 2024, https://arxiv.org/abs/2211.15775

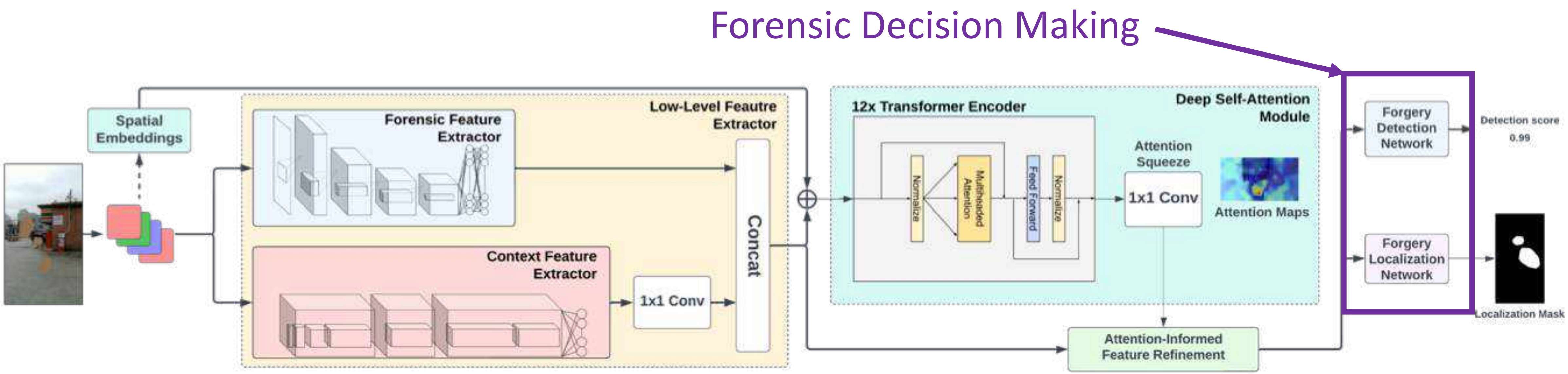
Deep Self-Attention Module

- Transformer built with 12 encoder blocks
- Jointly analyze sequence of
  - Forensic embeddings
  - Context embeddings
  - Spatial position embeddings
- Outputs spatial attention maps
  - Small weight to regions with low-quality info
  - Large weight to regions with high-quality & relevant info



High-Level Overview

- Final forensic decisions made using attention-refined features
- Separate networks for detection and localization
  - Disregard localization if no detection



T. D. Nguyen, S. Fang, and M. C. Stamm, "VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces," to appear at WACV 2024, https://arxiv.org/abs/2211.15775

Datasets

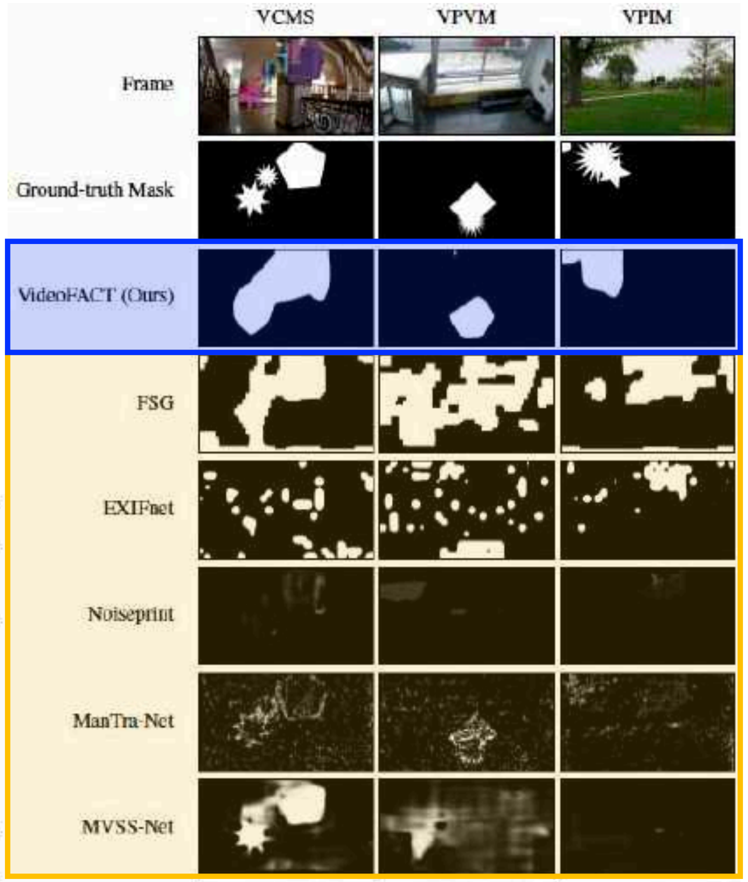
- Almost no public video forgery datasets
  - Only Adobe VideoSham (WACV 2023) for evaluation
  - None for training
- “Standard Manipulation” datastets created by us
- “In-the-Wild” datasets
  - AI-Based Inpainting
    - Created by us using E2FGVI & FuseFormer algorithms
  - Deepfakes
    - DeepFaceLab deepfakes created by us
    - FaceForensics++, Deepfake Detection Dataset (DFD)

