ACM multimedia

ACM MULTIMEDIA
CONFERENCE 2020

2020.acmmm.org

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms

Hua Qi*, Qing Guo*, Felix Juefei-Xu, Xiaofei Xie, Lei Ma†, Wei Feng, Yang Liu, Jianjun Zhao

* Both authors contributed equally to this research.
† Lei Ma is the corresponding author (malei@ait.kyushu-u.ac.jp).

A Novel DeepFake Detection Technique

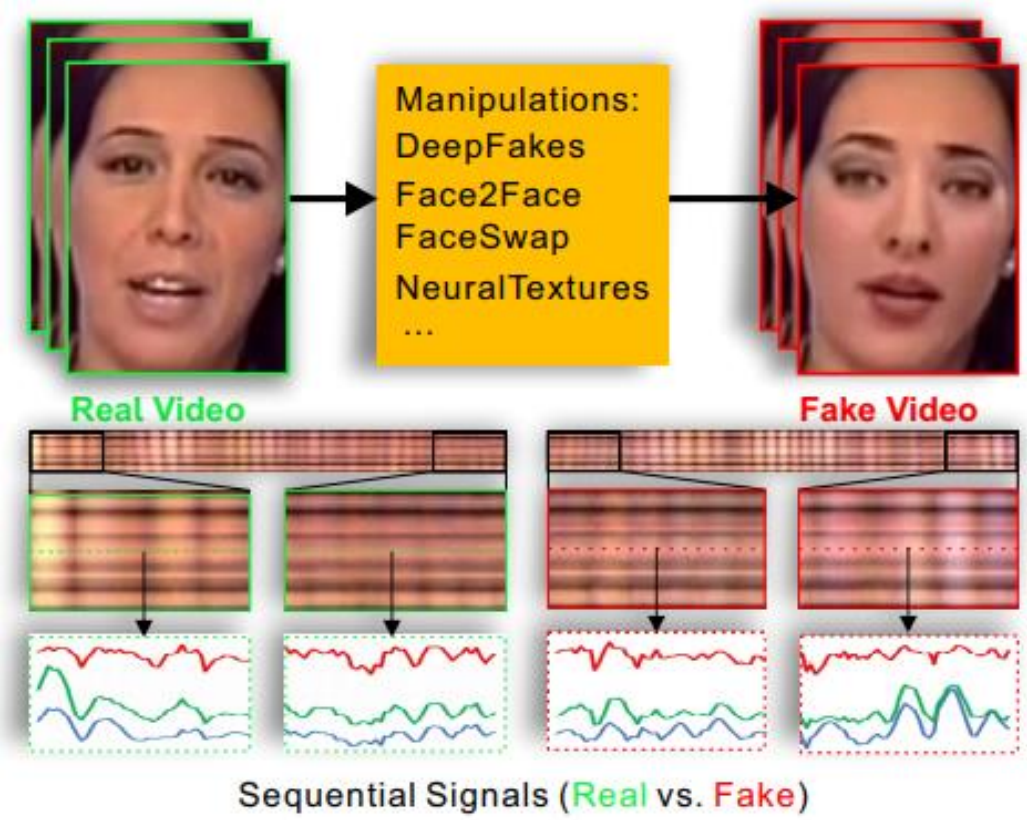
DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms

➢ Motivation

※ Detection methods based solely on the raw pixel-domain input become less effective as the DeepFake images and videos becoming more and more realistic.

※ Remote visual photoplethysmography (PPG) is made possible by monitoring the minuscule periodic changes of skin color due to blood pumping through the face from a video.

※ Current DeepFake methods do not explicitly preserve the pulse signal

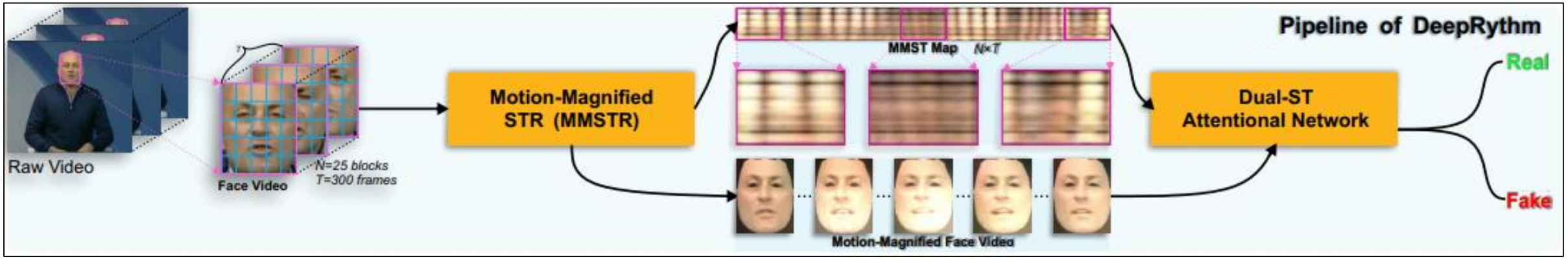


➢ Hypothesis

Heartbeat rhythms found in the real face videos will be disrupted or even broken in a DeepFake video.

