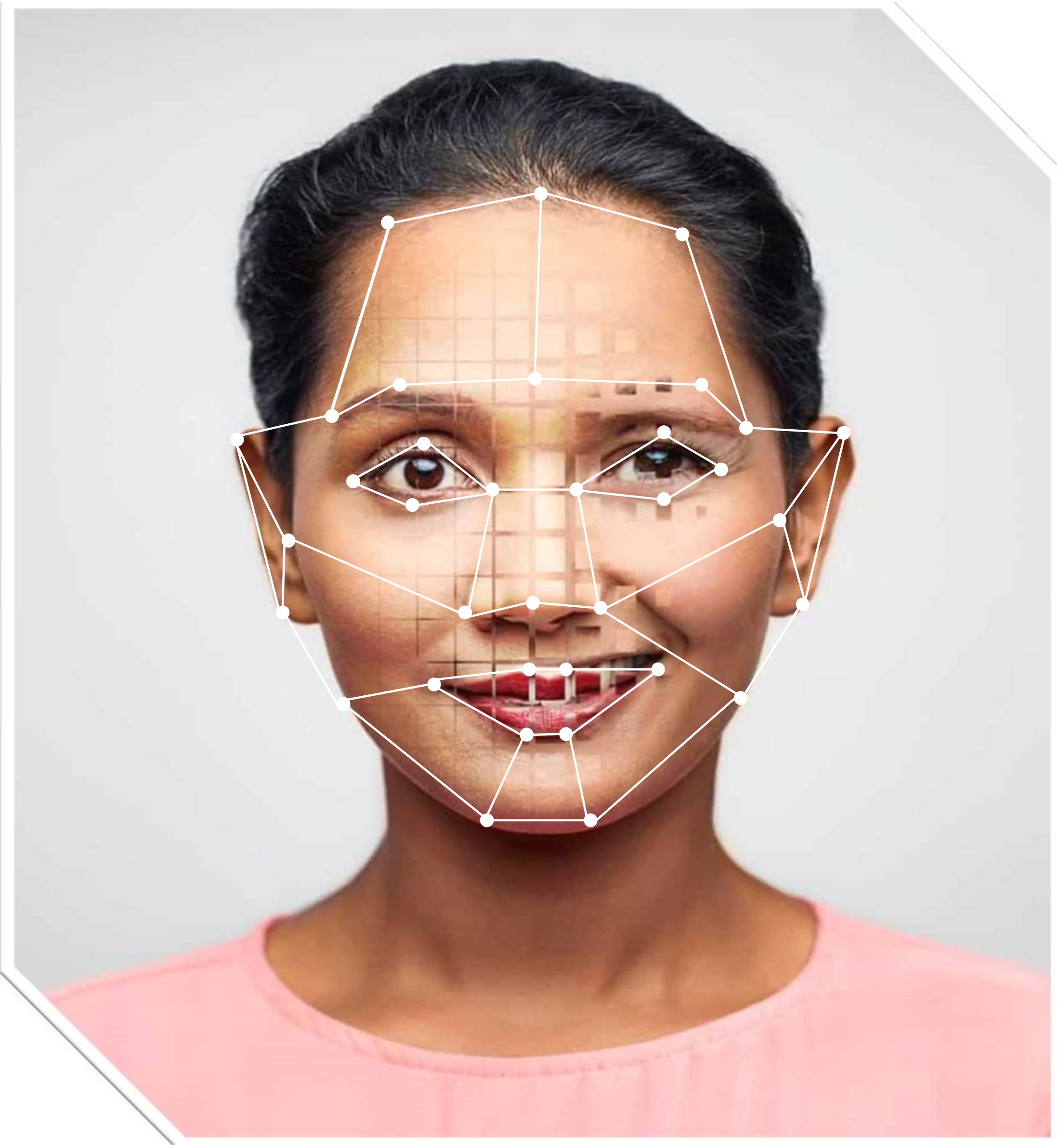


Detecting AI-Altered Media with Deep Learning

Advisor
Dr. Arnab Bose

Presented By
Rhys Chua
Jim Fang
Jon Huff

THE UNIVERSITY OF CHICAGO

August 29th, 2020:

A digitally altered video of Joe Biden is uploaded onto Twitter & Youtube

Within the span of 24 hours...

The video received 2.4M views on Twitter...

Within the span of 24 hours...

The video received 2.4M views on Twitter...

200K+ views on YouTube...

Within the span of 24 hours...

The video received 2.4M views on Twitter...

200K+ views on YouTube...

Shared tens of thousands of times on Facebook...

Within the span of 24 hours...

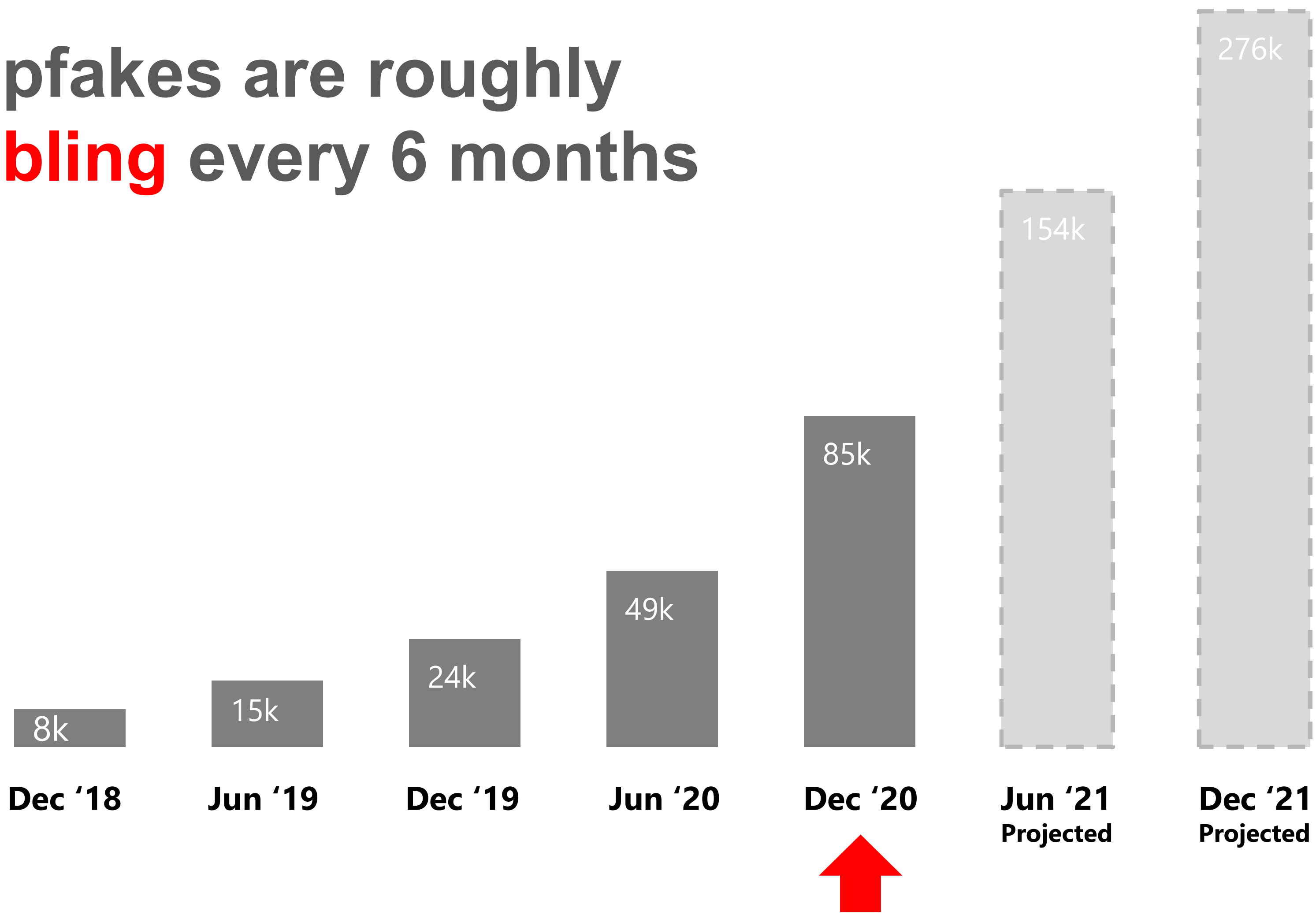
The video received 2.4M views on Twitter...

200K+ views on YouTube...

Shared tens of thousands of times on Facebook...

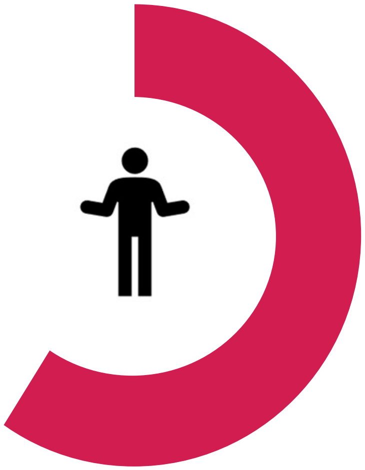
...And gets retweeted by the White House Deputy Chief of Staff & Director of Social Media, Dan Scanvino, which was then seen by millions more

Deepfakes are roughly **doubling** every 6 months




Every minute,
16 IDENTITIES
will be stolen
and repurposed.

The Public senses a looming threat...



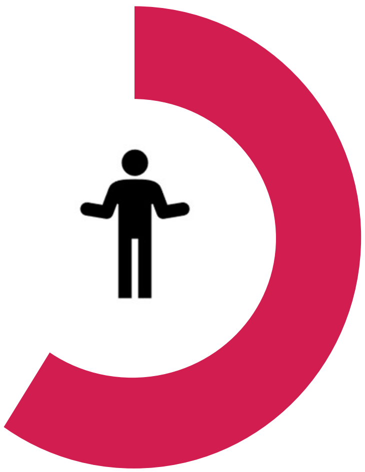
63%

of U.S. adults surveyed believe altered videos create **a great deal of confusion** about the facts of current events

77%

of U.S. adults surveyed **support restrictions** on publishing and accessing them

...and the U.S. Government Agrees



63%

of U.S. adults surveyed believe altered videos create **a great deal of confusion** about the facts of current events

77%

of U.S. adults surveyed **support restrictions** on publishing and accessing them

$68M

Spent by **DARPA** to research ways to fight threat of deepfakes in **2016-2018**

How do we combat the threat of misinformation?