T. D. Nguyen, S. Fang, and M. C. Stamm, “VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces,” to appear at WACV 2024, https://arxiv.org/abs/2211.15775

Results – Splicing & Editing

- Very strong detection & localization performance
  - VCMS – Splicing
  - VPVM – Editing
  - VPIM – Editing (Invisible)
- Existing detectors largely fail

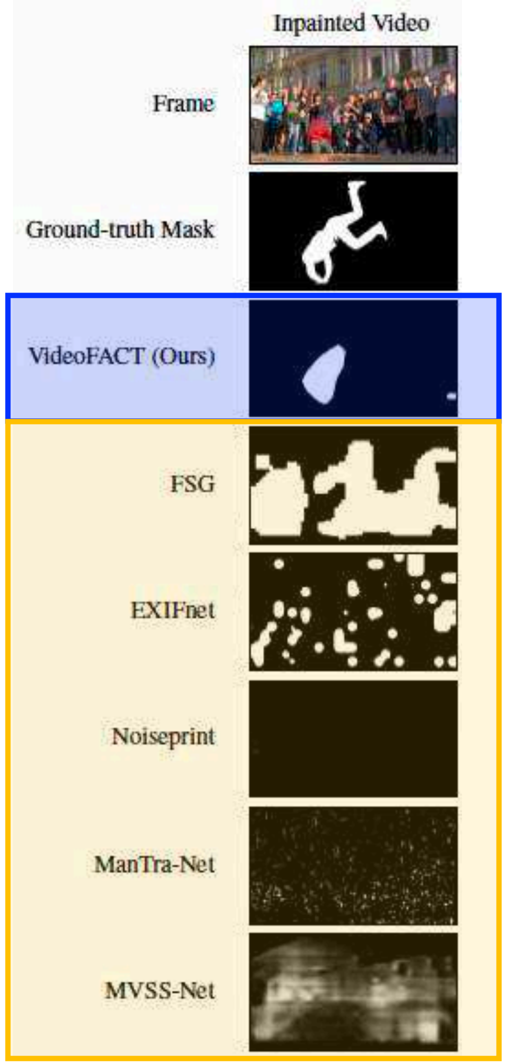
Method	VCMS				VPVM				VPIM			
	Det. mAP	*Det. ACC*	*Loc. MCC*	*Loc. F1*	*Det. mAP*	*Det. ACC*	*Loc. MCC*	*Loc. F1*	*Det. mAP*	*Det. ACC*	*Loc. MCC*	*Loc. F1*
FSG [40]	0.445	0.497	0.001	0.064	0.431	0.480	0.004	0.067	0.485	0.494	0.011	0.065
EXIFnet [26]	0.610	0.502	0.208	0.230	0.568	0.501	0.213	0.236	0.509	0.500	0.026	0.124
Noiseprint [12]	0.521	0.500	0.041	0.030	0.495	0.500	0.012	0.013	0.511	0.500	0.010	0.010
ManTra-Net [58]	0.451	0.500	0.079	0.114	0.526	0.500	0.110	0.145	0.513	0.500	0.025	0.064
MVSS-Net [8]	0.883	0.602	0.545	0.557	0.644	0.529	0.267	0.279	0.482	0.492	0.018	0.042
VideoFACT	0.995	0.987	0.530	0.526	0.980	0.950	0.676	0.697	0.869	0.797	0.515	0.547



Results - Inpainting

- Baseline VideoFACT: not trained on any inpainting data
  - Good detection & localization results
- VideoFACT-FT: fine tuned using very small training dataset
  - Excellent detection & localization results
- Existing approaches largely fail

