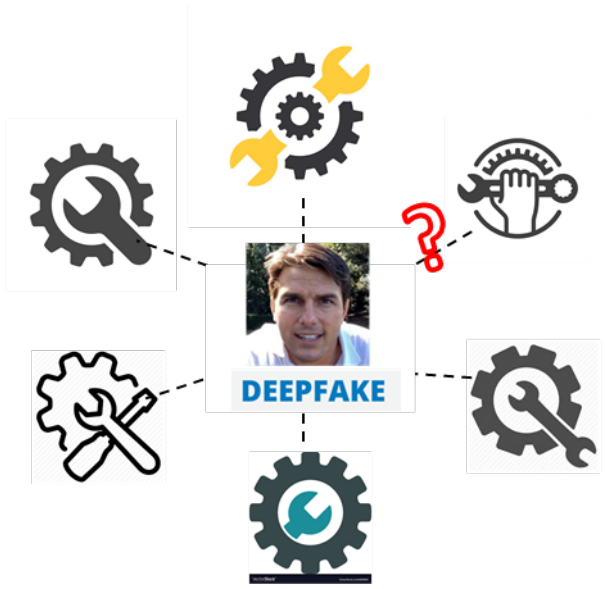
Generation Model Attribution of Face-swapped Deepfake Videos



Shan Jia (University at Buffalo)

Professor Xin Li (West Virginia University)

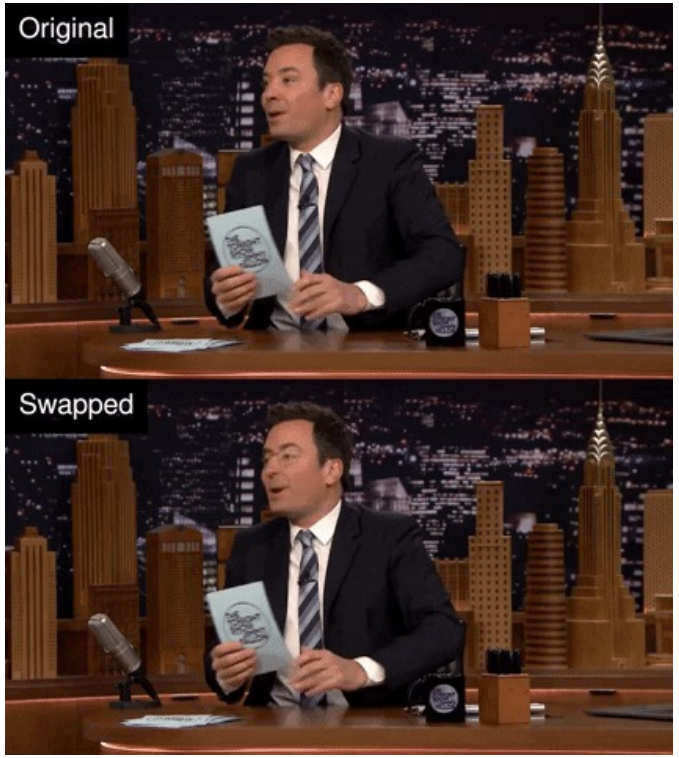
Professor Siwei Lyu (University at Buffalo)

12-07-2021

Deepfake Videos

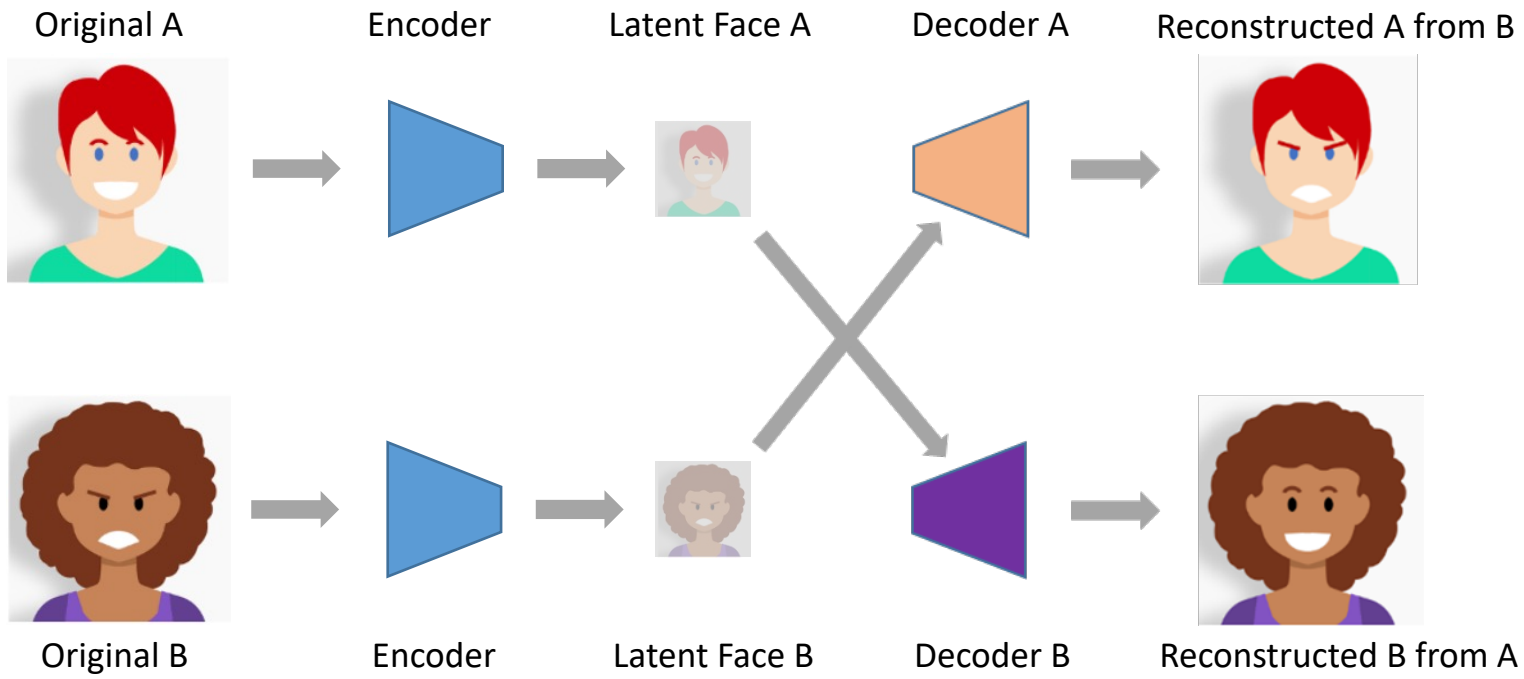
- Synthesized videos with face identity replacement based on deep learning (beginning from Reddit, Nov. 2017)
- Great threat! Can be used to create fake news, financial fraud, and malicious hoaxes, etc.

htt



Deepfake Videos Generation

- Deepfakes: using Autoencoder for face reconstruction and swapping
- Source video: the source content used to extract the identity that will be swapped onto the target video
- Target video: the base video in which a face will be swapped