The Solution

Build a deep learning pipeline to distinguish Fake, AI-altered videos from real videos using Deep Ensemble Learning

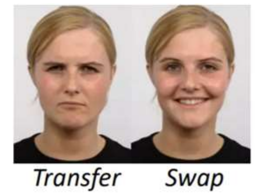
Visual

Audio

Visual

Face Reenactment

Face Replacement



Face Editing

Face Synthesis

Audio

Speech Synthesis
(Concatenative, Parametric, and Deep Learning)

Visual

- **Face Reenactment**

Mouth Expression Pose

- **Face Replacement**

Transfer Swap

- **Face Editing**

Article Age Beauty

- **Face Synthesis**

Audio

Speech Synthesis
(Concatenative, Parametric, and Deep Learning)

Goals of this Project

Build a deep learning pipeline to distinguish fake from real videos

Contribute something new to the research community **AND** **deliver a best-in-class deepfake detector** that stands up to the current SoTA on the latest datasets

Publish a working app + code that can detect deepfakes in real-time as a contribution to the research community & for commercial use

Contributions

	Tolosana Et Al. (2020)	1st Place DFDC Winner	Dessa (2019)	Wang Et Al. (2020)	Our Method
Utilizes CNN's and image preprocessing techniques (augmentation, cropping)	✓	✓	✓	✓	✓
Trained on datasets that use Face2Face, FaceSwap, Deepfake, and NeuralTextures techs.	✓	✓	✓	✓	✓
Transfer Learning using ***Xception***	✓		✓	✓	✓
Transfer Learning using ***Efficientnet***		✓			✓
Trained on ***Mixed Datasets*** for greater generalizability			✓		✓
Transfer Learning using ***3D CNN's***				✓	
Utilizes ***LSTM*** to process sequences of frames					✓

Contributions

	Tolosana Et Al. (2020)	1st Place DFDC Winner	Dessa (2019)	Wang Et Al. (2020)	Our Method
Utilizes CNN's and image preprocessing techniques (augmentation, cropping)	✓	✓	✓	✓	✓
Trained on datasets that use Face2Face, FaceSwap, Deepfake, and NeuralTextures techs.	✓	✓	✓	✓	✓
Transfer Learning using ***Xception***	✓		✓	✓	✓
Transfer Learning using ***Efficientnet***		✓			✓
Trained on ***Mixed Datasets*** for greater generalizability			✓		✓
Transfer Learning using ***3D CNN's***				✓	
Utilizes ***LSTM*** to process sequences of frames					✓
Ability to detect ***multiple subjects*** per frame					✓
Uses ***Ensemble Meta-learner*** that increases performance and allows plug/play new models					✓
Developed a ***functional app*** with built-in ***model interpretability*** algorithms					✓