Method	E2FGVI Inpainted Videos				FuseFormer Inpainted Videos			
	Det. mAP	*Det. ACC*	*Loc. MCC*	*Loc. F1*	*Det. mAP*	*Det. ACC*	*Loc. MCC*	*Loc. F1*
FSG [40]	0.386	0.452	0.208	0.302	0.351	0.484	**0.241**	**0.290**
EXIFnet [26]	0.635	0.501	0.160	0.244	0.506	0.507	0.146	0.225
Noiseprint [12]	0.601	0.500	0.091	0.232	0.471	0.500	0.001	0.200
ManTra-Net [58]	0.499	0.500	0.009	0.055	0.613	0.500	0.031	0.204
MVSS-Net [8]	0.341	0.435	0.058	0.227	0.230	0.359	0.029	0.206
VideoFACT	**0.782**	**0.687**	**0.225**	**0.309**	**0.652**	**0.527**	0.118	0.237
VideoFACT-FT	**0.908**	**0.820**	**0.411**	**0.445**	**0.948**	**0.846**	**0.361**	**0.411**



T. D. Nguyen, S. Fang, and M. C. Stamm, "VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces," to appear at WACV 2024, https://arxiv.org/abs/2211.15775

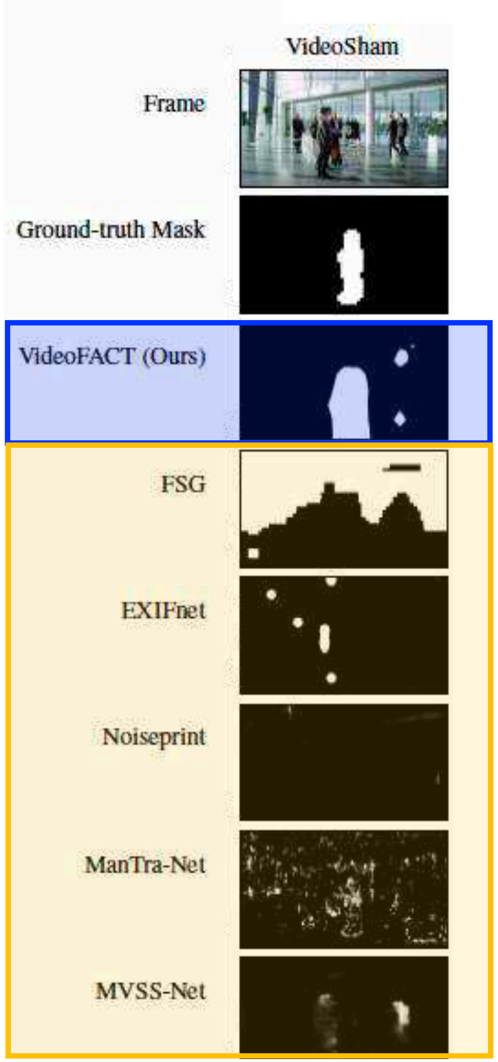
Results – Adobe VideoSham

- VideoSHAM contains multiple video manipulations
  - Color change, object add/remove, text add/remove, etc.

- VideoFACT not trained or finetuned on any of this data
  - Strongest reported results

- Existing approaches largely fail

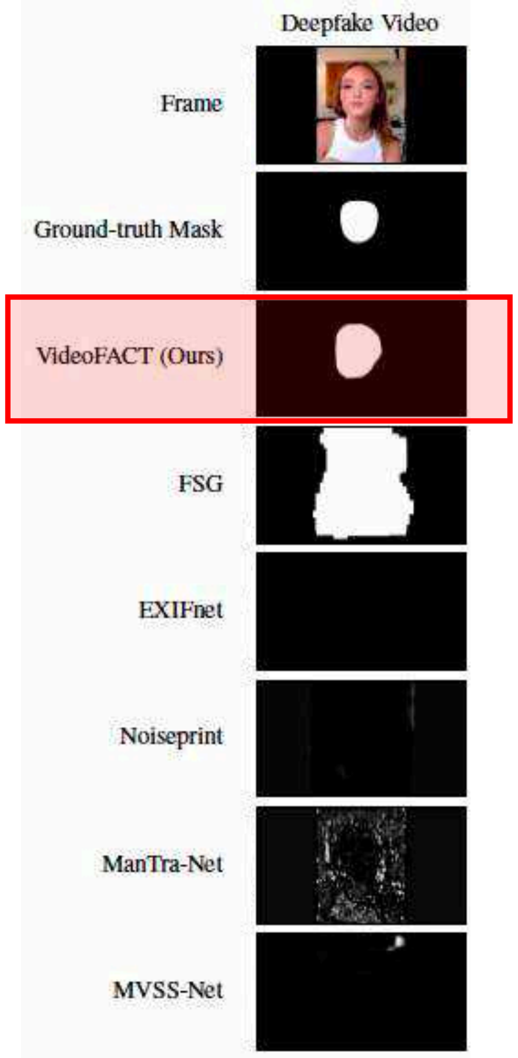
Method	VideoSham [42]			
	Det. mAP	*Det. ACC*	*Loc. MCC*	*Loc. F1*
FSG [40]	0.596	0.538	0.162	0.246
EXIFnet [26]	0.584	0.555	0.148	0.246
Noiseprint [12]	0.422	0.447	0.034	0.206
ManTra-Net [58]	0.551	0.553	0.009	0.058
MVSS-Net [8]	0.595	0.449	0.142	0.096
VideoFACT	**0.691**	**0.656**	**0.193**	**0.312**