DeepRhythm - Workflow

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



Video: $\mathcal{V} = \{I_i\}_{i=1}^{T}$ Real/Fake classification: $\phi(\cdot)$ Element-wise multiplication: $\odot$

➢ Generate MMST map

Use motion-magnified spatial-temporal representation (MMSTR) $X = mmstr(\mathcal{V}) \in \mathbb{R}^{T \times N \times C}$

➢ Classify MMST map

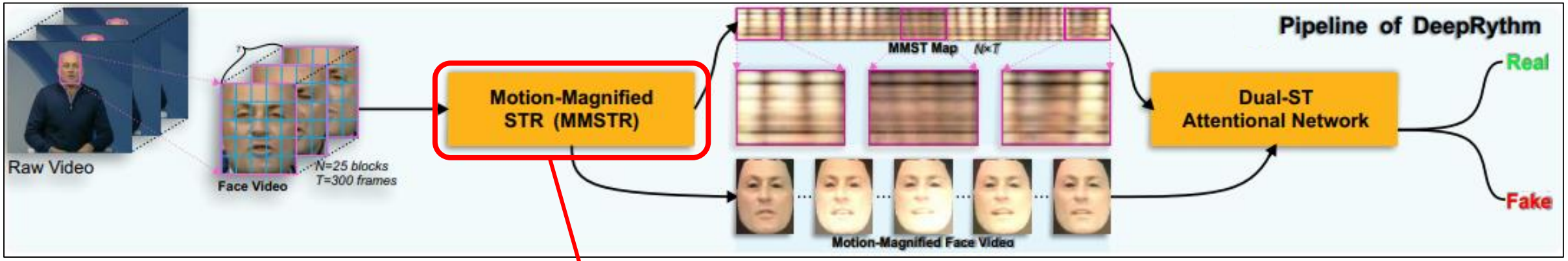
Use attention mechanism to alleviate interference

$$y = \phi(A \odot X), A \in \mathbb{R}^{T \times N}$$

$$A = t \cdot s^{\top}$$

DeepRhythm - Workflow

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



Video: $\mathcal{V} = \{I_i\}_{i=1}^{T}$ Real/Fake classification: $\phi(\cdot)$ Element-wise multiplication: $\odot$

➢ Generate MMST map

Use motion-magnified spatial-temporal representation (MMSTR) $X = mmstr(\mathcal{V}) \in \mathbb{R}^{T \times N \times C}$

➢ Classify MMST map

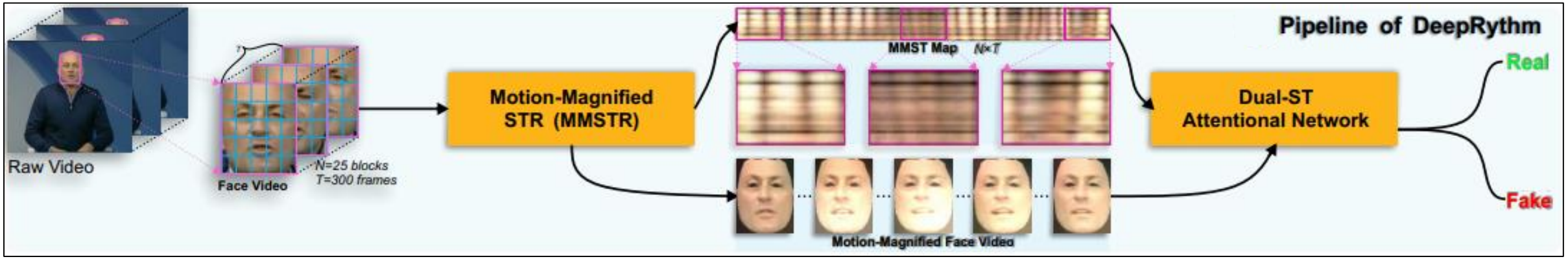
Use attention mechanism to alleviate interference

$$y = \phi(A \odot X), A \in \mathbb{R}^{T \times N}$$

$$A = t \cdot s^{\top}$$

DeepRhythm - Workflow

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



Video: $\mathcal{V} = \{I_i\}_{i=1}^{T}$ Real/Fake classification: $\phi(\cdot)$ Element-wise multiplication: $\odot$