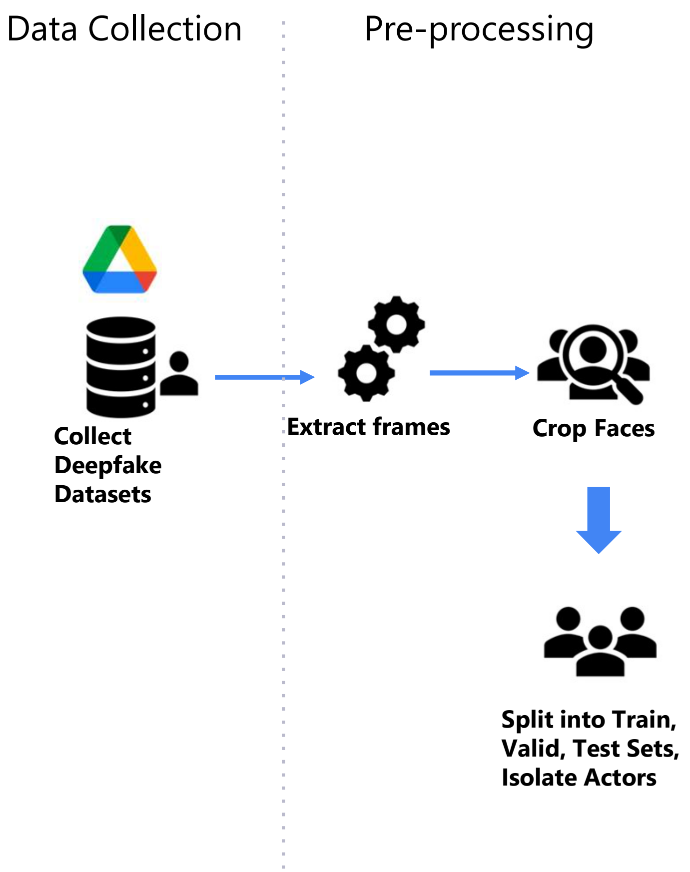
Implementation

Pipeline Overview

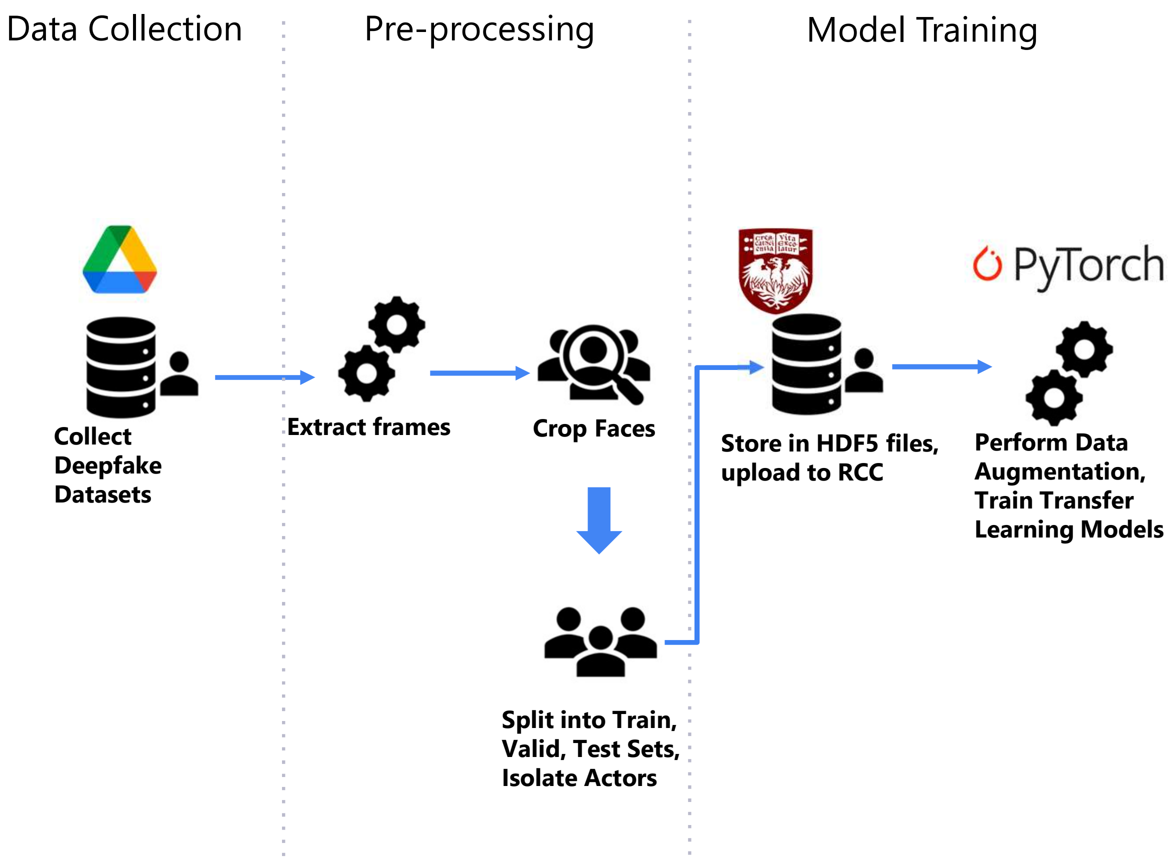
Data Collection

Collect Deepfake Datasets

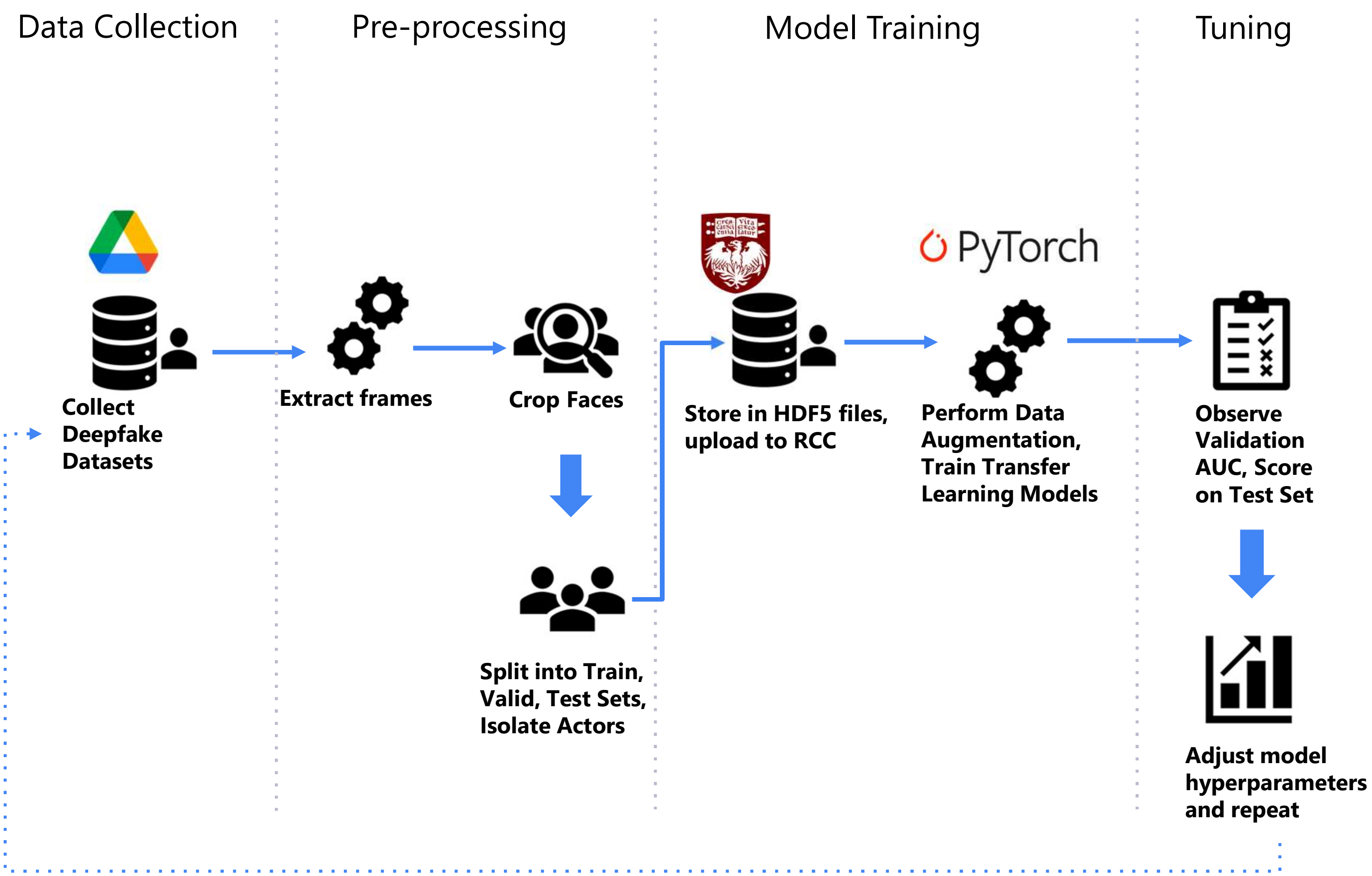
Pipeline Overview



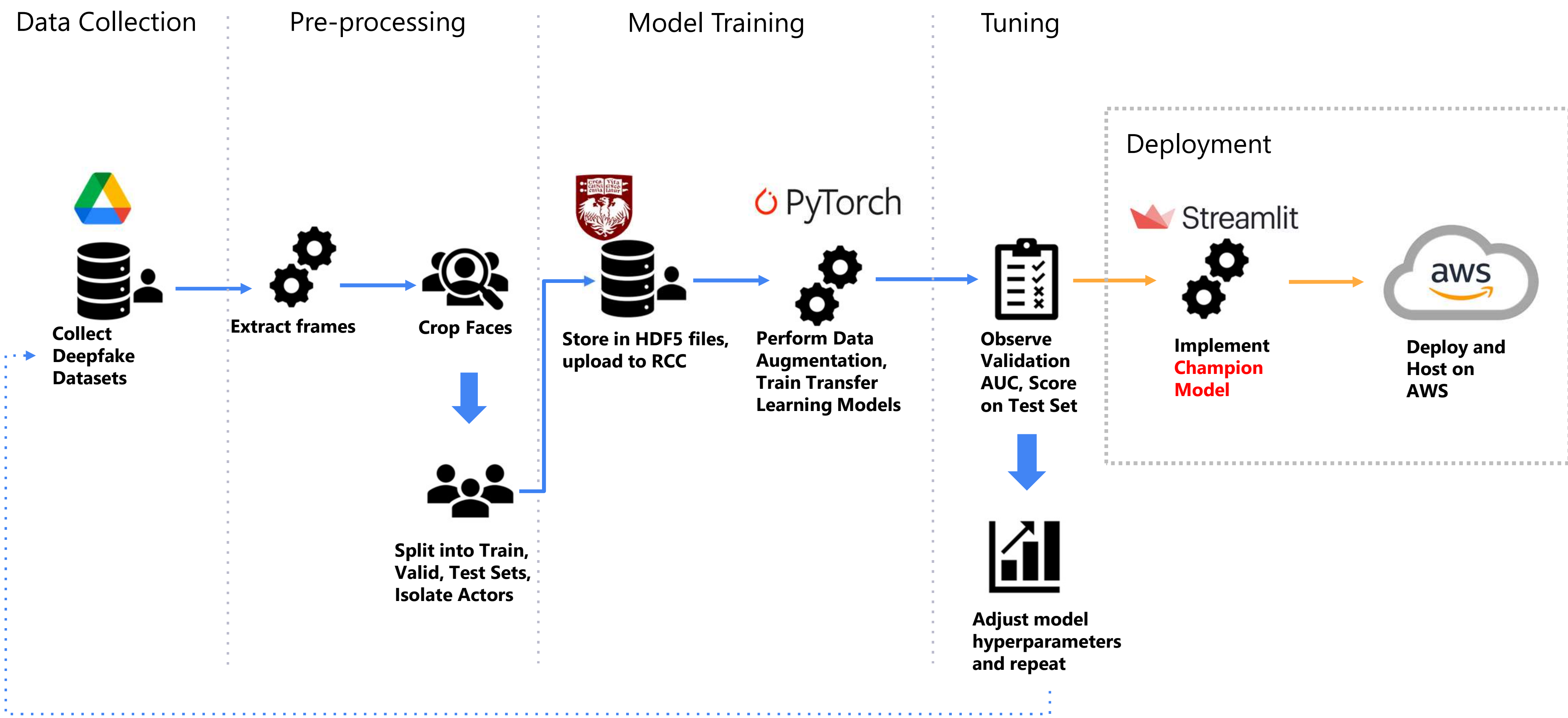
Pipeline Overview



Pipeline Overview



Pipeline Overview



Data Collection

DeepFake Datasets

	DFDC	FaceForensics++	Celeb-DF	Mixed Dataset
Main Focus	Compilation of Diff Datasets	Different forgery methods	Reduce Visual Quality Gap	FF++ and CDF
Generation	3rd	2nd	2nd	Created by Us
Size (# Videos)	25TB (129K)	39GB (5K)	34GB (6K)	280GB (8.4K)
Train/Val/Test %	58/21/21	72/14/14	68/16/16	70/15/15
Real/Fake Ratio	1:4.5	1:4	1:9	1:1
Method	Convolution Autoencoder	Generative Adversarial Network	Convolution Autoencoder	-
Technique	FaceSwap, Neural Talking Heads, Augmentation Techs	FaceSwap, Deepfakes, Face2Face, NeuralTextures	Increase resolution pixels, color transfer algorithm, face masking, temporal flickering	-

Real

Fake

Independent Variables:	Pixels from each frame of a sample video
Dependent Variables:	Fake (1) vs Real (0) – Image Label
Unit of Analysis:	300x300pixel RGB image

DeepFake Datasets

	DFDC	FaceForensics++	Celeb-DF	Mixed Dataset
Main Focus	Compilation of Diff Datasets	Different forgery methods	Reduce Visual Quality Gap	FF++ and CDF
Generation	3rd	2nd	2nd	Created by Us
Size (# Videos)	25TB (129K)	39GB (5K)	34GB (6K)	280GB (8.4K)
Train/Val/Test %	58/21/21	72/14/14	68/16/16	70/15/15
Real/Fake Ratio	1:4.5	1:4	1:9	1:1
Method	Convolution Autoencoder	Generative Adversarial Network	Convolution Autoencoder	-
Technique	FaceSwap, Neural Talking Heads, Augmentation Techs	FaceSwap, Deepfakes, Face2Face, NeuralTextures	Increase resolution pixels, color transfer algorithm, face masking, temporal flickering	-

Independent Variables:	Pixels from each frame of a sample video
Dependent Variables:	Fake (1) vs Real (0) – Image Label
Unit of Analysis:	300x300pixel RGB image

DeepFake Datasets

	DFDC	FaceForensics++	Celeb-DF	Mixed Dataset
Main Focus	Compilation of Diff Datasets	Different forgery methods	Reduce Visual Quality Gap	FF++ and CDF
Generation	3rd	2nd	2nd	Created by Us
Size (# Videos)	25TB (129K)	39GB (5K)	34GB (6K)	280GB (8.4K)
Train/Val/Test %	58/21/21	72/14/14	68/16/16	70/15/15
Real/Fake Ratio	1:4.5	1:4	1:9	1:1
Method	Convolution Autoencoder	Generative Adversarial Network	Convolution Autoencoder	-
Technique	FaceSwap, Neural Talking Heads, Augmentation Techs	FaceSwap, Deepfakes, Face2Face, NeuralTextures	Increase resolution pixels, color transfer algorithm, face masking, temporal flickering	-

Real

Fake

Independent Variables:	Pixels from each frame of a sample video
Dependent Variables:	Fake (1) vs Real (0) – Image Label
Unit of Analysis:	300x300pixel RGB image

DeepFake Datasets

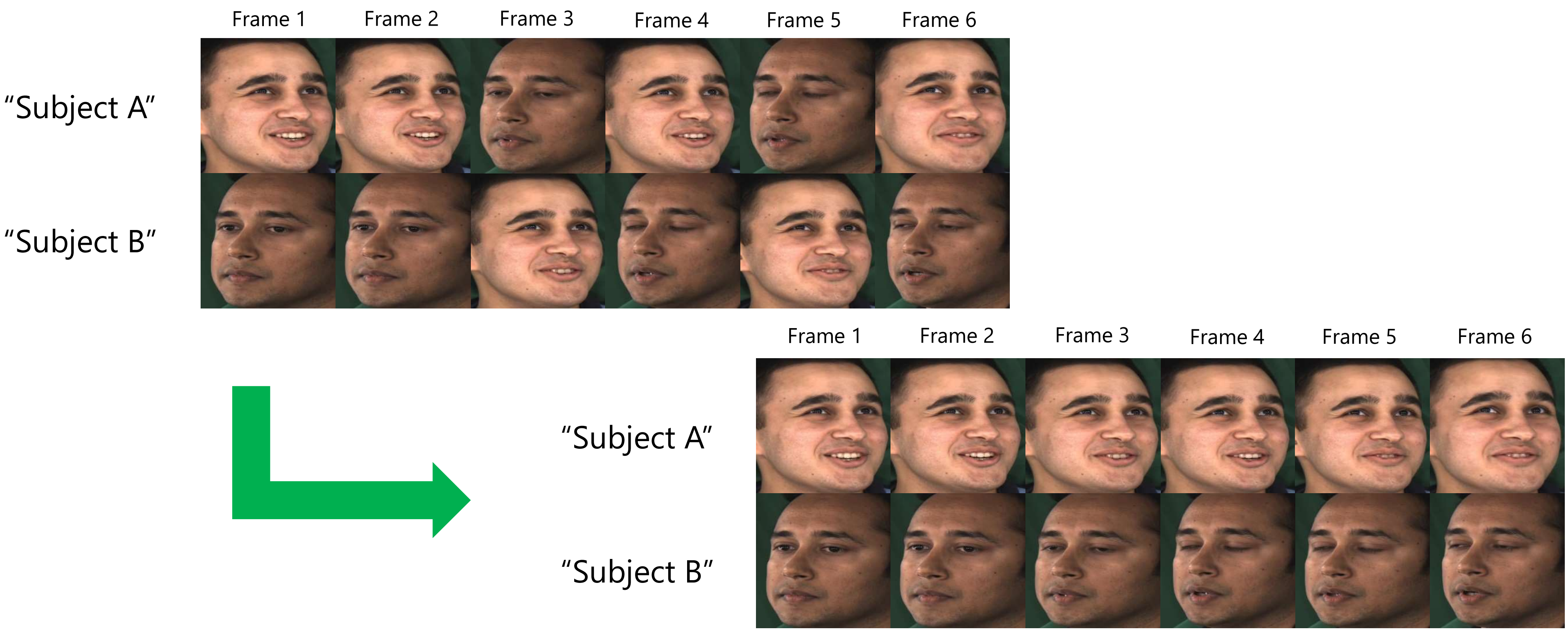
	DFDC	FaceForensics++	Celeb-DF	Mixed Dataset
Main Focus	Compilation of Diff Datasets	Different forgery methods	Reduce Visual Quality Gap	FF++ and CDF
Generation	3rd	2nd	2nd	Created by Us
Size (# Videos)	25TB (129K)	39GB (5K)	34GB (6K)	280GB (8.4K)
Train/Val/Test %	58/21/21	72/14/14	68/16/16	70/15/15
Real/Fake Ratio	1:4.5	1:4	1:9	1:1
Method	Convolution Autoencoder	Generative Adversarial Network	Convolution Autoencoder	-
Technique	FaceSwap, Neural Talking Heads, Augmentation Techs	FaceSwap, Deepfakes, Face2Face, NeuralTextures	Increase resolution pixels, color transfer algorithm, face masking, temporal flickering	-

Independent Variables: Pixels from each frame of a sample video
Dependent Variables: Fake (1) vs Real (0) – Image Label
Unit of Analysis: 300x300pixel RGB image

Video Processing

The Sorting Challenge

- We propose to analyze sequences of still frame images in videos for deepfake detection
- Naïve face detection algorithms do not automatically sort identities when multiple faces are detected in each frame. Mixed sequences of faces **contaminate** our datasets.