Results - Deepfakes

- Baseline VideoFACT performance is mixed
- VideoFACT-FT: fine tuned 10% of DFD & FF++ training datasets
  - Excellent detection & localization results
- VideoFACT-FT outperforms existing approaches
  - Splicing detectors largely fail
  - Outperforms existing deepfake detectors on this experiment

Method	DeepFaceLab Deepfake Videos				DFD [14]				FF++ [49]			
	Det. mAP	*Det. ACC*	*Loc. MCC*	*Loc. F1*	*Det. mAP*	*Det. ACC*	*Loc. MCC*	*Loc. F1*	*Det. mAP*	*Det. ACC*	*Loc. MCC*	*Loc. F1*
FSG [40]	0.450	0.515	0.204	0.137	0.449	0.325	0.097	0.043	0.509	**0.519**	0.144	0.113
EXIFnet [26]	0.447	0.492	0.180	0.133	0.489	0.258	0.095	0.051	0.487	**0.519**	0.141	0.073
Noiseprint [12]	0.591	0.500	0.010	0.062	0.489	0.252	0.000	0.021	0.486	0.518	0.000	0.066
ManTra-Net [58]	0.450	0.500	0.004	0.042	0.476	0.253	0.017	0.025	0.504	0.514	0.070	0.091
MVSS-Net [8]	0.464	0.498	0.199	0.189	**0.513**	**0.532**	**0.152**	**0.108**	0.499	0.487	0.133	0.164
VideoFACT	**0.666**	**0.648**	**0.415**	**0.410**	0.468	0.444	0.081	0.077	**0.529**	**0.519**	**0.160**	**0.167**
VideoFACT-FT	**0.988**	**0.922**	**0.745**	**0.732**	**0.937**	**0.804**	**0.536**	**0.490**	**0.916**	**0.837**	**0.661**	**0.645**
E.ViT [10]	0.896	0.805	N/A	N/A	0.811	0.737	N/A	N/A	0.764	0.676	N/A	N/A
CCE.ViT [10]	0.962	0.837	N/A	N/A	0.816	0.761	N/A	N/A	0.796	0.719	N/A	N/A
CNN Ensemble [6]	0.936	0.857	N/A	N/A	0.829	0.745	N/A	N/A	0.713	0.672	N/A	N/A



Summary

- H.264 significantly harms forgery detectors & localizers

- Can overcomes this using multiple strategies
  - Context embeddings
  - Self-attention

- Strong experimental performance – still much to do!