➢ Generate MMST map

Use motion-magnified spatial-temporal representation (MMSTR) $X = mmstr(\mathcal{V}) \in \mathbb{R}^{T \times N \times C}$

➢ Classify MMST map

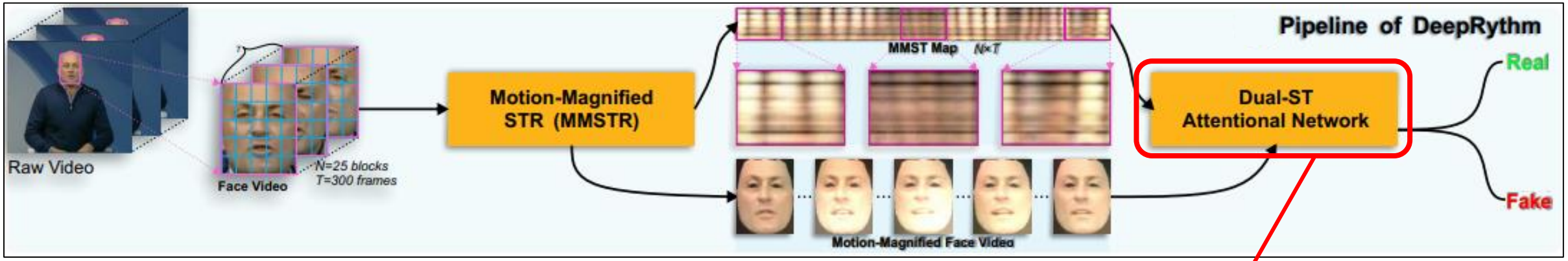
Use attention mechanism to alleviate interference

$$y = \phi(A \odot X), A \in \mathbb{R}^{T \times N}$$

$$A = t \cdot s^{\top}$$

DeepRhythm - Workflow

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



Video: $\mathcal{V} = \{I_i\}_{i=1}^{T}$ Real/Fake classification: $\phi(\cdot)$ Element-wise multiplication: $\odot$

➢ Generate MMST map

Use motion-magnified spatial-temporal representation (MMSTR) $X = mmstr(\mathcal{V}) \in \mathbb{R}^{T\times N\times C}$

➢ Classify MMST map

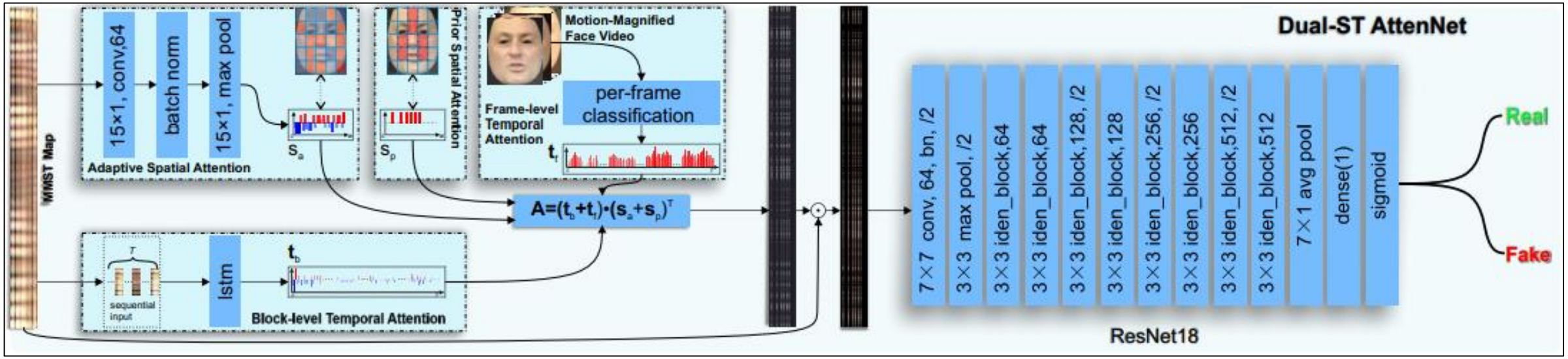
Use attention mechanism to alleviate interference