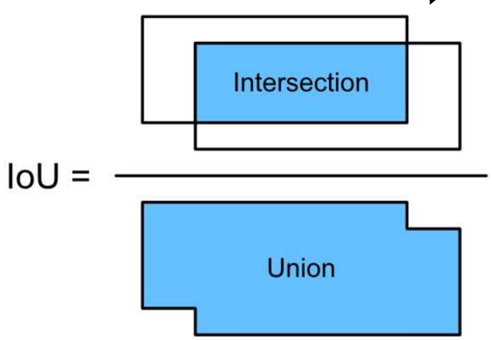


Solution: A Custom Face-Sorting Algorithm

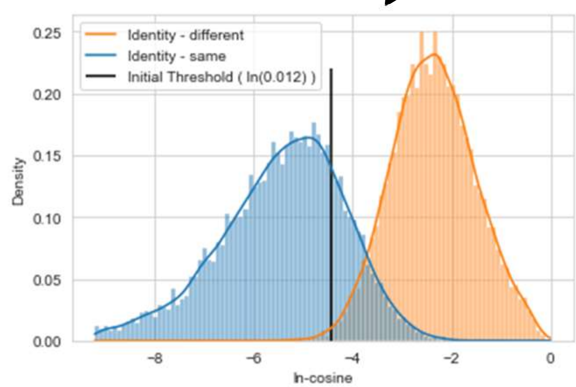
To overcome the issue of **contaminated sequences**, we designed and implemented a high-performance sorting algorithm which isolates sequences of faces based on:

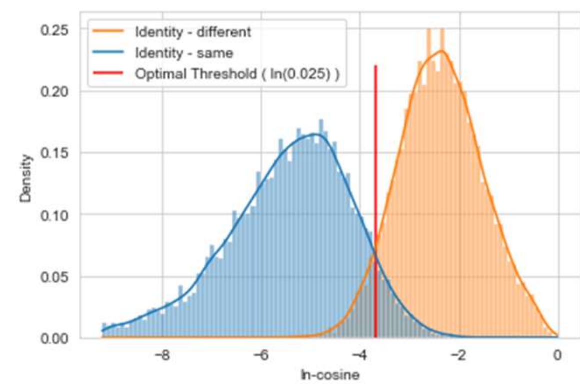
1) Sequential bounding box intersection-over-union threshold (IoU); the frames around detected faces must **overlap** frame-to-frame to be considered the **same person**

2)
- Cosine distance of facial embeddings for multiple faces derived from FaceNet
- The algorithm builds sequences by iteratively relaxing the cosine distance threshold up to a statistically determined value



+

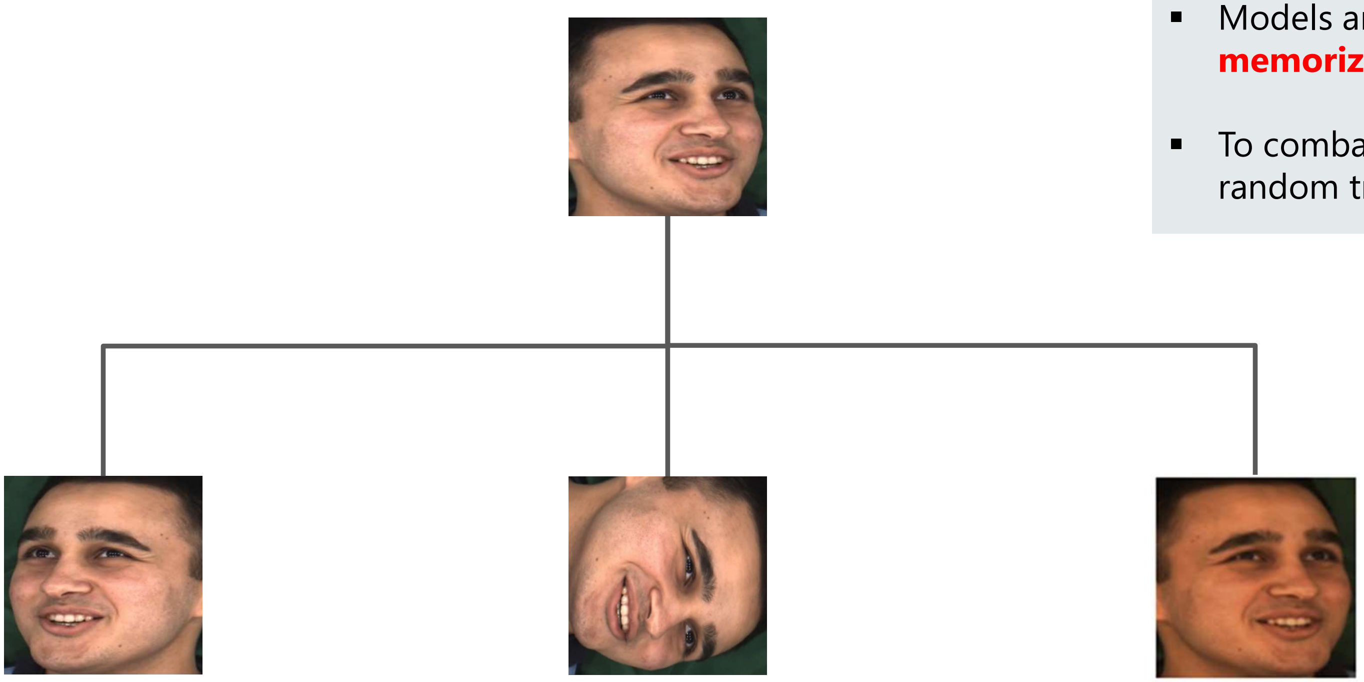




Data Transformations

- Models are prone to **overfitting** by **memorizing faces**
- To combat this, we can employ a series of random transformations

Original Image



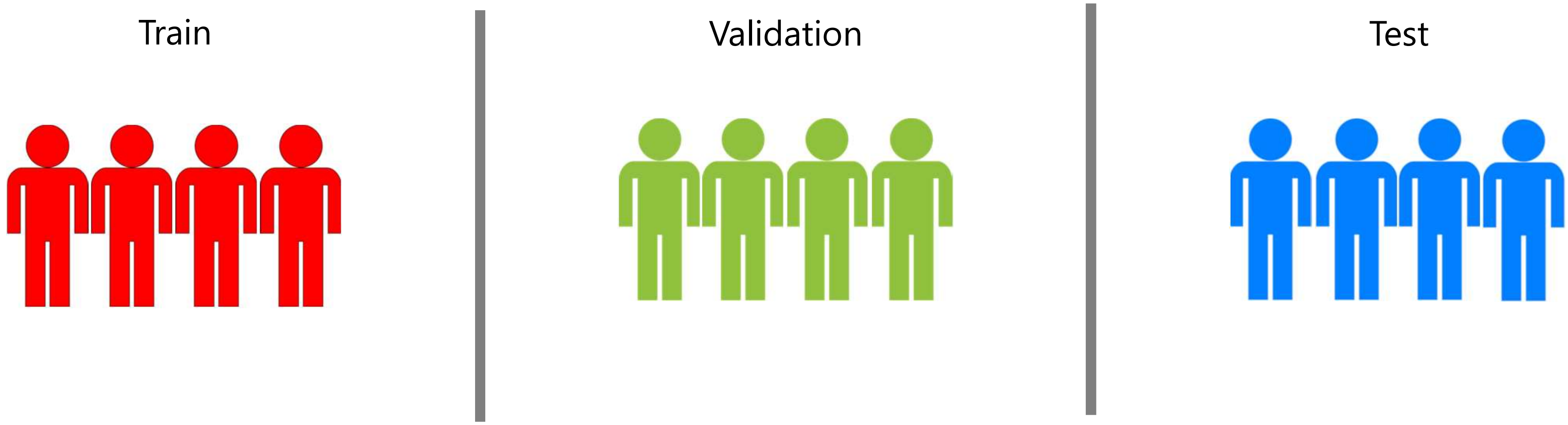
Random Horizontal Flip

Random Rotation/Scaling

Random Brightness, Contrast, Saturation Jitter, Pixelation

Dataset Splits

To prevent **overfitting** and **data leakage**, face identities were strictly isolated to each of the train, validation, and test datasets



Detection System

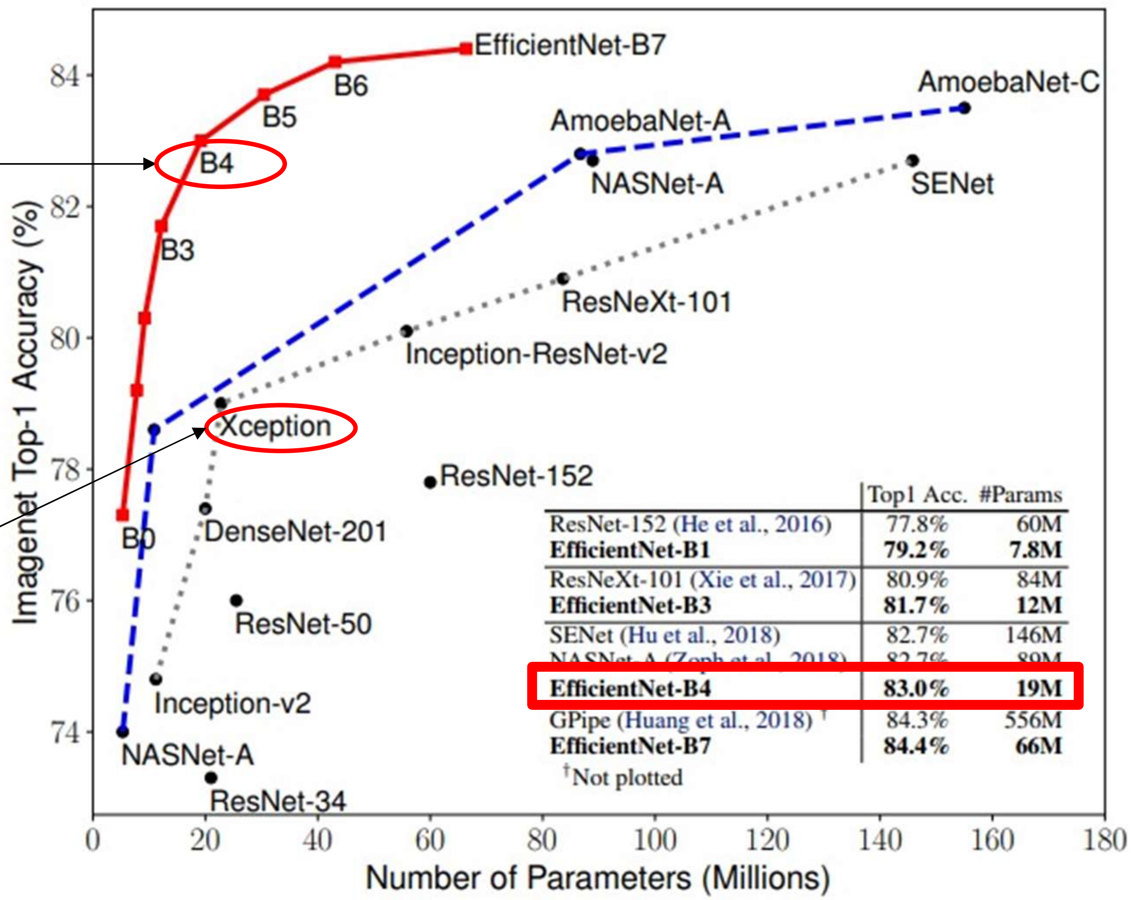
(Training + Tuning)

Model Selection

(results below based on ImageNet)

EfficientNet:
For our training process, Effnet-B4 provides the best balance of **speed** and **accuracy**, uniformly scales all dimensions of depth, width, & resolution using a compound coefficient