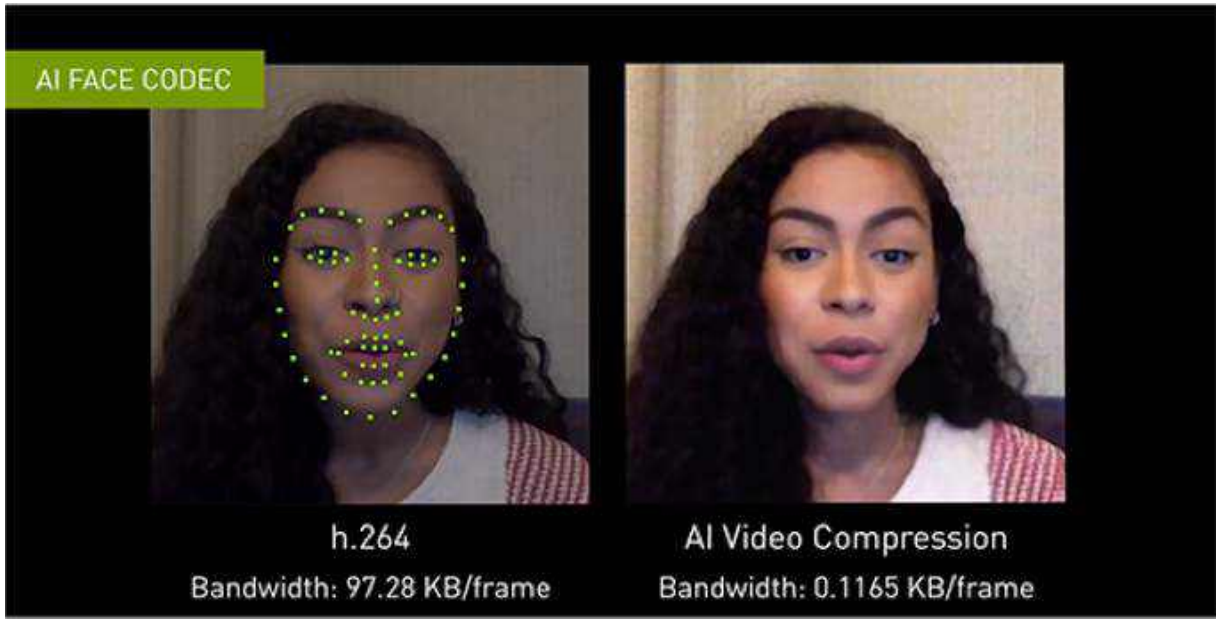
- Paper available at: https://arxiv.org/abs/2211.15775

Talking Head Videoconferencing

- Videoconferencing consumes significant bandwidth

- Recent research uses AI to compress talking head videos
  - Capture facial expression of sender
  - Use to synthesize face at receiver

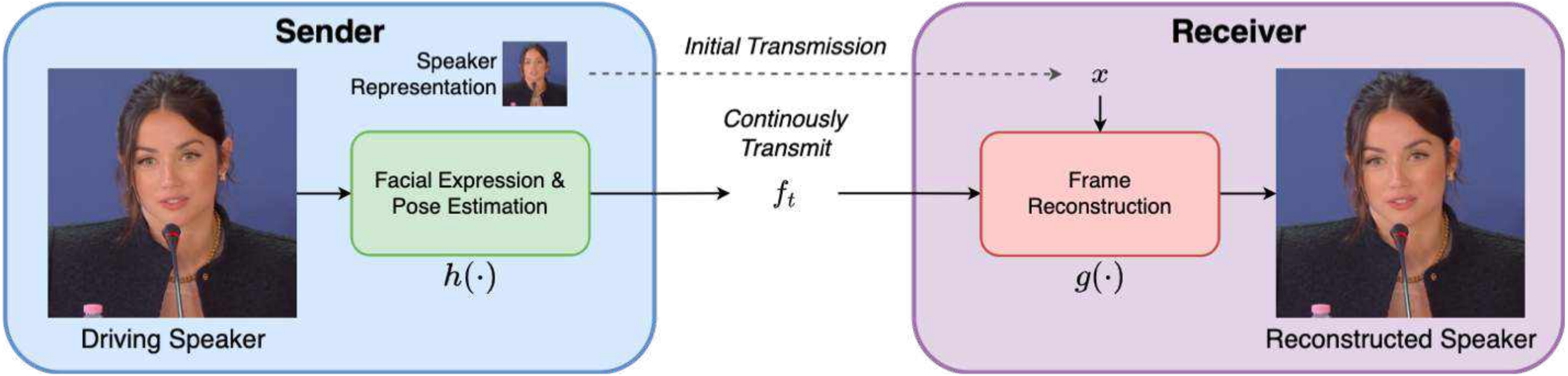
- Several recent approaches
  - NVIDIA Maxine
  - X2Face
  - DA-GAN
  - SAFA
  - Many more!