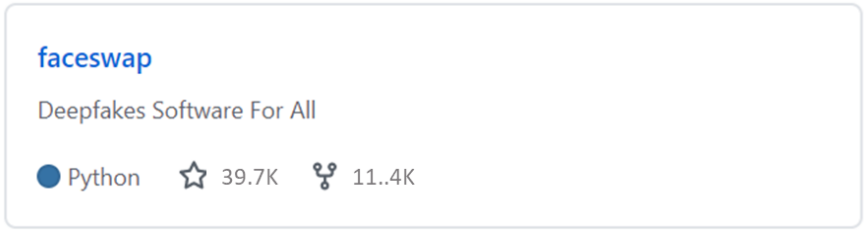


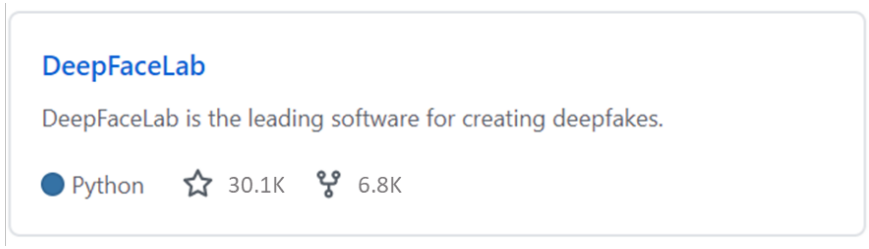
Research on Deepfake Videos

- Generation models/tools:

  Open-source tools: *Faceswap (11 models); DeepFaceLab (2 models), etc*.

  Publicly available datasets: *FaceForensics++, 2019, 4 models; Celeb-DF, 2020, 1 model; DeeperForensics-1.0, 2020, 1 model; DFDC, 2020, 8 models; etc*.





Dataset	#Deepfakes	#Sub	#Model	Model label?
UADFV, 2019	49	49	1	-
Deepfake-TIMIT, 2019	640	43	1	-
FaceForensics++, 2019	3,000	977	1	-
DFDC Preview, 2019	5,244	66	2	No
DeepfakeDetection, 2019	3,068	28	-	No
Celeb-DF, 2020	5,639	59	1	-
DeeperForensics-1.0, 2020	1,000	100	1	-
DFDC, 2020	104,500	960	6	No
WildDeepfake, 2020	3,509	-	-	No
FakeAVCeleb, 2021	20,000+	600+	3	No
DFDM (ours)	**10,320**	**59**	**8**	**Yes**

- Limitations:
- Most with only one Deepfakes generation models;
- No model labels for model attribution.

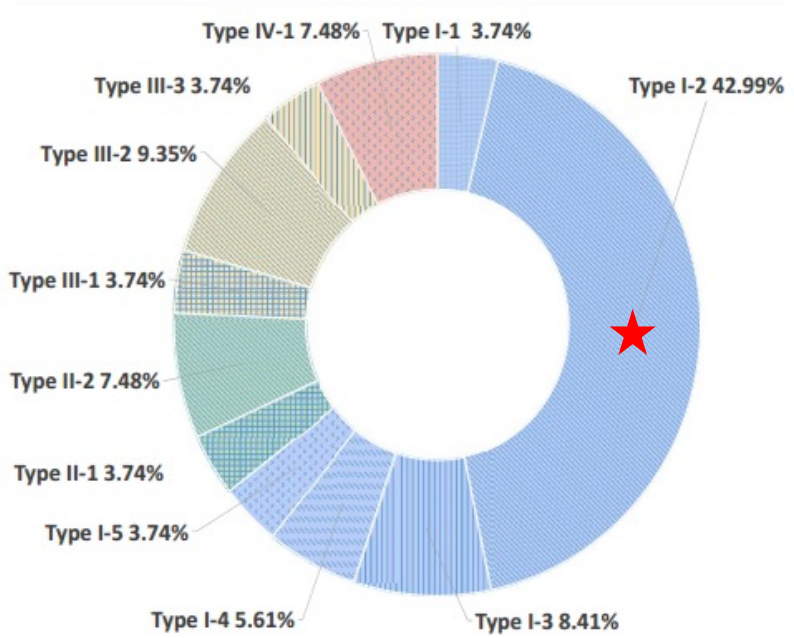
Research on Deepfake Videos

- Detection methods:

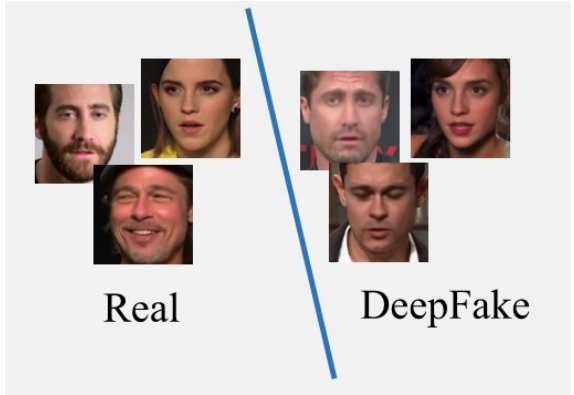
  Most detection methods focus on binary classification, i.e., real vs. fake;
  MoseNet, 2018; HeadPose, 2019; Xception, 2019; Capsule, 2019; DSP-FWA, 2020; Face X-ray, 2020; MAT, 2021; etc.

  Showing performance differences on datasets with different manipulation tools.

Type ID	Name
Type I1	Image Forensics based Detection
Type I2	DNN-based Detection ★
Type I3	Obvious Artifacts Clues
Type I4	Detection and Localization
Type I5	Facial Image Preprocessing
Type II1	GAN-based Artifacts
Type II2	Frequency Domain
Type III1	Visual-audio Inconsistency
Type III2	Visual Inconsistency
Type III3	Biological Signal in Video
Type IV1	Others



Juefei-Xu, et al., 2021



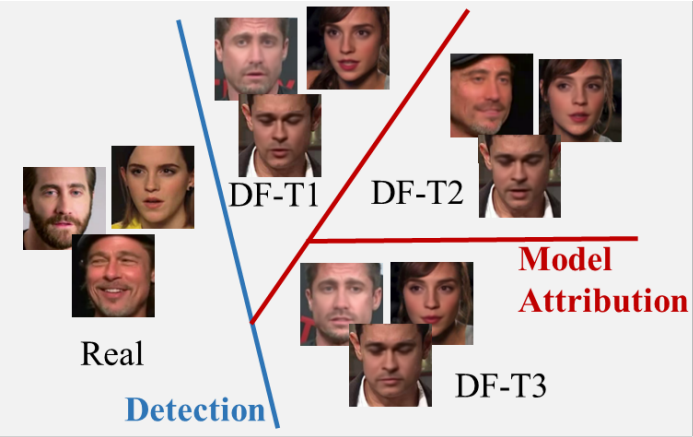
- Limitations:
- Only for binary classification