$$y = \phi(A \odot X), A \in \mathbb{R}^{T\times N}$$

$$A = t \cdot s^{\top}$$

DeepRhythm - Attention Mechanism

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



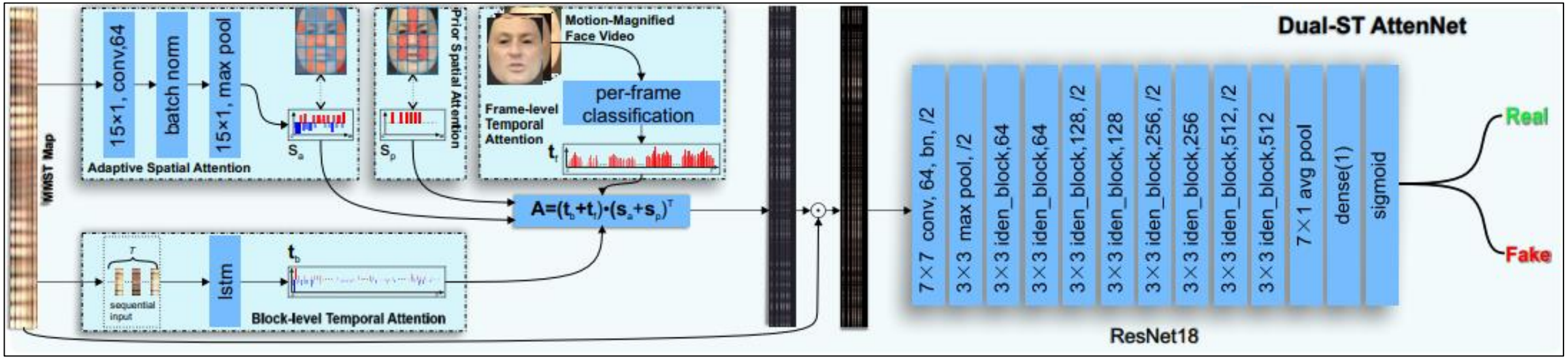
➢ Four Attentions

$A = t \cdot s^{\top}$

- Dual-Spatial Attention $s = s_a + s_p$
  - Adaptive Spatial Attention s_a
  - Prior Spatial Attention s_p
- Dual-Temporal Attention $t = t_b + t_f$
  - Block-level Temporal Attention t_b
  - Frame-level Temporal Attention t_f

DeepRhythm - Attention Mechanism

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



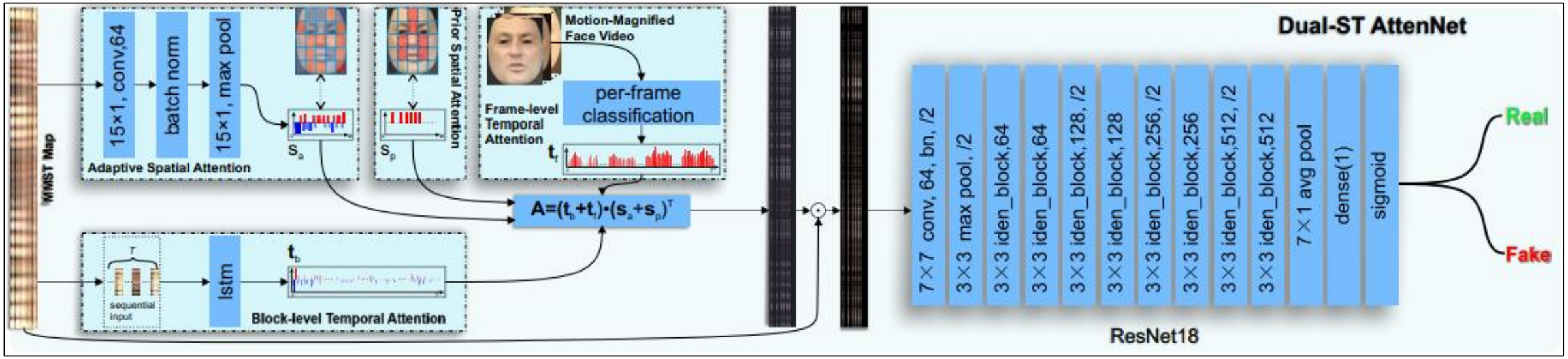
➢ Four Attentions

$$A = t \cdot s^{\top}$$

- Dual-Spatial Attention $s = s_a + s_p$
  - Adaptive Spatial Attention s_a
  - Prior Spatial Attention s_p
- Dual-Temporal Attention $t = t_b + t_f$
  - Block-level Temporal Attention t_b
  - Frame-level Temporal Attention t_f