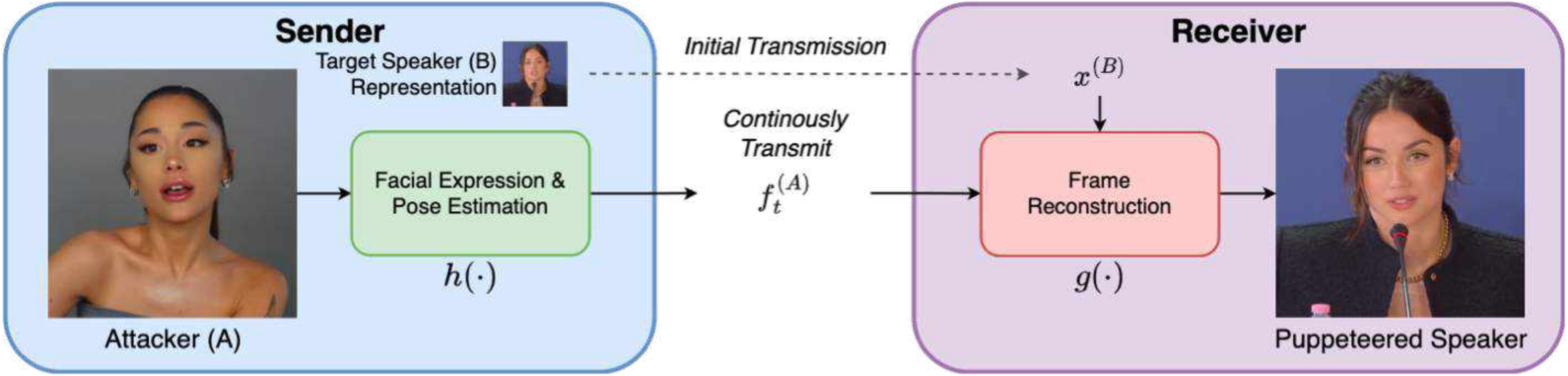
Puppeteering Attacks

- Problem: *Puppeteering attacks*
- "Driving" speaker controls target face like a puppet in real time
- Deepfake detectors can't protect against this
  - *Everything is a deepfake!*