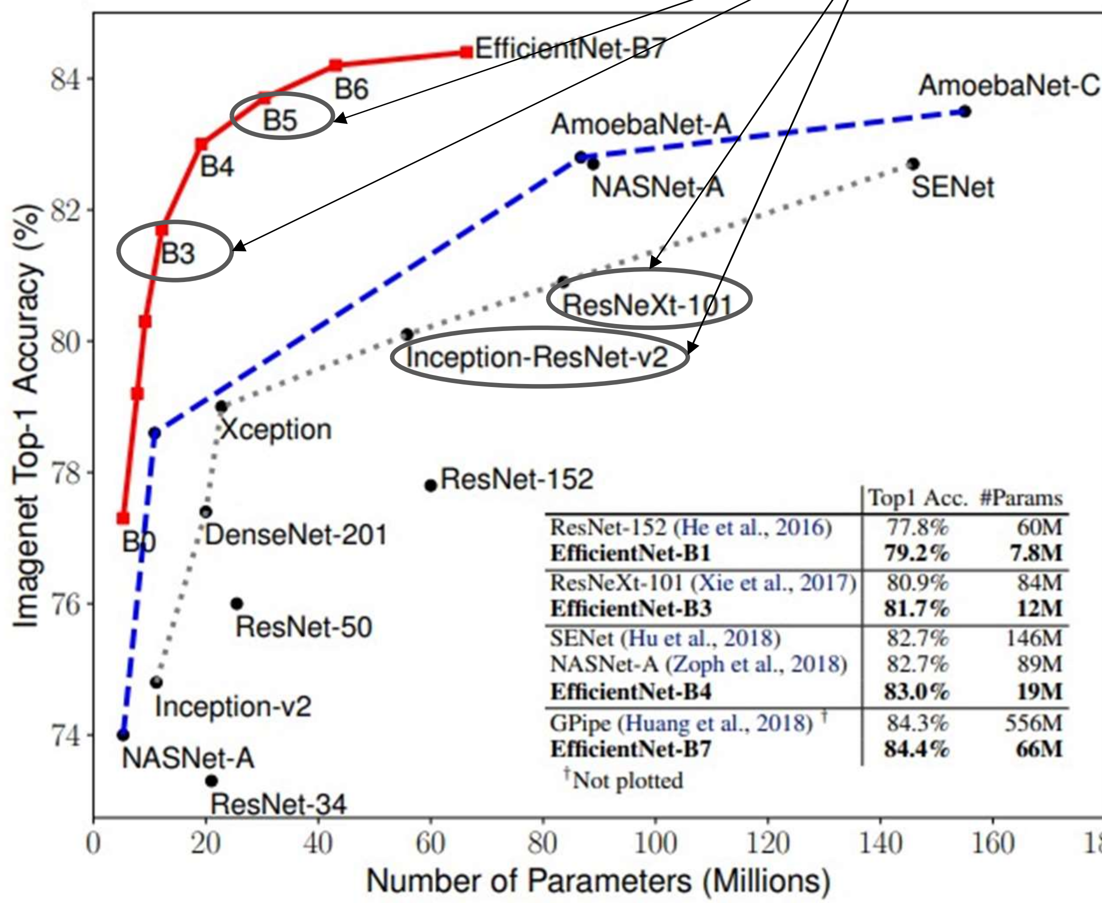
XceptionNet:
Favored by researchers, introduced CNN based entirely on depthwise separable convolution layers



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.2%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.7%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**83.0%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

† Not plotted

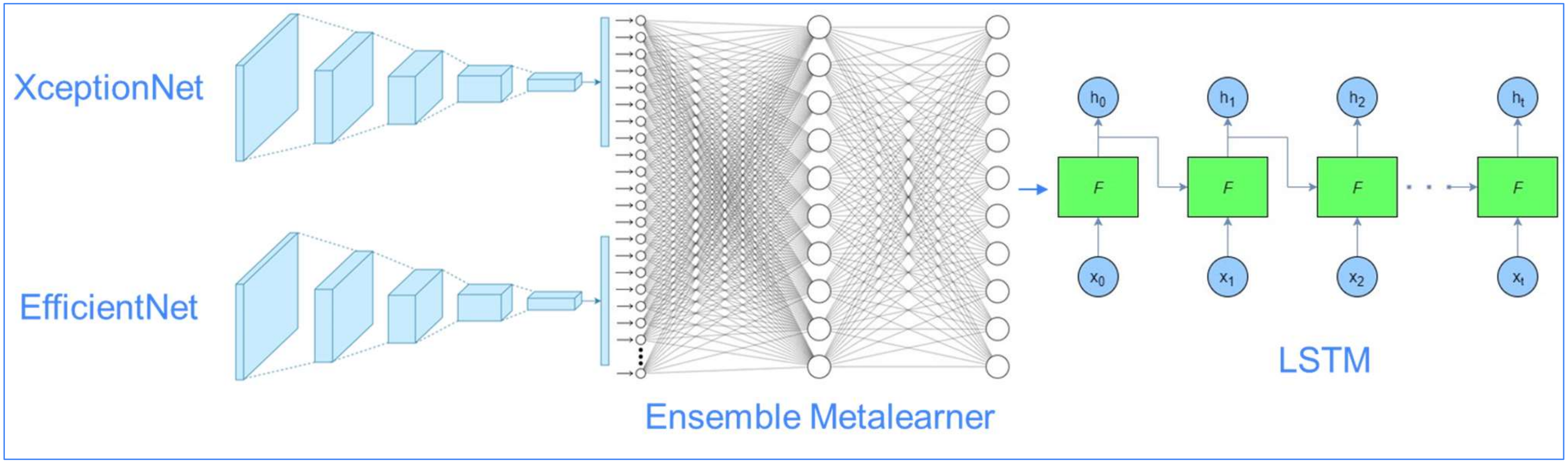
Initially considered but fell out of favor during model iteration due to substandard performance



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.2%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.7%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**83.0%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

†Not plotted

Our Novel Architecture



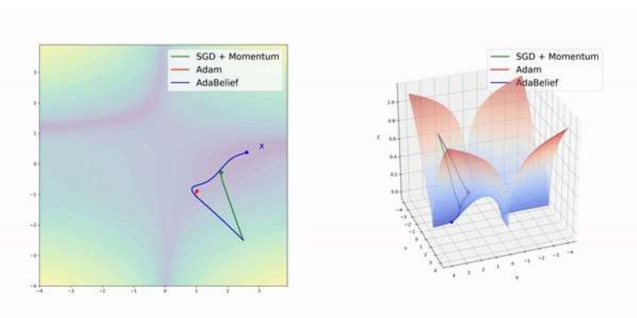
Model Assumptions

1. Ensembles provide better predictive power
2. Sequence classification will add robustness
3. May not generalize well to unseen forgery methods

Model Training & Tuning

Image Classifiers

- XceptionNet and EfficientNet (b4)
- All layers unfrozen
- AdaBelief optimizer
- Weight decay added
- Fully custom multithreaded streaming data loader



Meta-Learner

- Xception and EfficientNet fully trained and frozen
- Classifier final layers replaced with trainable dense layers
- Outputs fed into ensemble meta-learner
- Meta-learner trained using same hyperparameters as the classifiers

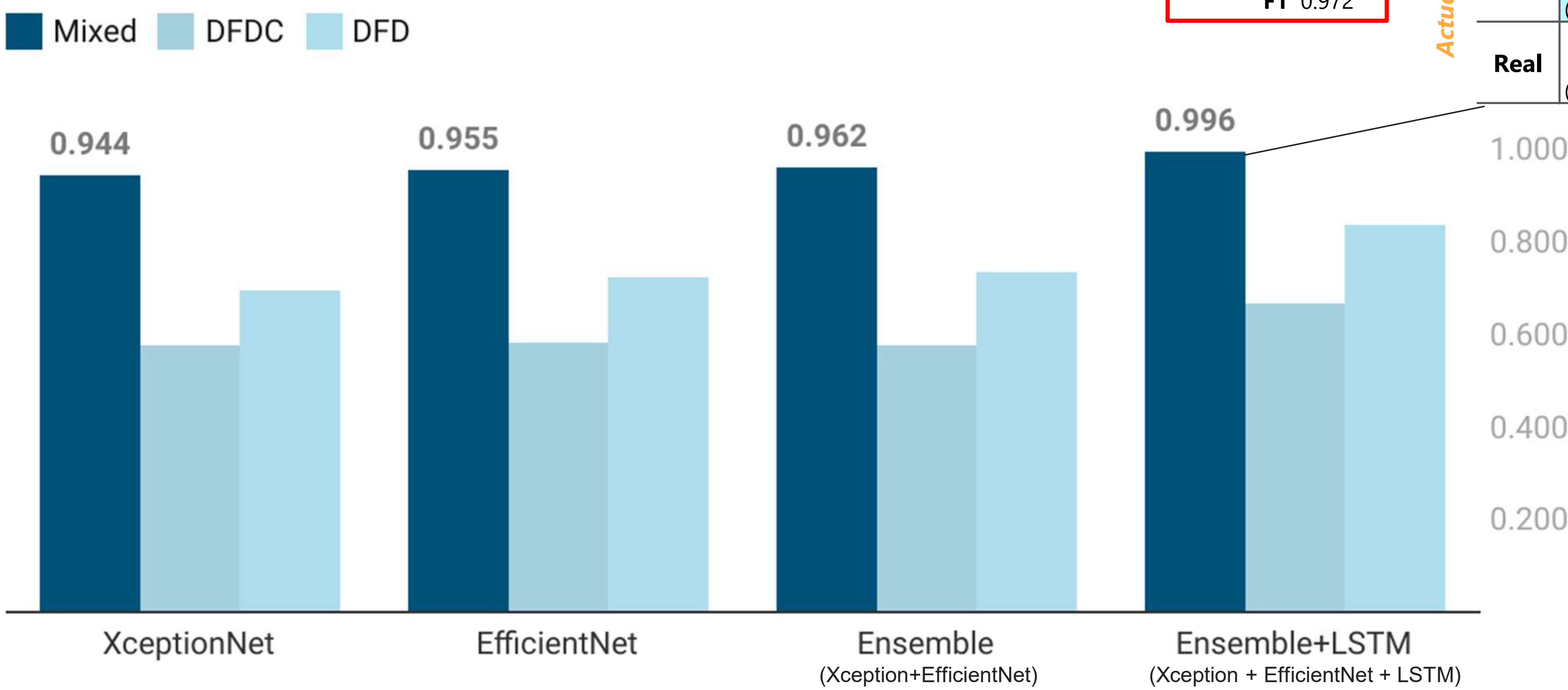
LSTM

- Analyzes a sequence of outputs from the meta-learner
- Training hyperparameters remained unchanged
- Classifies each 30-frame sequence as “real” or “fake”
- Model probabilities calibrated using temperature scaling

AUC for In-Distribution Test Sets

Precision 0.982
Recall 0.963
F1 0.972

		Predicted	
		Fake	**Real**
Actual	**Fake**	**TP** 4125 (0.98)	**FN** 75 (.02)
	Real	**FP** 157 (0.04)	**TN** 4043 (0.96)