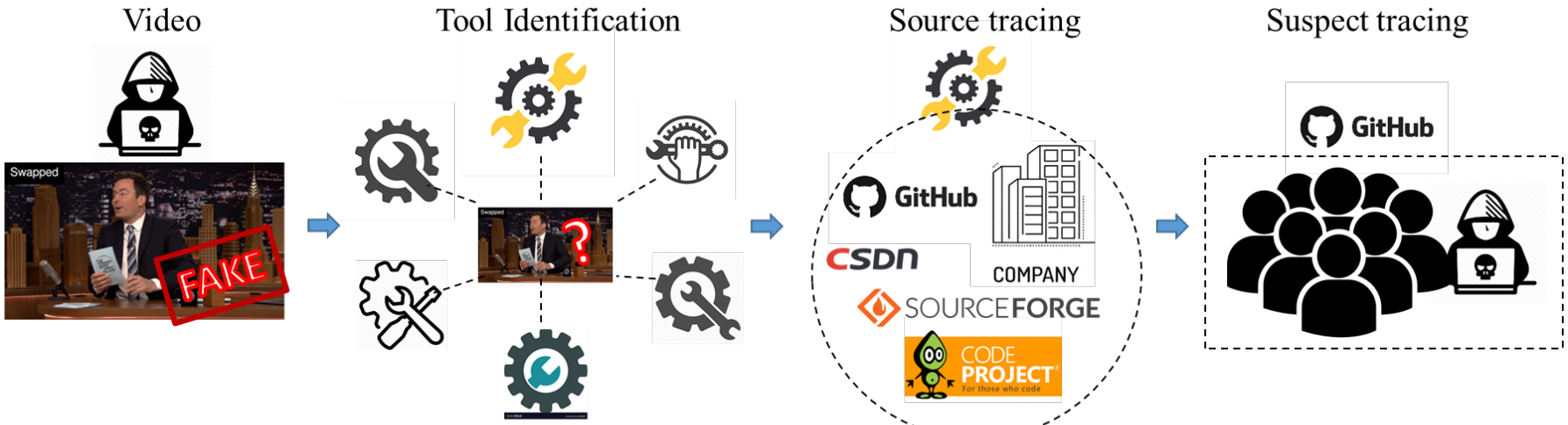
Research on Deepfake Videos

- Motivations: Beyond detection, model attribution of Deepfakes also matters!

To explore if and how different generation models influence the Deepfake videos;

To identify deepfakes source (model attribution) for further forensics.



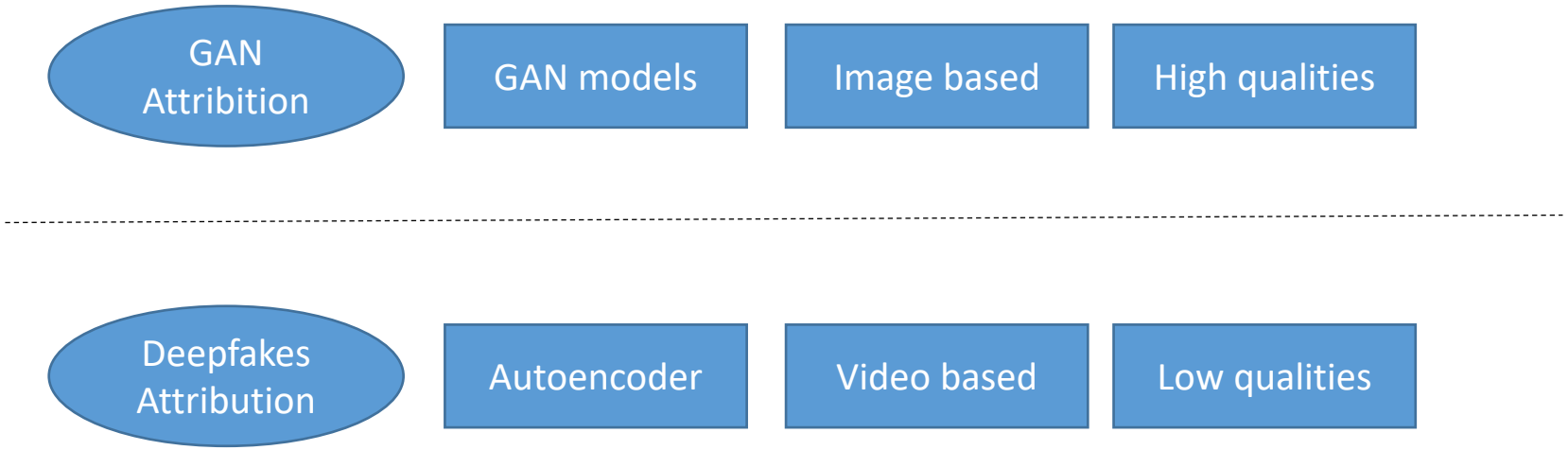


Application example in forensics

Generation Model Attribution

- Model attribution: identifying the source of Deepfakes
- GAN model attribution:

  Marra, et al., 2019; Yu, et al., 2019; Goebel, et al., 2020; Wang, et al., 2020; Girish, et al., 2021; etc.
- No previous work on Deepfakes model attribution



Model Attribution of Deepfake videos

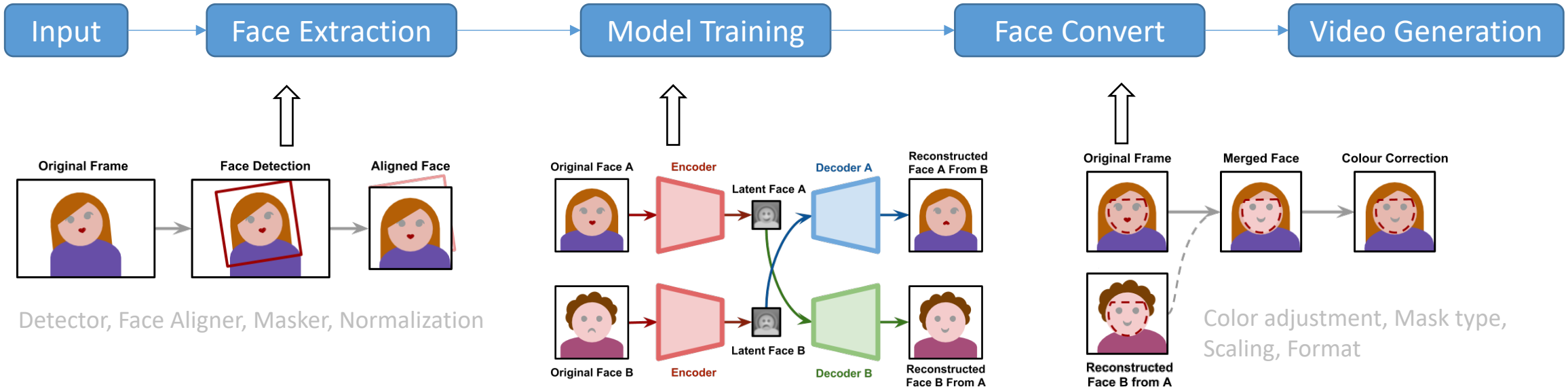
- Our research:

  1) A new dataset: DFDM (Deepfakes generated by different models)

  2) A novel method: DMA-STA (Deepfakes model attribution based on spatial and temporal attention)

DFDM Dataset Generation

- Deepfakes generation process: model-level (keep other modules the same);



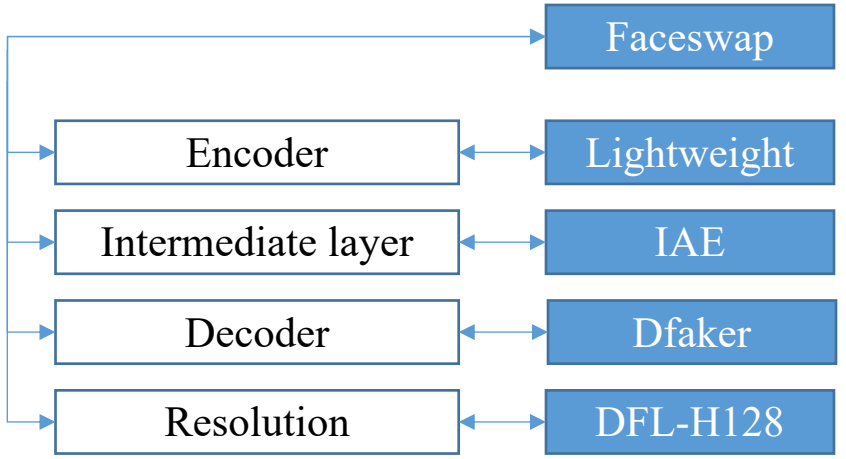
DFDM Dataset Generation Models

- Faceswap platform based: the most popular, with various models;
- Five models: Faceswap; Lightweight; IAE; Dfaker; DFL-H128.
- Selection criteria:

  1) Use Faceswap as baseline, select models with only one variation for the most subtle model attribution;

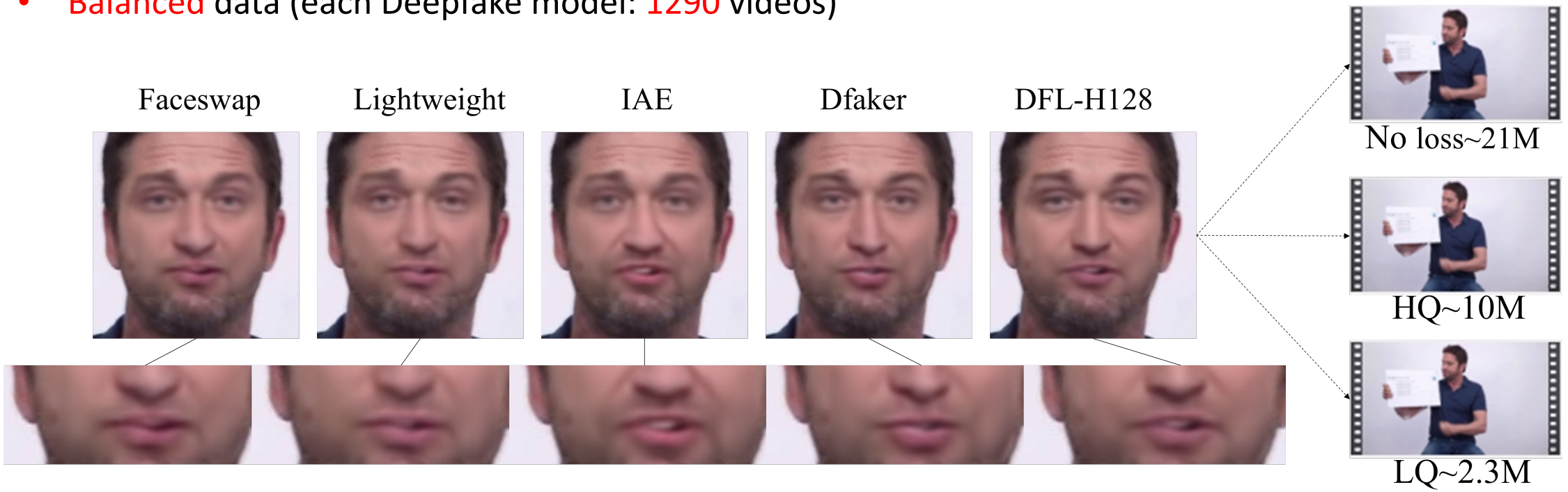
  2) Select variations in encoder, decoder, intermediate layer, input resolution, respectively.

Model	Input	Output	Encoder	Decoder	Others
Faceswap	64	64	4Conv+1Ups	3Ups+1Conv	/
Lightweight	64	64	3Conv+1Ups	3Ups+1Conv	For low-end card
IAE	64	64	4Conv	4Ups+1Conv	Intermediate layers; Shared E&D
Dfaker	64	128	4Conv+1Ups	4Ups+3Residual+1Conv	
DFL-H128	128	128	4Conv+1Ups	3Ups+1Conv	Same with Original



DFDM Dataset Details

- Original videos: 590 videos from Celeb-DF, 2020
- 6450 Deepfakes: 430 videos * 5 generation tools * 3 compression rates
- Balanced data (each Deepfake model: 1290 videos)

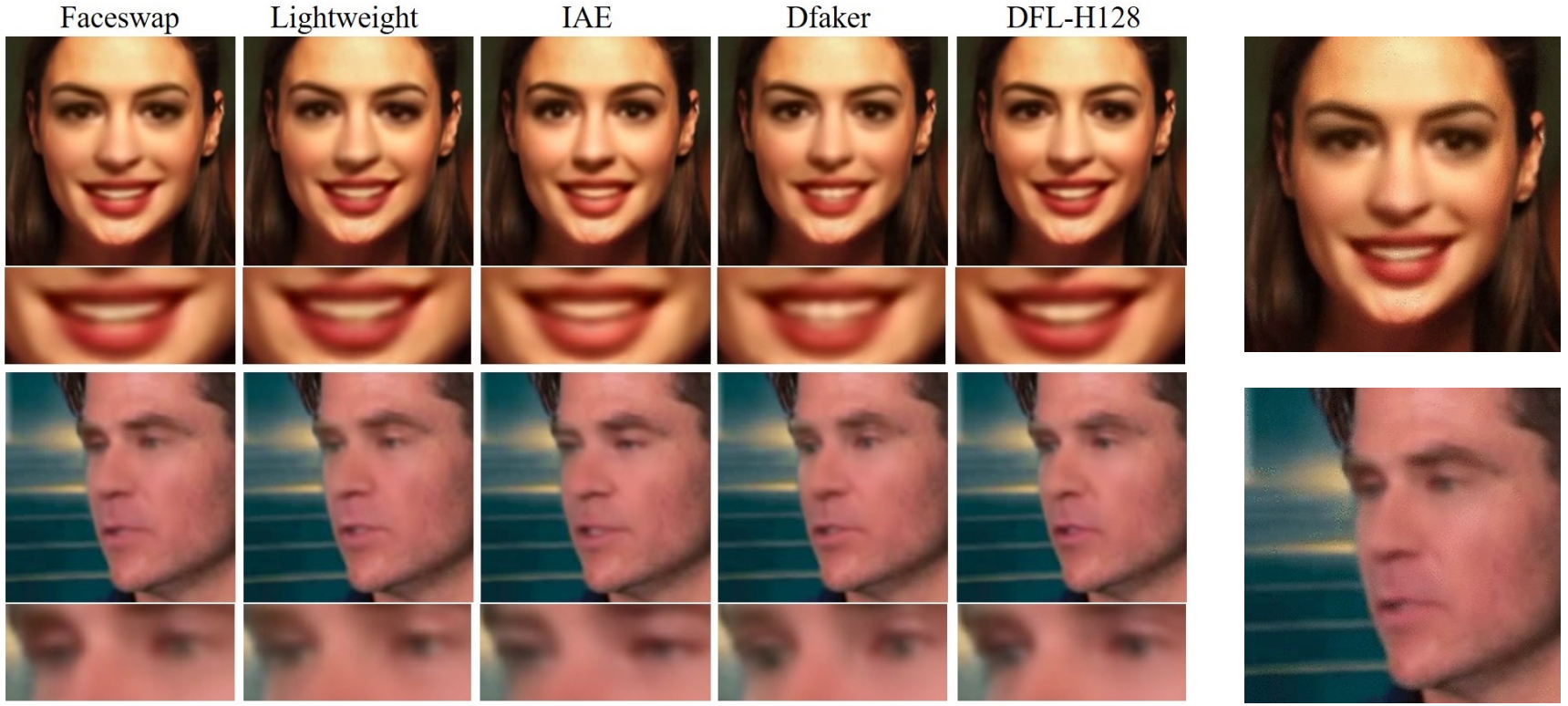


Observable Visual Differences:

- Local: Eye direction; Nose shape; Mouth region; Teeth;
- Global: Skin texture; Blurriness; Sharpness.