AI Videoconferencing: Closer Look

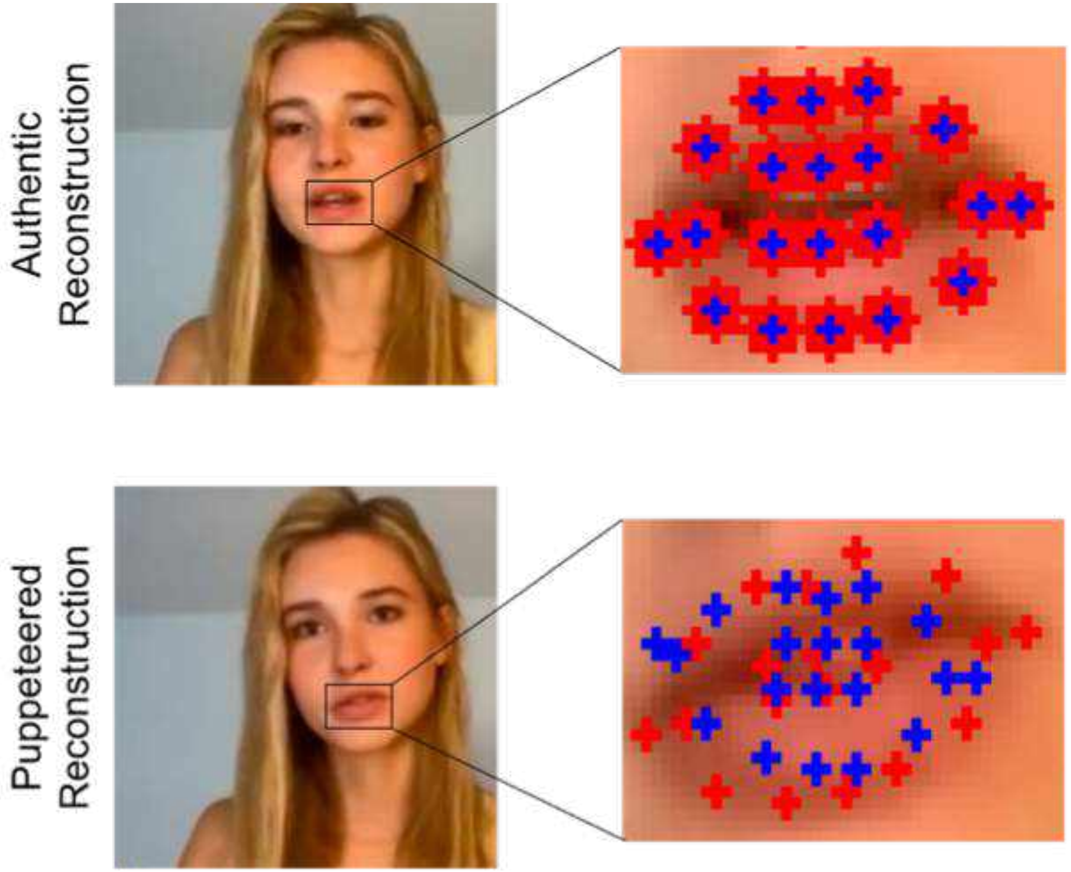


Pupetteering Attack



Key observaction

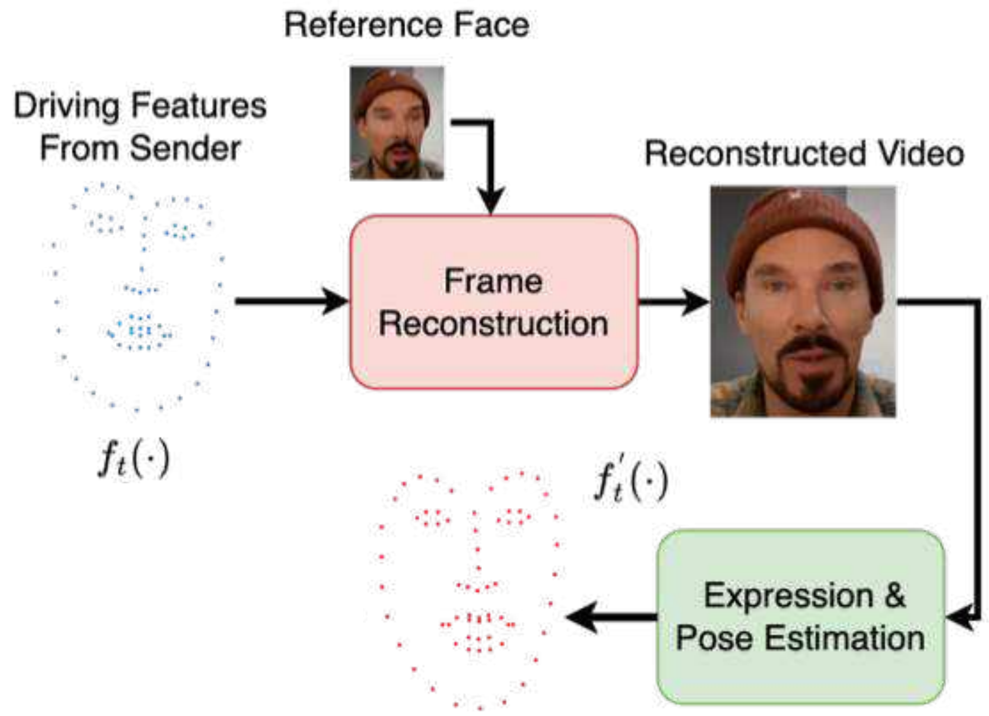
- Can compare facial landmarks from sender and those in reconstructed speaker
- Self-driven video
  - Landmarks are tightly coupled
- Puppeteered video
  - Difference in landmark positions
  - Caused by differences in facial geometry



Puppeteering Detection

- Reconstruct speaker at receiver
- Pass reconstructed face through encoder
- Obtain facial expression and pose estimation vector (landmarks)