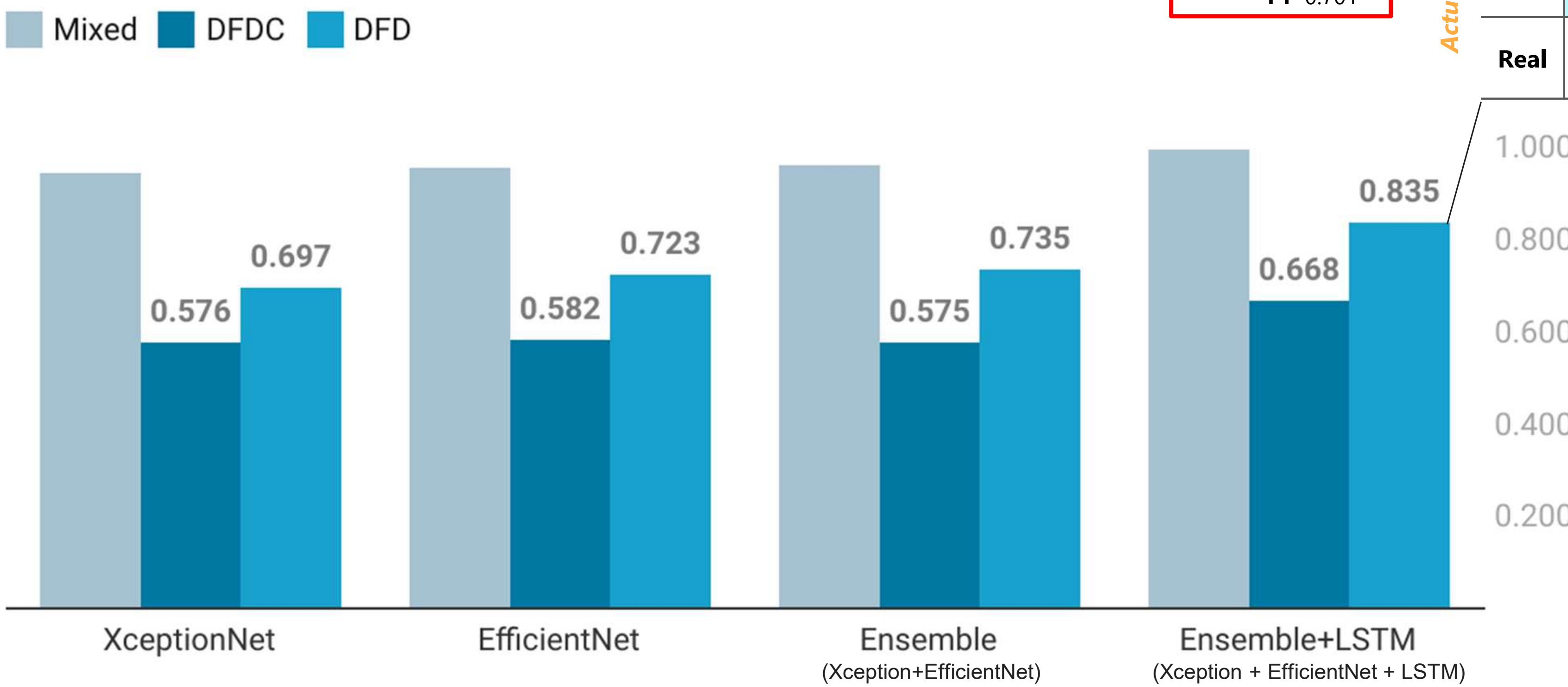


- Mixed: (Train/Test/Validation, in-distribution) Dataset comprised of randomly sampled videos from curated Celeb-DF and FF++ full datasets (Deepfakes, Face2Face, FaceSwap, FaceShifter, Neural Textures)
- DFDC: (Holdout, out-of-distribution) Deepfake Detection Challenge, random sample from full dataset
- DFD: (Holdout, out-of-distribution) FF++ Deepfake Detection; curated full dataset (original new actors different from FF++, not YouTube videos)

AUC for Out-of-Distribution Data

Precision	0.588
Recall	0.869
F1	0.701

		Predicted	
		Fake	**Real**
Actual	**Fake**	**TP** 2693 (0.87)	**FN** 407 (.13)
	Real	**FP** 1887 (0.46)	**TN** 2248 (0.54)



- Mixed: (Train/Test/Validation, in-distribution) Dataset comprised of randomly sampled videos from curated Celeb-DF and FF++ full datasets (Deepfakes, Face2Face, FaceSwap, FaceShifter, Neural Textures)
- DFDC: (Holdout, out-of-distribution) Deepfake Detection Challenge, random sample from full dataset
- DFD: (Holdout, out-of-distribution) FF++ Deepfake Detection; curated full dataset (original new actors different from FF++, not YouTube videos)

With Augmented Data

(Flickrfaces + Addl. Transformation – 2x train/val/test)

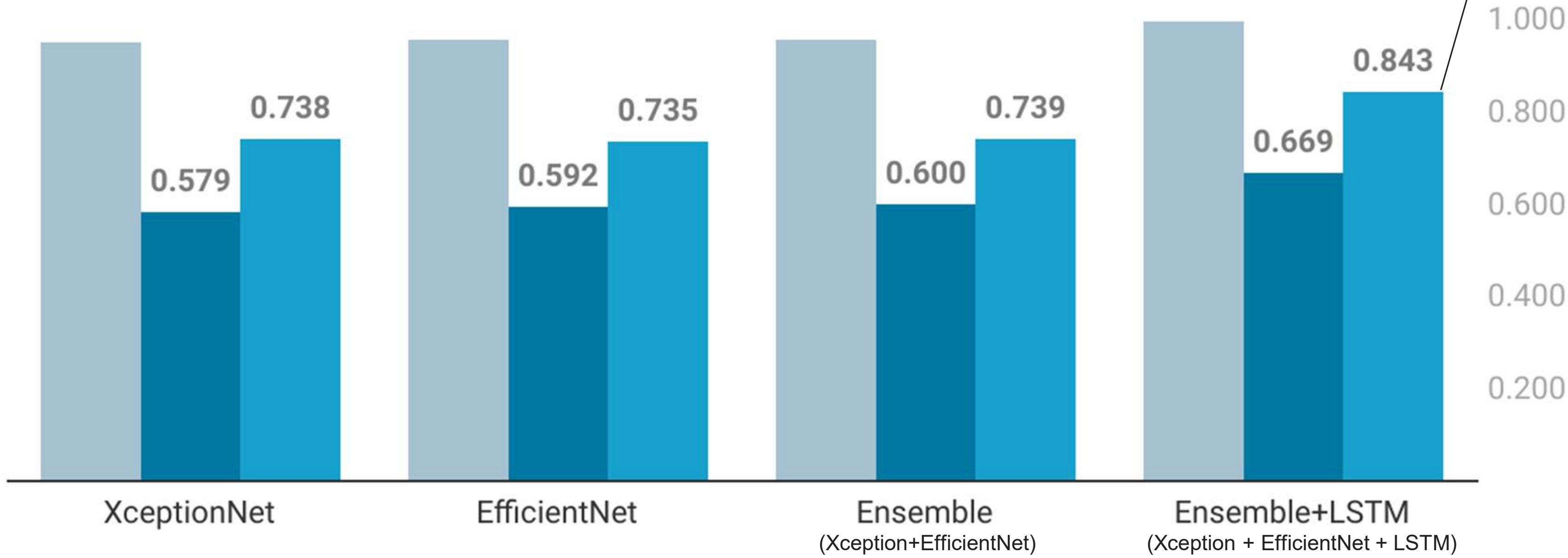
Mixed-Aug DFDC DFD

Before	
Precision	0.588
Recall	0.869
F1	0.701

After	
Precision	0.651
Recall	0.848
F1	0.736

Predicted / *Actual*

	Fake	**Real**
Fake	**TP** 2629 (0.85)	**FN** 471 (.15)
Real	**FP** 1408 (0.34)	**TN** 2727 (0.66)



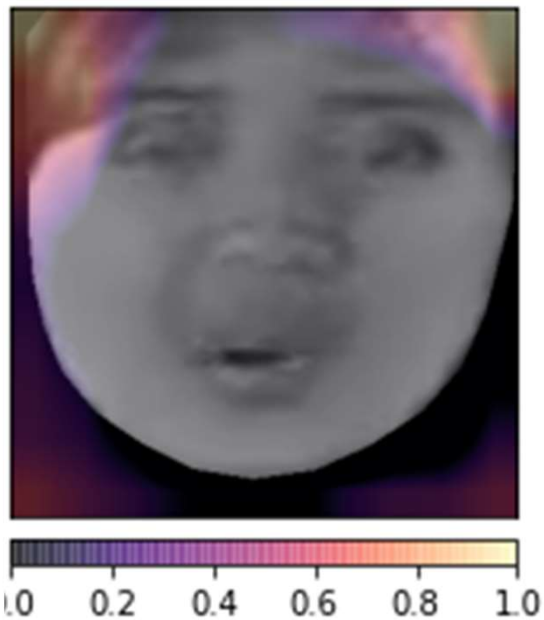
- Mixed: (Train/Test/Validation, in-distribution) Dataset comprised of randomly sampled videos from curated Celeb-DF and FF++ full datasets (Deepfakes, Face2Face, FaceSwap, FaceShifter, Neural Textures)
- DFDC: (Holdout, out-of-distribution) Deepfake Detection Challenge, random sample from full dataset
- DFD: (Holdout, out-of-distribution) FF++ Deepfake Detection; curated full dataset (original new actors different from FF++, not YouTube videos)

Comparison to SoTA Detectors on OOD Datasets

	Test Set	OOD	AUC
Tolosana et al (2020)	CDF		0.999
Oscar et al (2020)	CDF		0.997
Ours (ID)	Mixed		**0.996**
Ours (OOD)	DFD	✓	**0.843**
Lingzhi et al (2020)	CDF	✓	0.806
Yuval et al (2020)	CDF	✓	0.660
Dessa (2019)	FF++	✓	0.630

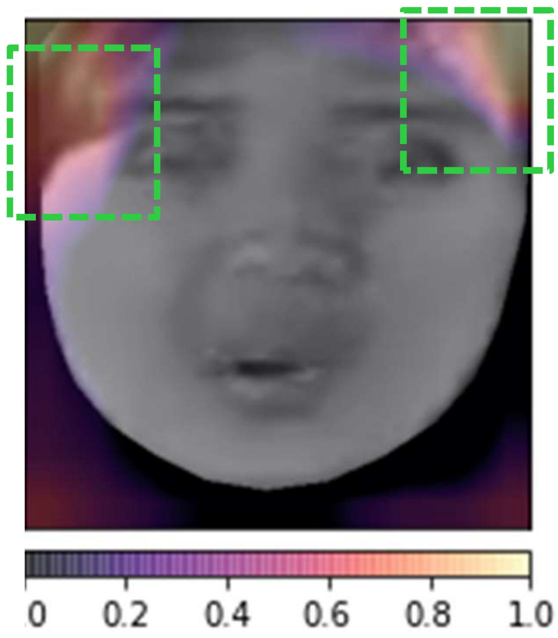
Model Interpretability

True Positive (Fake Image – Easy Detection)

Original Image	GradCAM
Predicted Probability: 1.0 	

Brighter region = more positive contribution to final prediction (steeper gradient to final conv. layer)

True Positive (Fake Image – Easy Detection)

Original Image	GradCAM
Predicted Probability: 1.0 	

Our detector easily discerns this poorly rendered deepfake (lighting differences)