DFDM Dataset Examples

- More face examples



- Inter-class: Do different generation models result in statistically distinguishable Deepfake videos?
- Intra-class: Are the differences of the same generation model consistent and detectable from the input video?

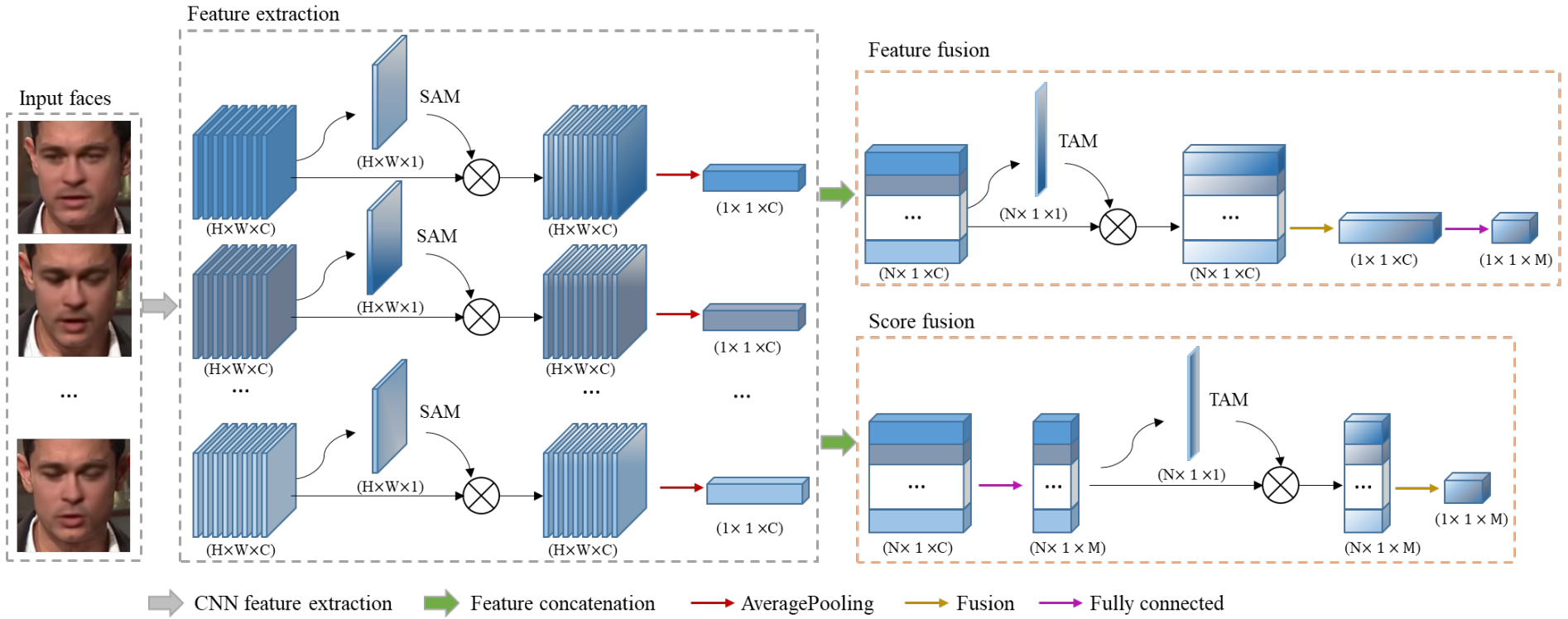
DMA-STA: Model Attribution Method

- Motivations:

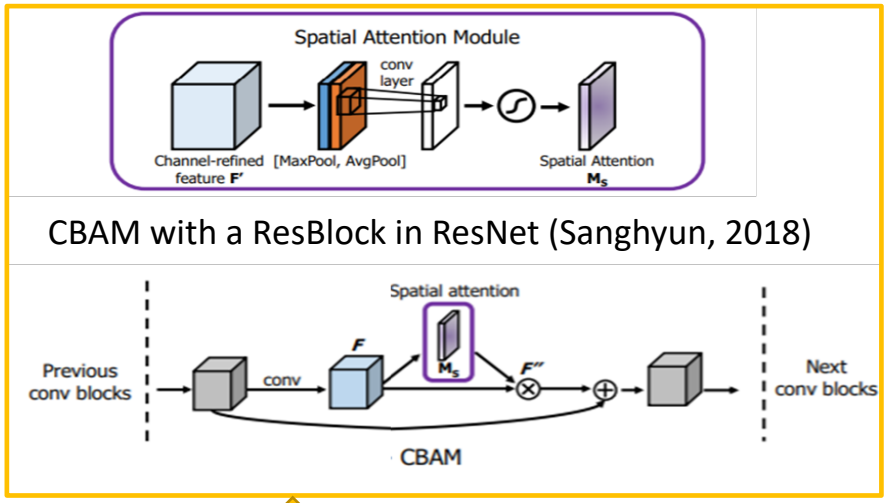
  Model attribution with subtle differences among different categories;

  Video-level analysis helps extract robust differences than frame-level analysis;

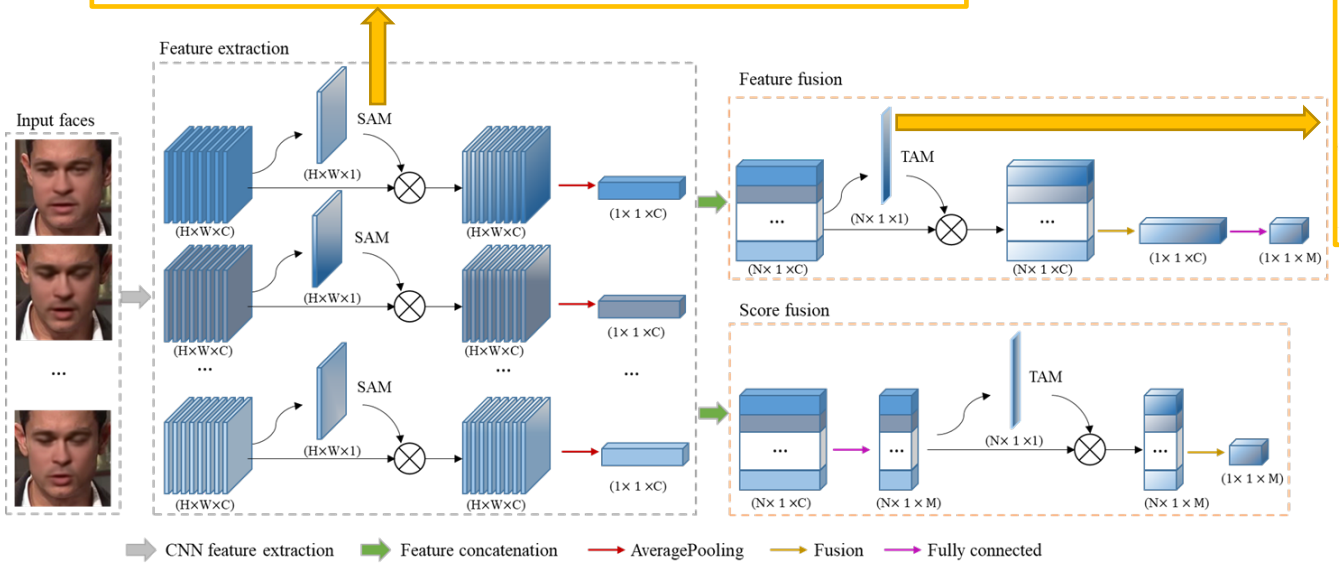
- DMA-STA: combining spatial attention (frame-level) with temporal attention (video-level) for discriminative features extraction

