NEURAL INFORMATION PROCESSING SYSTEMS

Through the Lens: Benchmarking Deepfake Detectors Against Moiré-Induced Distortions

Razaib Tariq*, Minji Heo*, Simon S.Woo‡ and Shahroz Tariq‡

*** Equal Contribution**
‡ Corresponding Author

SUNG KYUN KWAN UNIVERSITY

DASH-LAB Data-driven AI Security HCI SUNGKYUNKWAN UNIVERSITY

CSIRO DATA 61

DeepMoiréFake
Dataset

DeepMoiréFake
GitHub

Background

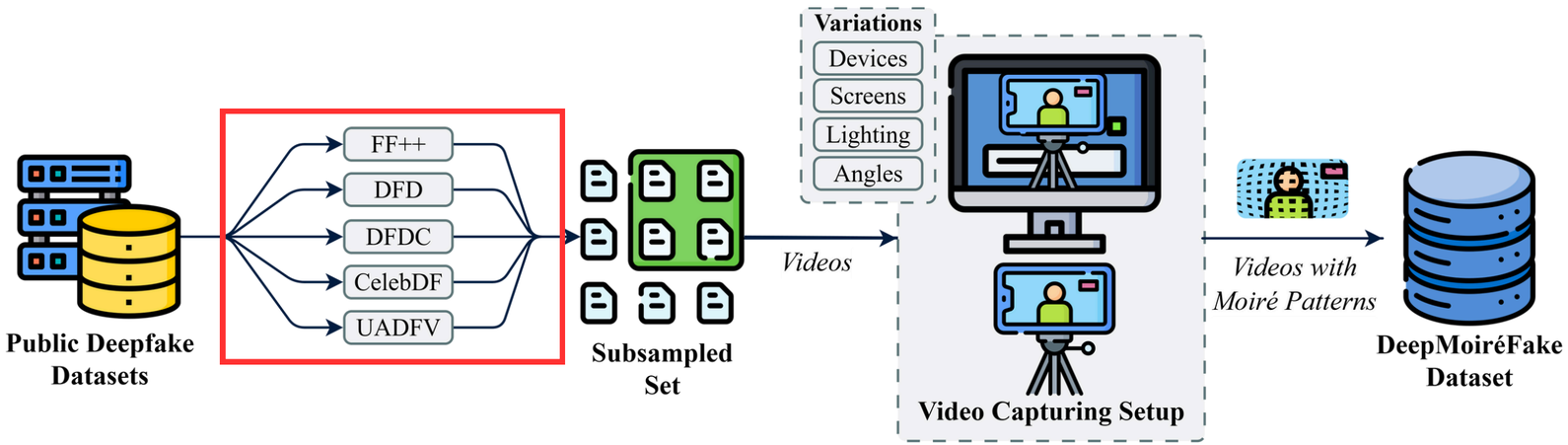
- Various Artifacts affect deepfakes in Real-World settings.
- Our Work focuses on Moiré Patterns.
- The interference is caused by the smartphone camera and the screen.