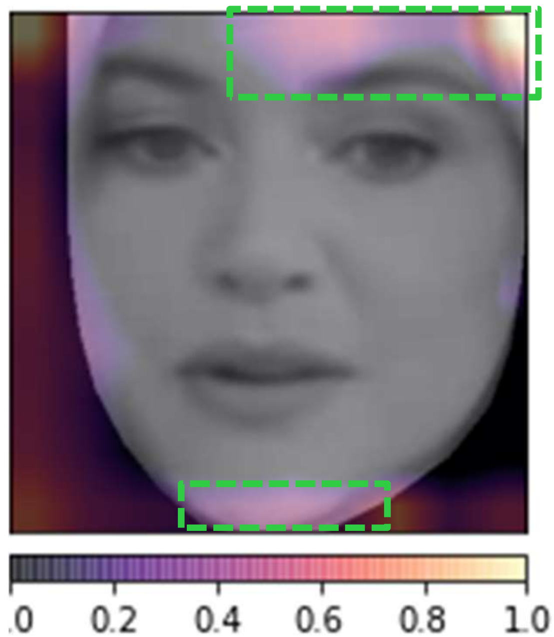
GradCAM, a **model interpretability algorithm**, reveals the facial regions that are most positively attributed to the prediction of "Fake"

True Positive (Fake Image – Difficult Detection)

Original Image

Predicted Probability: 1.0

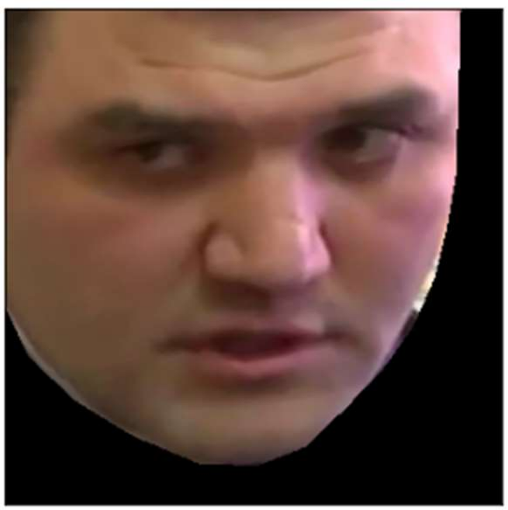
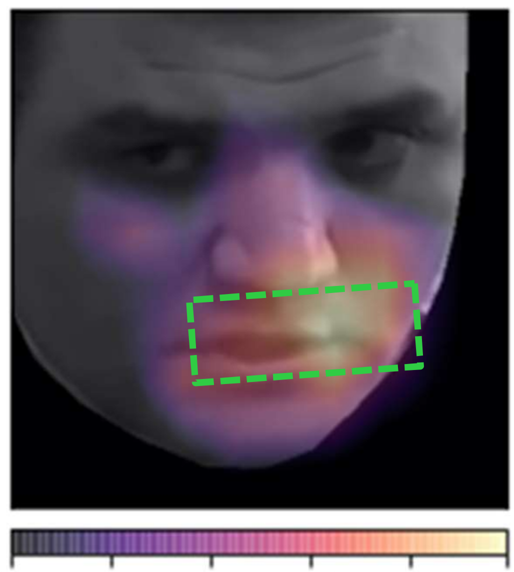
GradCAM



Here, the difference is less easily discerned, but our model is just as confident

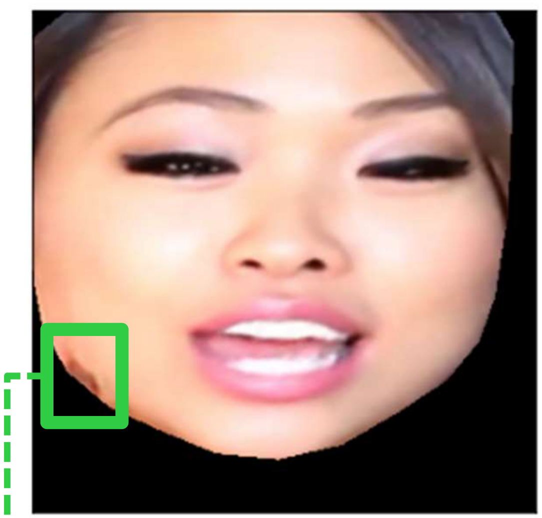
We expect artifacts/latent features to occur around the chin and forehead region due to the techniques mentioned in Face2Face's paper, which is confirmed by GradCAM

True Negative (Real Image)

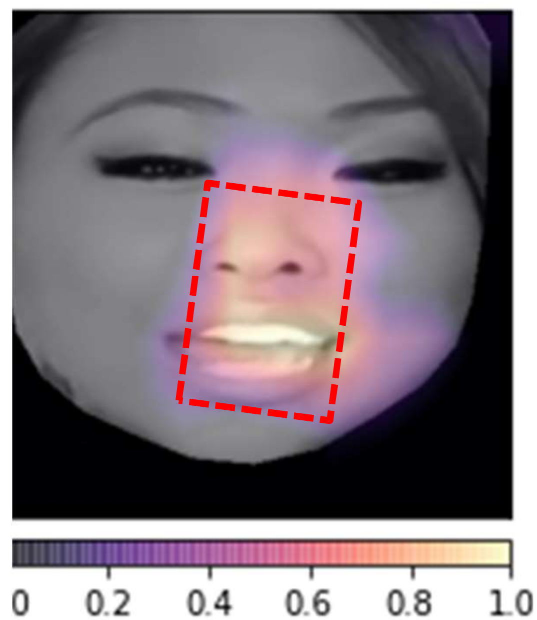
Original Image	GradCAM
Predicted Probability: 1.0 	

False Negative (Fake Image)

Original Image	GradCAM
Predicted Probability: 0.99	



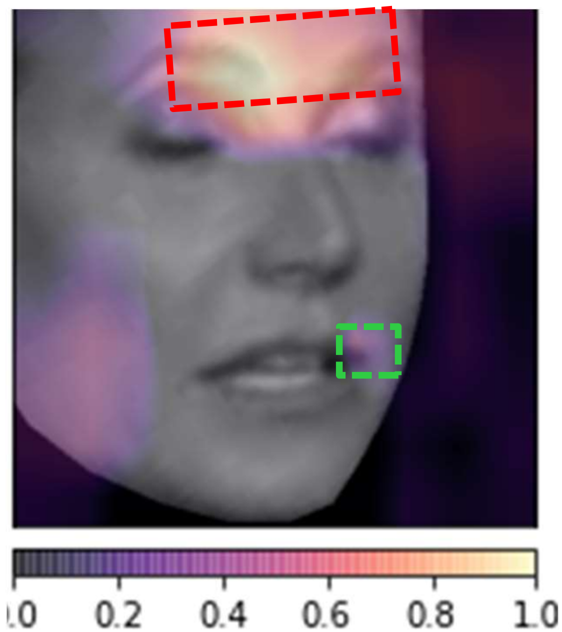
Needs to be tuned to better identify artifacts such as this



Here, the model is extremely confident that this is a real image, however it is **wrong**

Our model potentially placed too much weight on this region or was tricked by expert blending for this frame

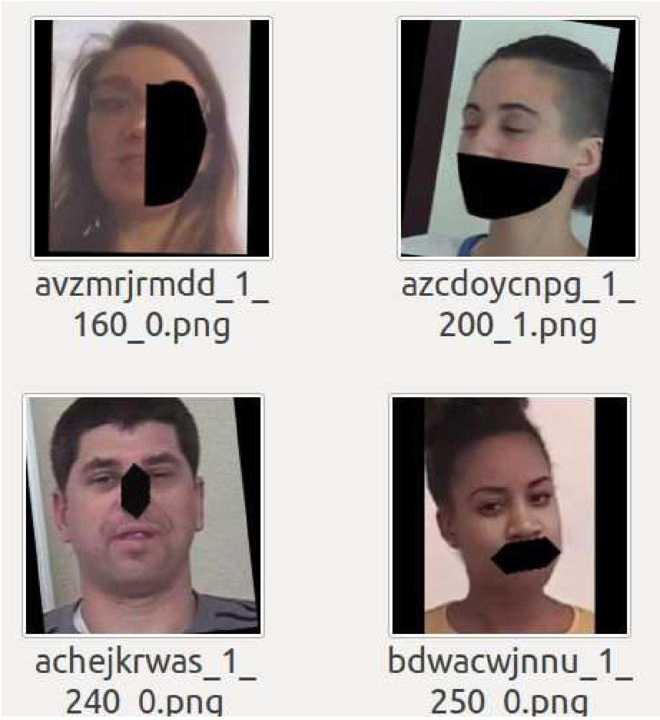
False Positive (Real Image)

Original Image	**GradCAM**
Predicted Probability: 1.0 	

The model gets confused here and labels a real image as fake

GradCAM tells us that the upper browline + upper left lip appeared similar to previously trained fake frames, potentially due to poor image quality or some naturally occurring blemish caused by movment

Future Improvements Based on GradCAM



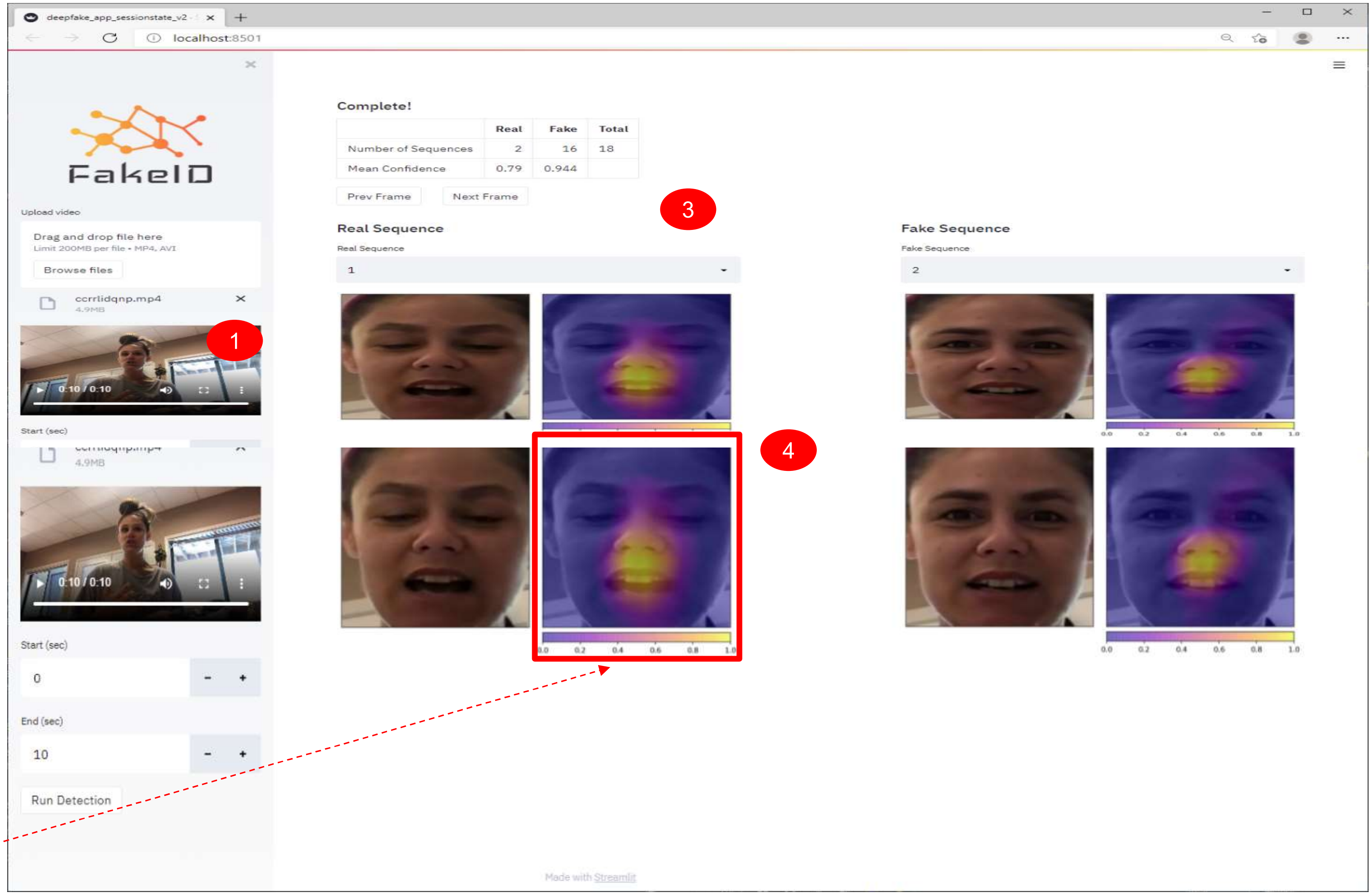
1) Blackout random regions of the face
2) Focus on conv layers that address outer edges of face
3) Augment dataset with more images that have that type of artifact/blemish