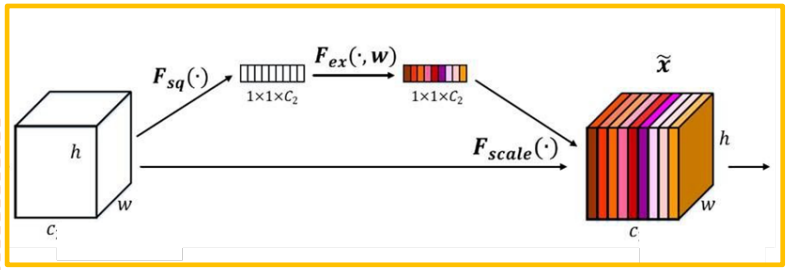


DMA-STA: Model Attribution Method



- CNN feature extractor: ResNet50
- SAM: spatial attention map
- TAM: temporal attention map





TAM based on SEnet (Hu, et al., 2018)

$$f'_v = \frac{\sum_{i=1}^{n} \alpha_i f_i}{\sum_{i=1}^{n} \alpha_i}$$

Experimental Evaluation (Ablation Studies)

- Frame number influence
- Data: DFDM HQ dataset (2150 videos, 145~740 frames, train:test = 7:3)

Classification Accuracy (%)

# Frame	Feature Fusion						Score Fusion					
	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128
1	45.58	41.13	31.78	51.16	55.04	50.39	44.81	35.48	33.33	48.06	60.47	48.06
3	62.33	59.69	31.01	79.84	85.27	58.06	61.71	49.61	41.09	74.42	79.84	66.13
5	64.19	58.14	37.98	69.35	79.84	78.29	65.27	51.16	44.96	75.97	86.05	70.97
10	71.94	63.57	58.91	66.67	82.95	87.60	69.15	33.33	74.42	68.22	87.60	85.48
15	69.46	45.16	57.36	82.95	90.70	72.87	72.25	45.97	62.79	80.62	87.60	86.05
20	70.23	51.94	70.54	59.68	89.15	82.17	68.53	56.45	51.16	78.29	90.70	68.22

- Larger frame number, higher classification accuracies;
- Using 10 frames to balance the accuracy and efficiency.

Experimental Evaluation (Ablation Studies)

- Ablation studies: Attention scheme influence
- Data: DFDM HQ dataset (2150 videos, train:test = 7:3)

Classification Accuracy (%)

Attention Scheme	Feature Fusion						Score Fusion					
	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128
ResNet50	68.02	54.84	57.36	70.54	89.92	70.54	67.60	51.16	54.26	79.84	82.95	72.87
SA+Ave	68.84	58.87	51.16	76.74	80.62	79.07	68.89	54.26	54.26	72.58	82.17	84.50
CBAM[1]	68.53	52.42	63.57	69.77	84.5	74.42	68.06	58.14	43.41	75.81	82.95	82.95
SA+FA[2]	66.82	64.34	42.64	76.61	74.42	79.07	67.29	54.03	53.49	65.89	83.72	81.4
Ours	71.94	63.57	58.91	66.67	82.95	87.60	69.15	33.33	74.42	68.22	87.60	85.48

[1] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." 2018 CVPR.
[2] Meng, Debin, et al. "Frame attention networks for facial expression recognition in videos." 1029 ICIP.

Experimental Evaluation (Comparison)

- Comparison results of Deepfakes model attribution performance
- Data: HQ DFDM (5 classes)
- All feature fusion based (10 frames)

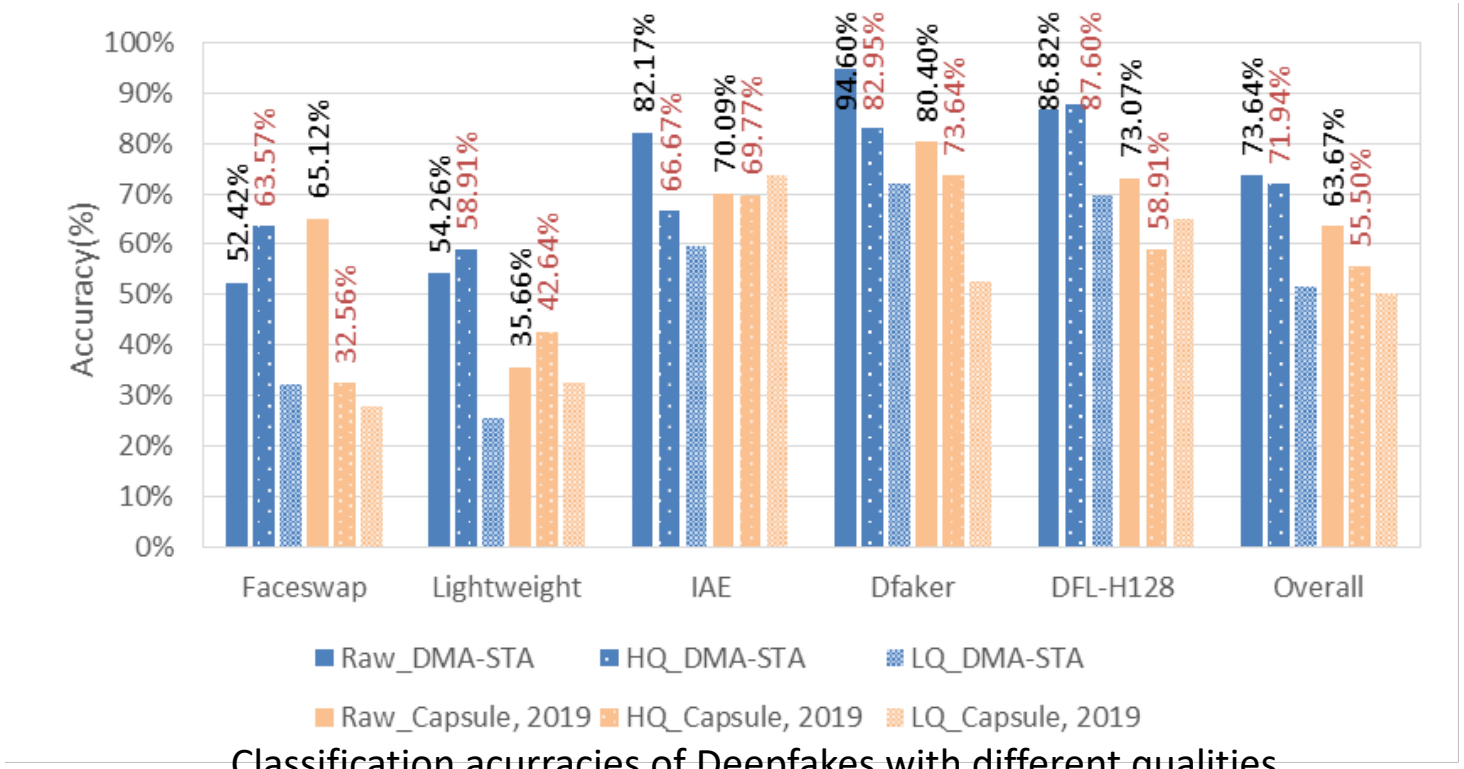
Classification Accuracy (%)

Method	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128
MesoInception-4, 2018	20.93	6.98	2.33	79.07	79.07	4.65
Capsule, 2019	55.50	32.56	42.64	69.77	73.64	58.91
Xception, 2019	20.93	0.77	0.00	12.4	12.4	19.38
R3D, 2020	21.40	27.13	25.58	15.5	20.16	18.61
DSP-FWA, 2020	23.41	17.05	7.75	43.41	40.31	8.87
Ours: DMA-STA	71.94	63.57	58.91	66.67	82.95	87.60

- Most failed in model attribution task (<25%):
- The proposed DFA-STA scheme achieved the best results.