DeepRhythm - Attention Mechanism

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



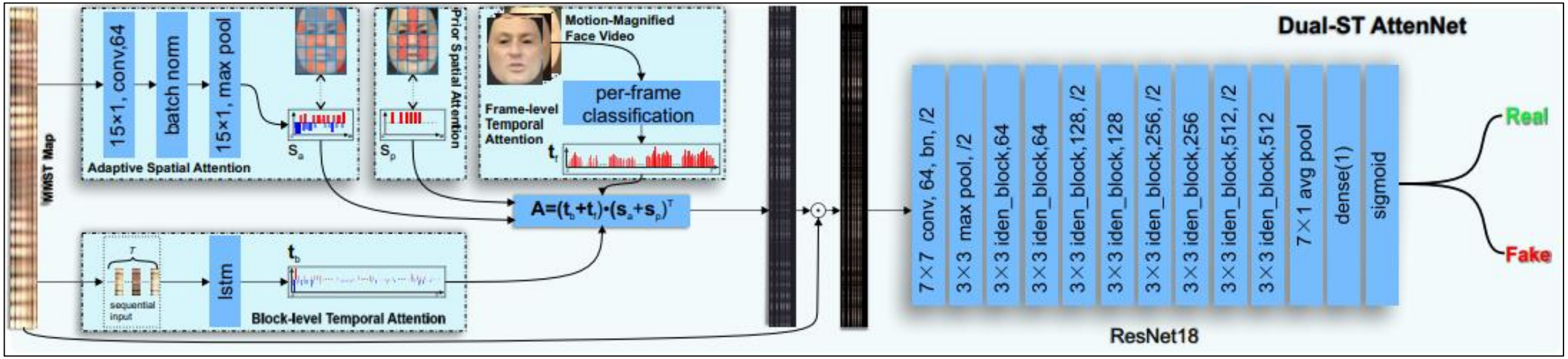
➢ Four Attentions

$A = t \cdot s^{\top}$

- Dual-Spatial Attention $s = s_a + s_p$
  - Adaptive Spatial Attention s_a
  - Prior Spatial Attention s_p
- Dual-Temporal Attention $t = t_b + t_f$
  - Block-level Temporal Attention t_b
  - Frame-level Temporal Attention t_f

DeepRhythm - Attention Mechanism

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



➢ Four Attentions

$A = t \cdot s^{\top}$

- Dual-Spatial Attention $s = s_a + s_p$
  - Adaptive Spatial Attention s_a
  - Prior Spatial Attention s_p
- Dual-Temporal Attention $t = t_b + t_f$
  - Block-level Temporal Attention t_b
  - Frame-level Temporal Attention t_f

Experimental Results - Baseline Comparison

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms

	train on sub-datasets				train on ALL dataset					
test on	DFD	DF	F2F	FS	DFD	DF	F2F	FS	ALL	DFDC
Bayer and Stamm	0.52	0.503	0.505	0.505	0.501	0.52	0.503	0.505	0.5	0.5
Inception ResNet V1	0.794	0.783	0.788	0.778	0.919	0.638	0.566	0.462	0.774	0.597
Xception	0.98	0.995	0.985	0.98	0.965	0.984	0.984	0.97	0.978	0.612
MesoNet	0.804	0.979	0.985	0.995	0.958	0.822	0.813	0.783	0.909	0.745
DeepRhythm (ours)	0.987	1.0	0.995	1.0	0.975	0.997	0.989	0.978	0.98	0.641

➢ Baseline

※ We choose the state-of-the-art DeepFake detection methods, i.e., Bayer's method[1], Inception ResNet V1[2], Xception[3] and MesoNet[4] as baselines, FaceForensics++ and DFDC as training and testing datasets.

※ The result demonstrates the **effectiveness** of our MMST representation and the **generalization capability** of our method across various DeepFake techniques.

[1] Belhassen Bayar and Matthew Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. 5–10.

[2] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learn-ing.AAAI Conference on Artificial Intelligence (02 2016).

[3] Francois Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. 1800–1807.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS).

Experimental Results – Ablation Study

	train on ALL sub-dataset				
test on	DFD	DF	F2F	FS	ALL
DR-st	0.522	0.497	0.497	0.492	0.512
DR-mmst	0.814	0.684	0.635	0.64	0.84
DR-mmst-A	0.849	0.77	0.736	0.716	0.847
DR-mmst-B	0.872	0.745	0.731	0.731	0.85
DR-mmst-AP	0.879	0.816	0.766	0.756	0.867
DR-mmst-BF	0.97	0.969	0.954	0.959	0.966
DR-mmst-APBF	0.965	0.959	0.954	0.965	0.964
DR-mmst-APBF-e2e	0.972	0.98	0.964	0.959	0.98

A - adaptive spatial attention
P - prior spatial attention
B - block-level temporal attention
F - frame-level temporal attention

➢ Ablation Study

※ The ST map has little discriminative power for DeepFake detection.

※ Adding adaptive spatial attention (DR-mmst-A) and block-level temporal attention (DR-mmst-B) do help improve the model's accuracy.