$$f_t' = h(I_t')$$



Puppeteering Detection

- Measure biometric difference between landmark vectors

$$d_t = m(f_t, f_t') = \|f_t - f_t'\|_2^2$$

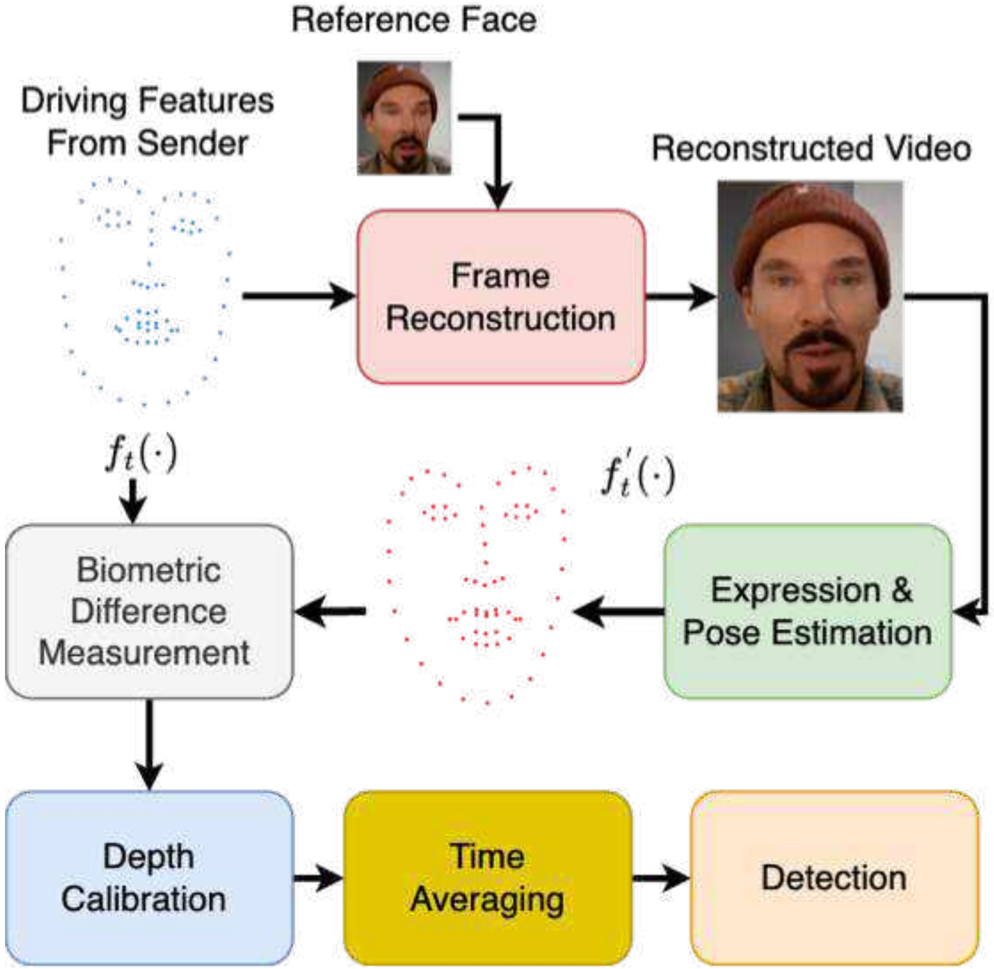
- Control for depth

$$c_t = d_t \left(\frac{r_t}{r_0}\right)$$

- Average over time

$$\Delta_t = \frac{1}{W} \sum_{\ell=0}^{W-1} c_{t-\ell}$$

- Threshold for detection



Puppeteering Dataset

- Created dataset of 1728 puppeteered videos
  - Public videos of celebrities
- Created using four different systems
  - DaGAN
  - SAFA
  - X2Face
  - ReenactGAN

Dataset available at: https://gitlab.com/MISLgit/talking-head-puppeteering-defense/

Experimental Results

	Proposed	CNN Ensemble	Efficient ViT	Cross-Efficient ViT
DaGAN	99.31%	66.80%	76.26%	69.81%
Reenact GAN	94.83%	69.73%	76.96%	68.58%
X2Face	99.80%	68.24%	79.00%	78.15%
SAFA	98.92%	67.35%	74.86%	67.81%
Average	98.03%	68.03%	76.77%	71.09%

- Strong detection performance across all talking head video systems
- Significantly outperforms deepfake detectors (as expected)
  - Higher performing deepfake detectors misclassify self-reenacted videos as real!

Experimental Results

	White male	White Female	Asian Male	Asian Female	Black Male	Black Female	Hispanic Male	Hispanic Female
DaGAN	98.37%	99.60%	97.04%	98.13%	98.26%	99.15%	99.71%	99.42%
Reenact GAN	94.10%	93.58%	95.27%	95.74%	93.54%	96.08%	94.37%	96.84%
X2Face	99.37%	98.26%	99.46%	97.35%	98.14%	99.31%	98.46%	99.02%
SAFA	99.74%	99.91%	97.10%	98.75%	98.48%	99.23%	98.20%	98.61%
Average	97.99%	97.84%	97.22%	97.49%	97.19%	98.44%	97.44%	98.47%

- Examined system for algorithmic bias

- Consistent performance across race/ethnicity and sex

Summary

- Many video forgery types beyond just deepfakes
- Detect multiple forgery types by using forensic traces, context, and self-attention
- Detect puppeteering by exploiting mismatch in implicit biometric information
- Much more work to be done!

Video Forensics Beyond Deepfakes

Matthew C. Stamm
Multimedia & Information Security Lab
Drexel University
mstamm@drexel.edu