Experimental Evaluation (Quality)

- Comparison results of Deepfakes model attribution on videos with different qualites;
- DMA-STA method vs Capsule Network (2019);



Classification acurracies of Deepfakes with different qualities

- Compression rates influence the model attribution performance a lot!
- Our method:

  Raw (73.64%)

  HQ (71.94%)

  LQ (51.58%)

Experimental Evaluation (Comparison)

- Comparison of Deepfakes model attribution & detection performance
- Data: HQ DFDM + Real videos in Celeb-DF (6 classes)
- All feature fusion based (10 frames)

Classification Accuracy (%)

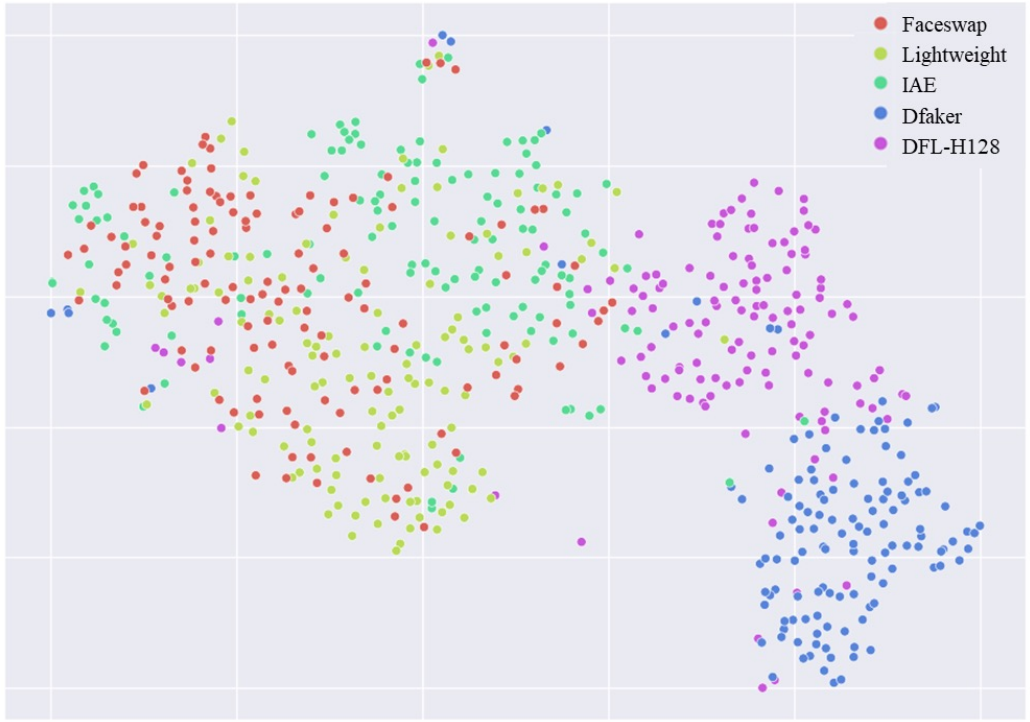
Method	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128	Real
MesoInception-4, 2018	33.33	14.73	1.55	18.6	18.6	64.34	98.45
Capsule, 2019	64.99	51.16	43.41	55.04	70.54	69.77	100.00
Xception, 2019	20.54	63.57	0.00	0.00	0.00	0.77	41.09
R3D, 2020	26.10	33.33	24.81	24.81	8.53	9.30	55.81
DSP-FWA, 2020	45.48	31.01	43.41	32.5	36.43	39.53	92.25
Ours: DMA-STA	74.03	51.16	58.91	65.12	87.60	81.40	100.00

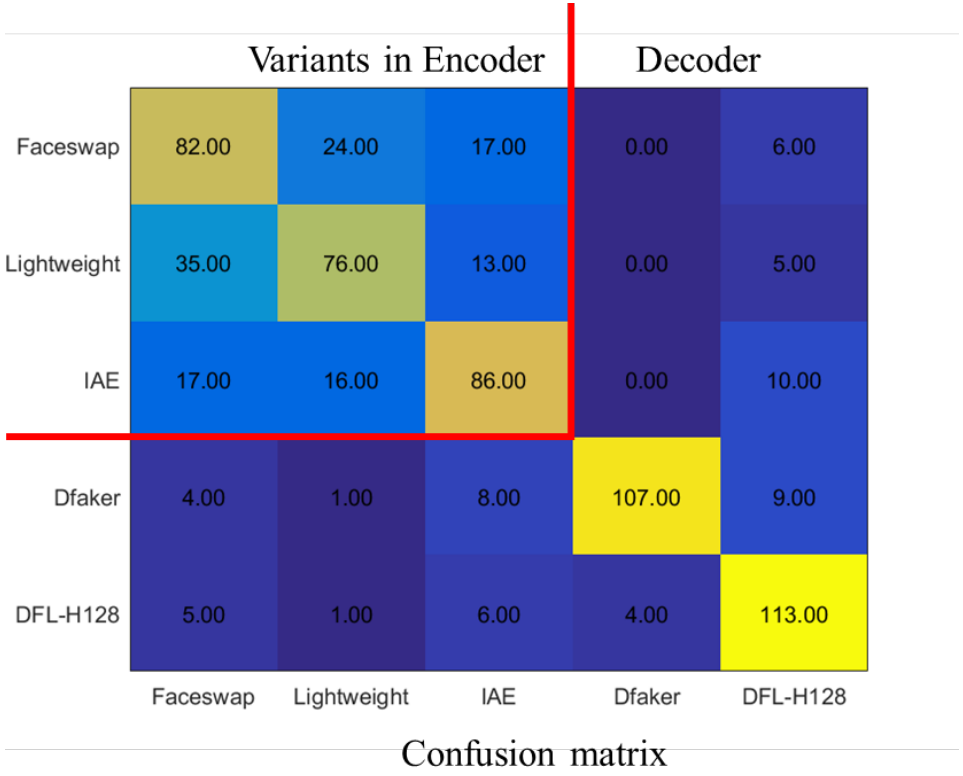
- The proposed DFA-STA scheme achieved the best results, especially with 100% accuracy for Real videos detection.