Experimental Results – Ablation Study

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms

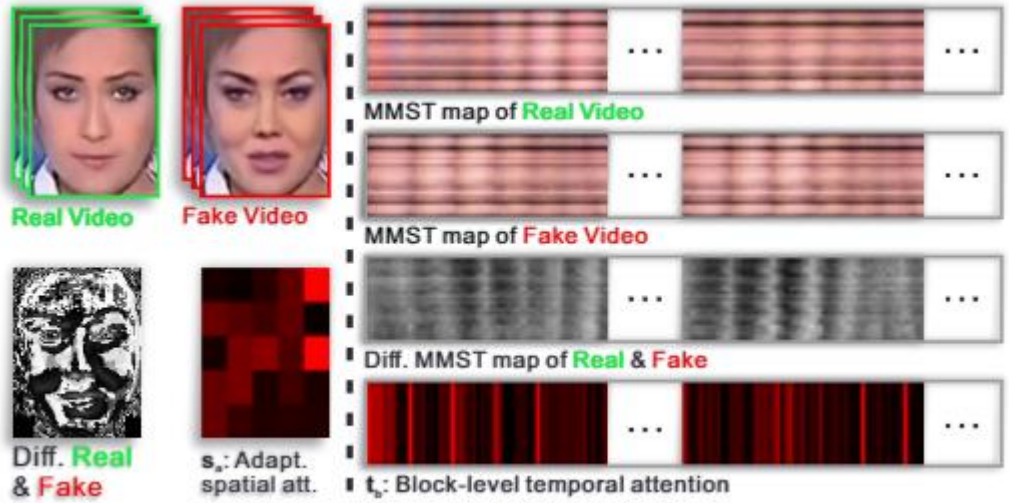
	train on ALL sub-dataset				
test on	DFD	DF	F2F	FS	ALL
DR-st	0.522	0.497	0.497	0.492	0.512
DR-mmst	0.814	0.684	0.635	0.64	0.84
DR-mmst-A	0.849	0.77	0.736	0.716	0.847
DR-mmst-B	0.872	0.745	0.731	0.731	0.85
DR-mmst-AP	0.879	0.816	0.766	0.756	0.867
DR-mmst-BF	0.97	0.969	0.954	0.959	0.966
DR-mmst-APBF	0.965	0.959	0.954	0.965	0.964
DR-mmst-APBF-e2e	0.972	0.98	0.964	0.959	0.98

A - adaptive spatial attention
P - prior spatial attention
B - block-level temporal attention
F - frame-level temporal attention

➢ Effectiveness of single attention

※ The main changes caused by the fake is around the nose, which is identical to the estimated adaptive spatial attention.

※ The estimated temporal attention has high values at the peaks of the difference MMST map.



Experimental Results – Ablation Study

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms

	train on ALL sub-dataset				
test on	DFD	DF	F2F	FS	ALL
DR-st	0.522	0.497	0.497	0.492	0.512
DR-mmst	0.814	0.684	0.635	0.64	0.84
DR-mmst-A	0.849	0.77	0.736	0.716	0.847
DR-mmst-B	0.872	0.745	0.731	0.731	0.85
DR-mmst-AP	0.879	0.816	0.766	0.756	0.867
DR-mmst-BF	0.97	0.969	0.954	0.959	0.966
DR-mmst-APBF	0.965	0.959	0.954	0.965	0.964
DR-mmst-APBF-e2e	0.972	0.98	0.964	0.959	0.98

A - adaptive spatial attention
P - prior spatial attention
B - block-level temporal attention
F - frame-level temporal attention

➢ Ablation Study

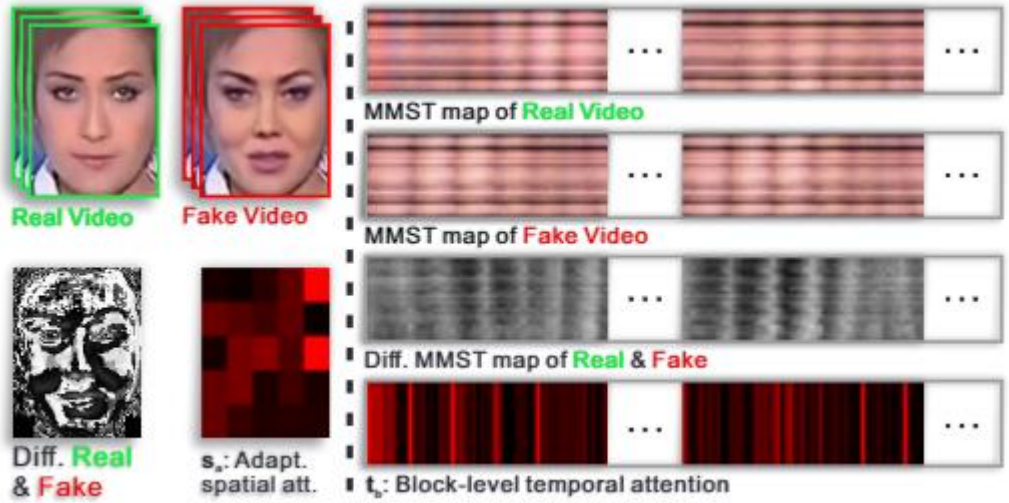
※ DR-mmst-APBF-e2e achieves the highest accuracy on all testing datasets.