Deployment (Detection Web App)

FakeID

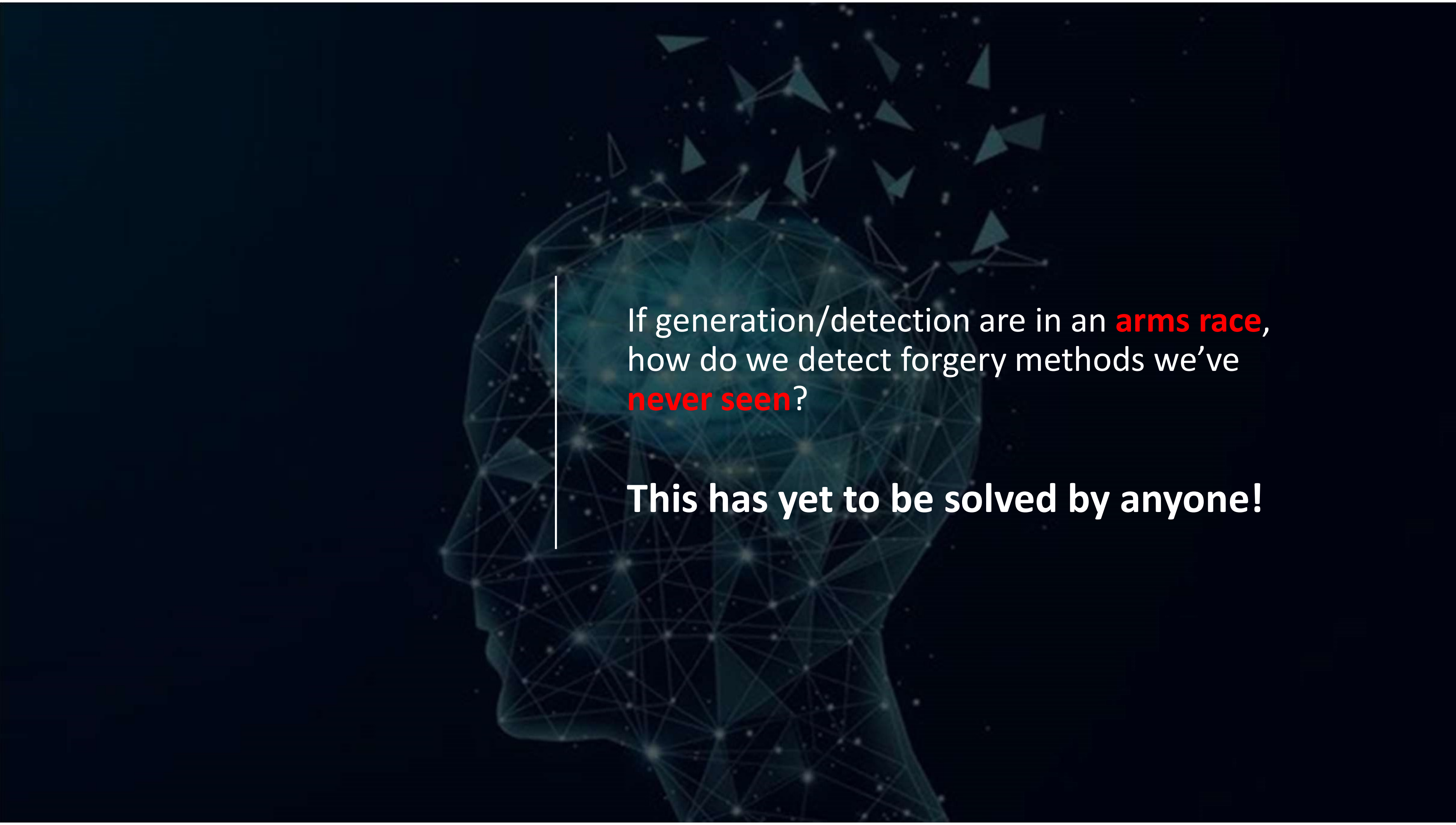
Powered by

Streamlit



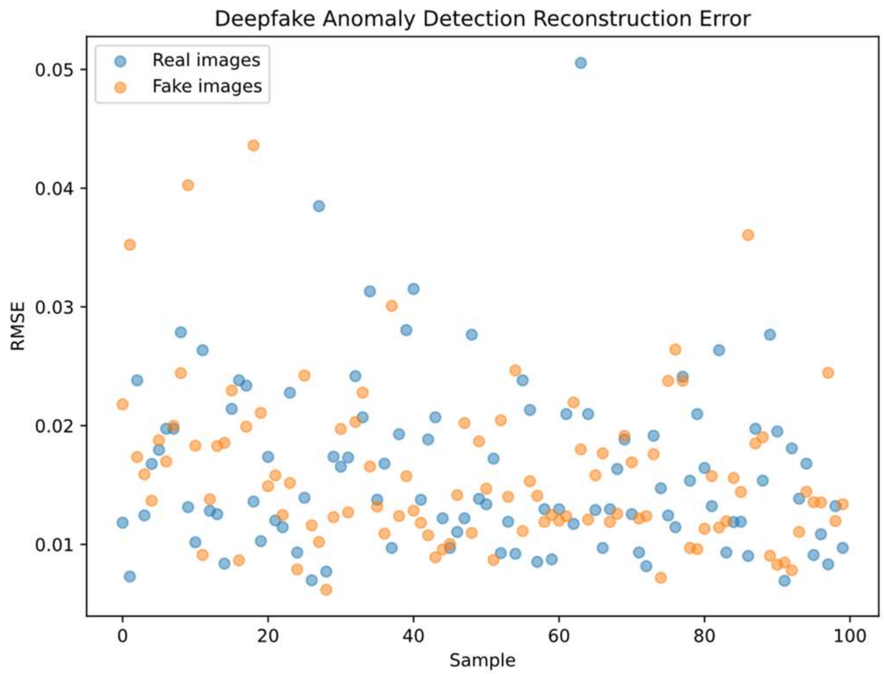
1. User uploads and selects a subset of a video
2. Video is parsed and still frames propagate through the pipeline
3. The app then displays basic classification results
4. GradCAM interpretability algorithm results are displayed

Addressing Generalizability

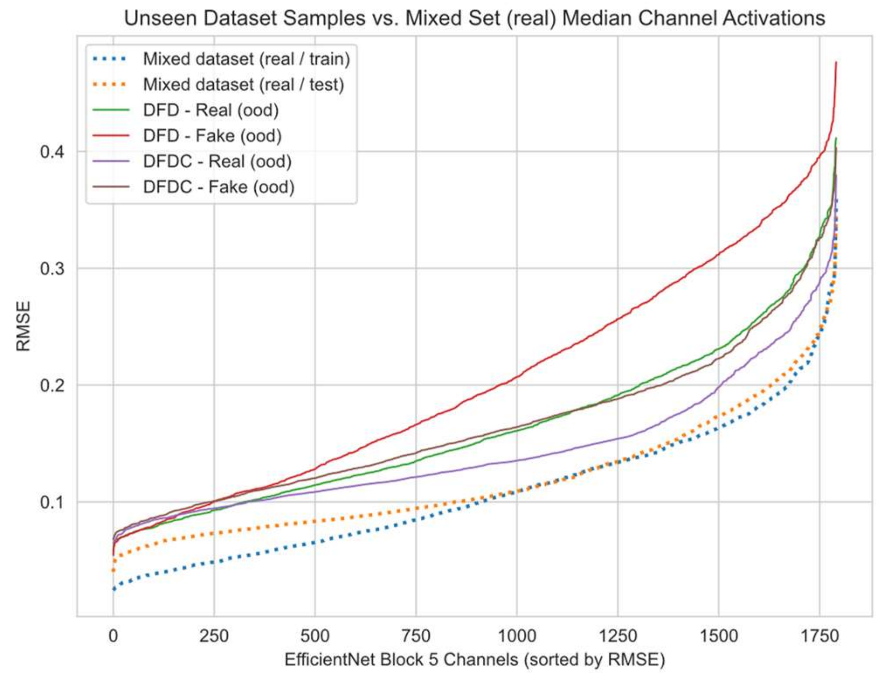

If generation/detection are in an **arms race**, how do we detect forgery methods we've **never seen**?
This has yet to be solved by anyone!

Two Approaches

Image outlier detection: We know what authentic video frames look like, **flag** videos that do not conform



Neural activation analysis: Model misclassifications on never-before-seen forgery methods may be identified by investigating the **inner workings** of our detection model



Final Thoughts and Future Work

Contributions

	Tolosana Et Al. (2020)	1st Place DFDC Winner	Dessa (2019)	Wang Et Al. (2020)	Our Method
Utilizes CNN's and image preprocessing techniques (augmentation, cropping)	✓	✓	✓	✓	✓
Trained on datasets that use Face2Face, FaceSwap, Deepfake, and NeuralTextures techs.	✓	✓	✓	✓	✓
Transfer Learning using ***Xception***	✓		✓	✓	✓
Transfer Learning using ***Efficientnet***		✓			✓
Trained on ***Mixed Datasets*** for greater generalizability			✓		✓
Transfer Learning using ***3D CNN's***				✓	
Utilizes ***LSTM*** to process sequences of frames					✓
Ability to detect ***multiple subjects*** per frame					✓
Uses ***Ensemble Meta-learner*** that increases performance and allows plug/play new models					✓
Developed a ***functional app*** with built-in ***model interpretability*** algorithms					✓

Future Work

Improve on Deep Learning Models

3D CNN
Capture spatio-temporal features

Siamese Network
Distinguishes unique facial features

Optical Flow
Granular pixel-to-pixel prediction

Additional Datasets for Diversification

Demographics
Race
Age

Deeper Forensics
Full face swapping
Additional data augmentation

Online Deep Learning
Increase data availability for model training

Thank you!