Visualization Results

- Model attribution performance (5 classes)

Method	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128
Ours: DMA-STA	71.94	63.57	58.91	66.67	82.95	87.60



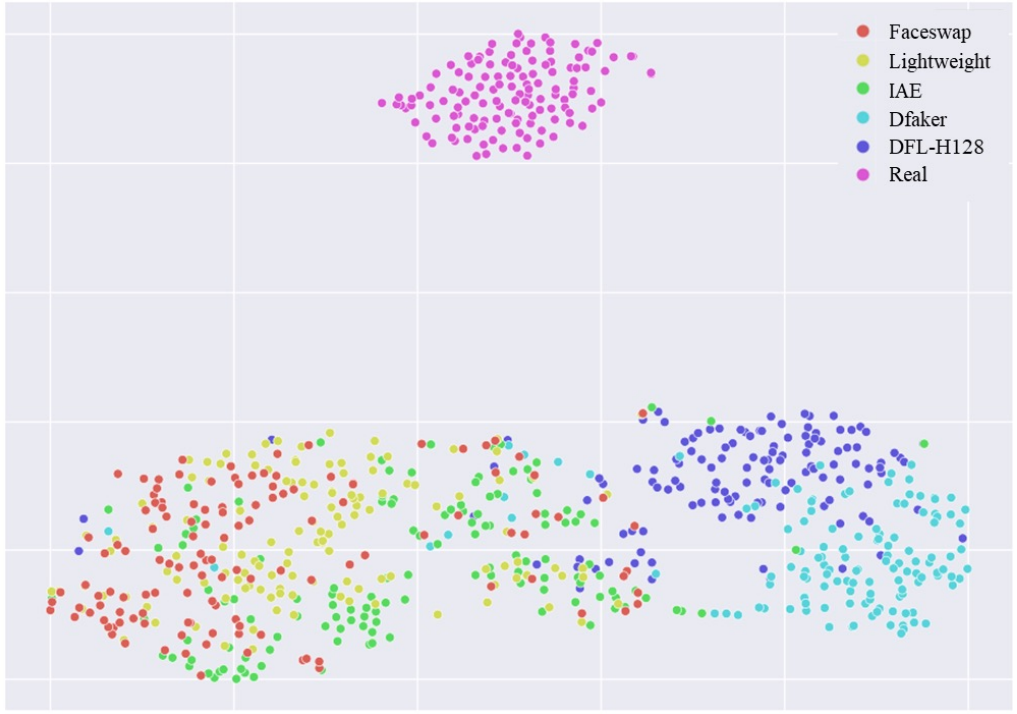


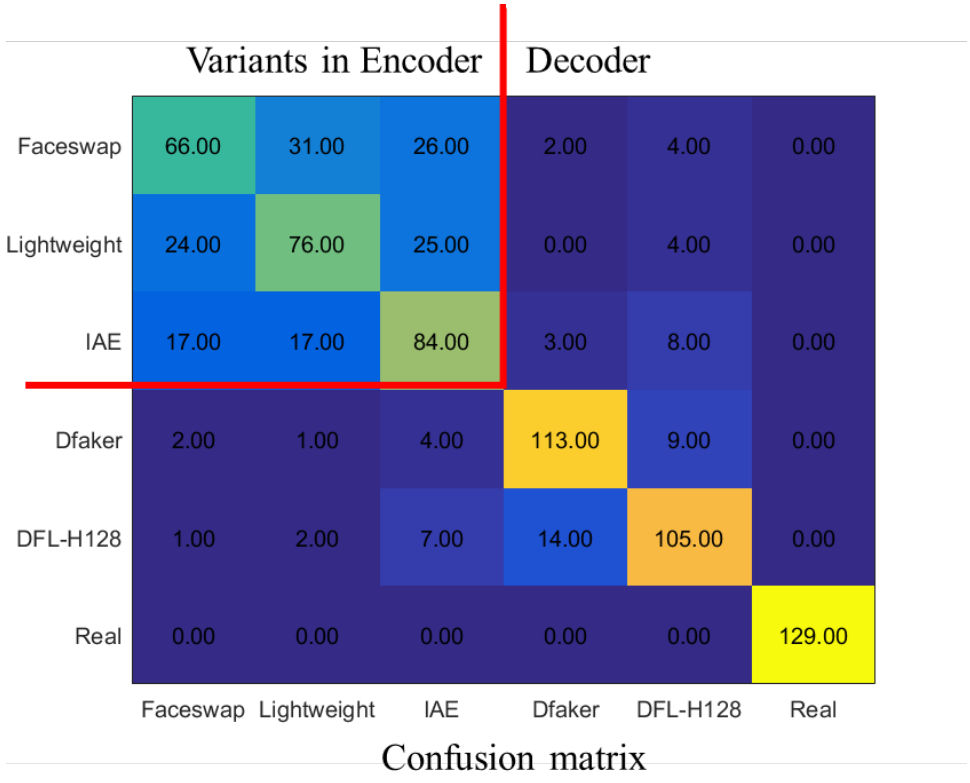
Confusion matrix

Visualization Results

- Model attribution & detection performance (6 classes)

Method	Overall	Faceswap	Lightweight	IAE	Dfaker	DFL-H128	Real
Ours: DMA-STA	74.03	51.16	58.91	65.12	87.60	81.40	100.00



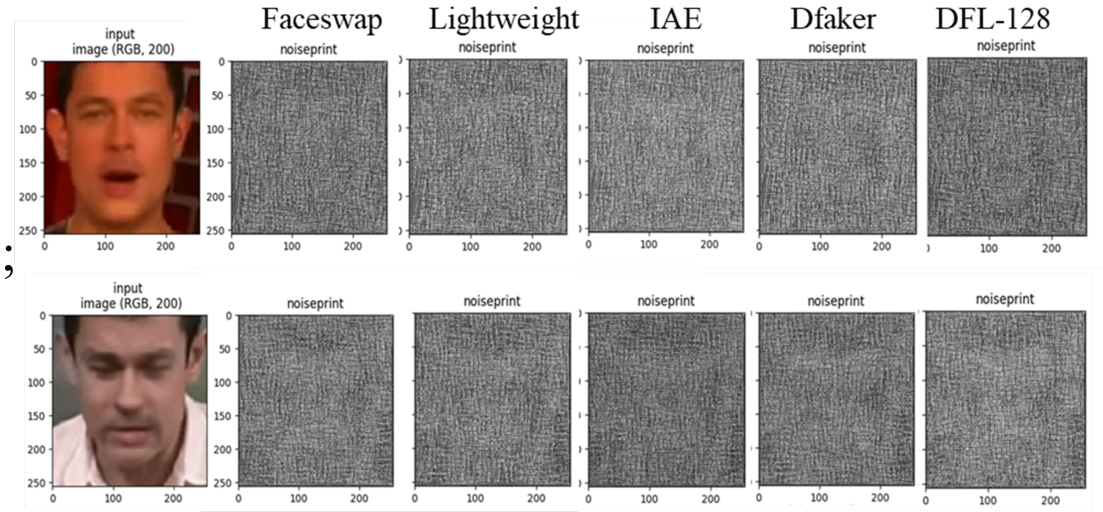


Conclusion

- The first work to explore model attribution of Deepfake videos;
- A new dataset with Deepfakes generated from different models, DFDM, showing visual differences among Deepfakes from different models;
- A new Deepfakes model attribution method based on spatial and temporal attention, DMA-STA, achieving over 70% accuracy in identifying the Deepfakes.

Future work

- Represent the artifacts/fingerprint of Deepfakes generation models;
- More Deepfake videos from more models (> 1 variation);
- Design open-set model attribution method.



Noise patterns from Siamese network based tool in [Cozzolino and Verdoliva, 2020]

Thank You!

Questions?

shanjia@buffalo.edu