※ The result indicates that training four attention separately might not mine the potential power of the four attention effectively, and training them together helps get the maximum effect.

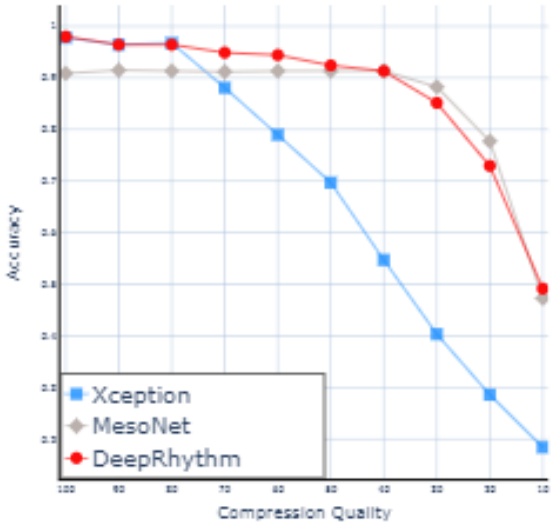
➢ Effectiveness of single attention

※ The main changes caused by the fake is around the nose, which is identical to the estimated adaptive spatial attention.

※ The estimated temporal attention has high values at the peaks of the difference MMST map.

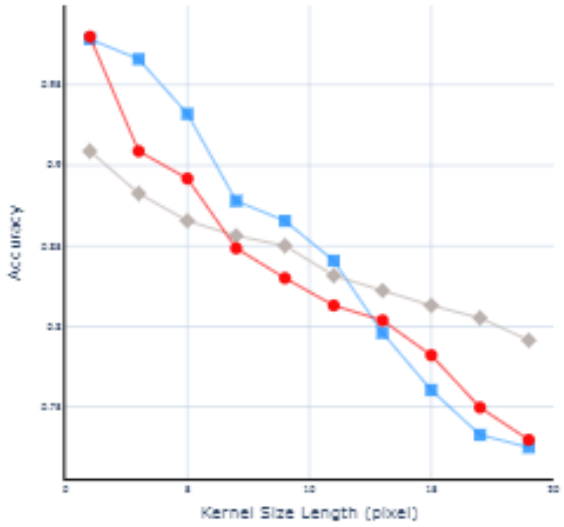


Experimental Results - Robustness



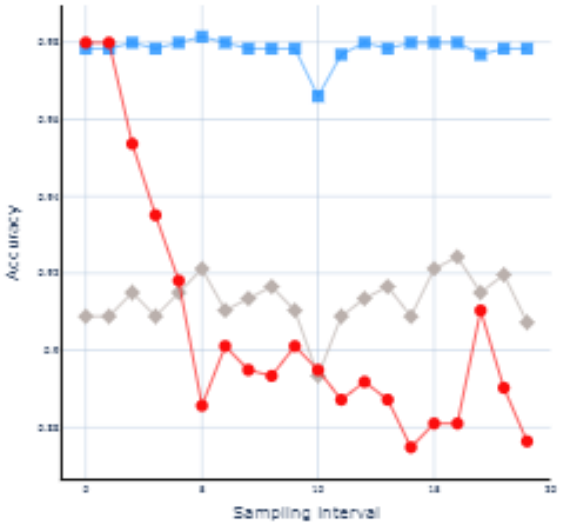
Xception	MesoNet	DeepRhythm
0.67±0.283	0.852±0.133	0.871±0.145

(a) JPEG



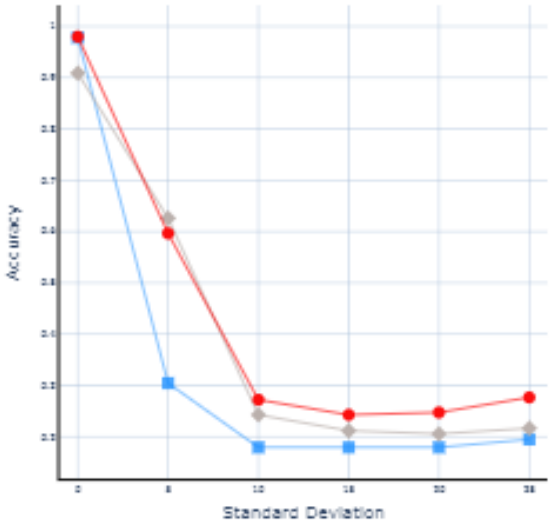
Xception	MesoNet	DeepRhythm
0.847±0.088	0.843±0.035	0.834±0.072

(b) Blur



Xception	MesoNet	DeepRhythm
0.847±0.088	0.843±0.035	0.834±0.072

(c) Sampling



Xception	MesoNet	DeepRhythm
0.847±0.088	0.843±0.035	0.834±0.072

(d) Gaussian Noise

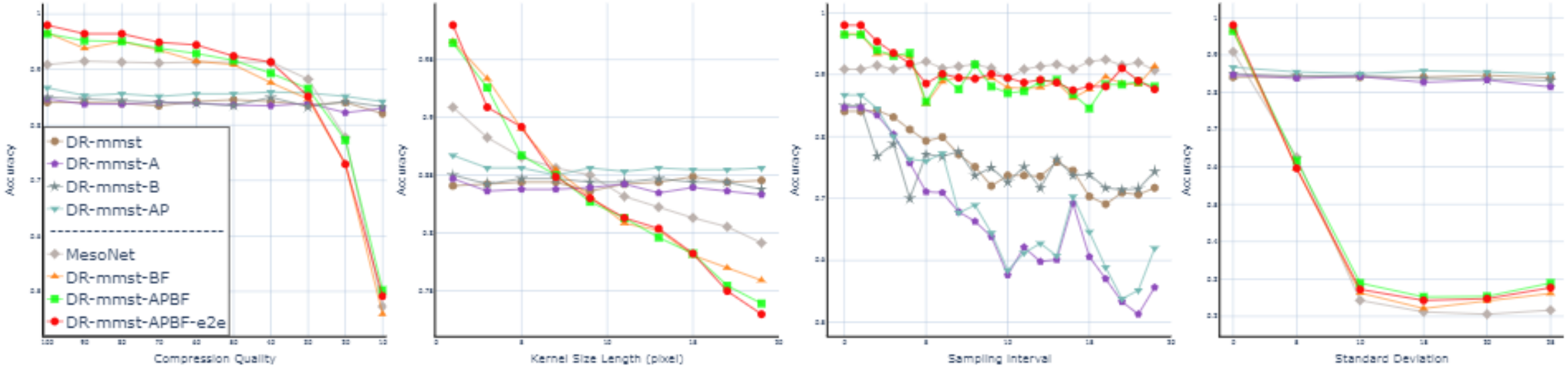
➢ Analysis

※ Our method exhibits strong robustness on JPEG compression and Gaussian noise, but do not perform well on temporal sampling when compared with Xception and MesoNet.

※ We could mitigate this issue with the video frame interpolation techniques in the future work.

Experimental Results - Robustness

DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms



(a) JPEG (b) Blur (c) Sampling (d) Gaussian Noise

➢ Analysis

※ The MMSTR is calculated by average pooling pixel values in ROI blocks, thus is insensitive to local pixel variation caused by JPEG compression, Gaussian blur, and Gaussian noise.

※ The frame-level temporal attention, generated by MesoNet, helps our methods be robust to temporal sampling and achieve the best performance but is sensitive to local pixel variation.

※ Combining these two modules shows comprehensive robustness across all degradations.

Conclusion

➢ Main Contributions

※ We propose DeepRhythm, the very first method for effective detection of DeepFake with the heartbeat rhythms.

※ To characterize the sequential signals of face videos, we propose the motion-magnified spatial-temporal representation (MMSTR) that provides powerful discriminative features for high accurate DeepFake detection.

※ To fully utilize the MMSTR, we propose dual-spatial-temporal attention network to adapt to dynamically changing faces and various fake types. Experimental results on FaceForensics++ and DeepFake Detection Challenge-preview dataset demonstrate that our method not only outperforms state-of-the-art methods but is robust to various degradations.

➢ Future Works

※ Studying the combined effort of DeepRhythm with other DeepFake detectors.

※ Investigating of how DeepRhythm can be applied further to other domains.

※ Using tracking methods to mine more discriminative spatial-temporal features.

ACM multimedia

ACM MULTIMEDIA
CONFERENCE 2020

2020.acmmm.org

Thank You!

Hua Qi*, Qing Guo*, Felix Juefei-Xu, Xiaofei Xie, Lei Ma†, Wei Feng, Yang Liu, Jianjun Zhao

* Both authors contributed equally to this research.
† Lei Ma is the corresponding author (malei@ait.kyushu-u.ac.jp).