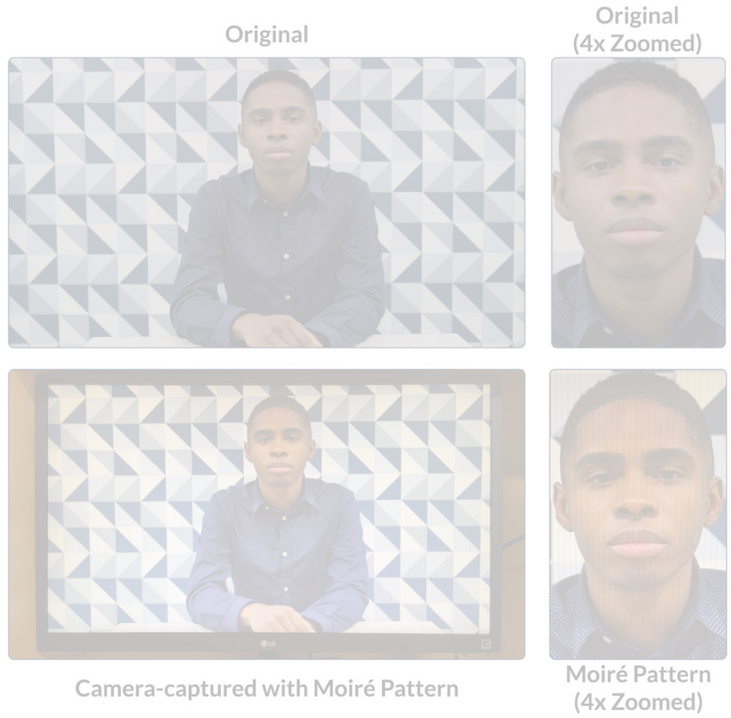
Background

- Various Artifacts affect deepfakes in Real-World settings.
- Our Work focuses on Moiré Patterns.
- The interference is caused by the smartphone camera and the screen.

DeepMoiréFake Generation Pipeline

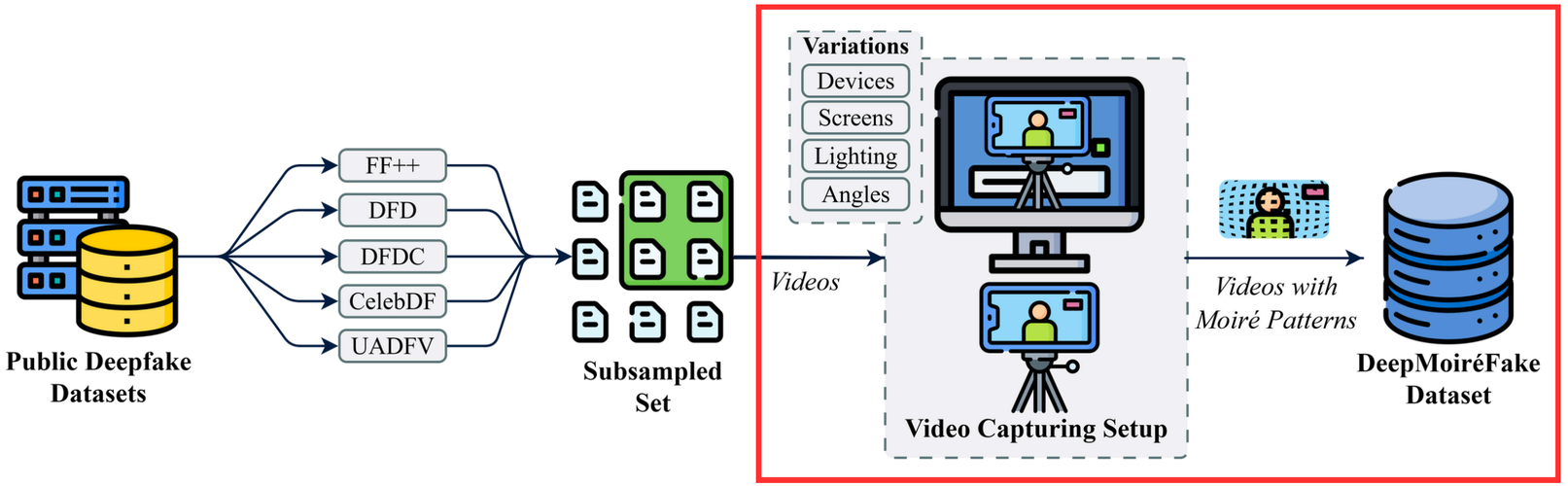
- We focus on five datasets that cover various ethnicities and genders.
- We captured the moiré pattern under 2 devices, 4 screens, 2 lightning, and 4 different camera angles.
- A 4x zoomed image of both the original and the Moiré.

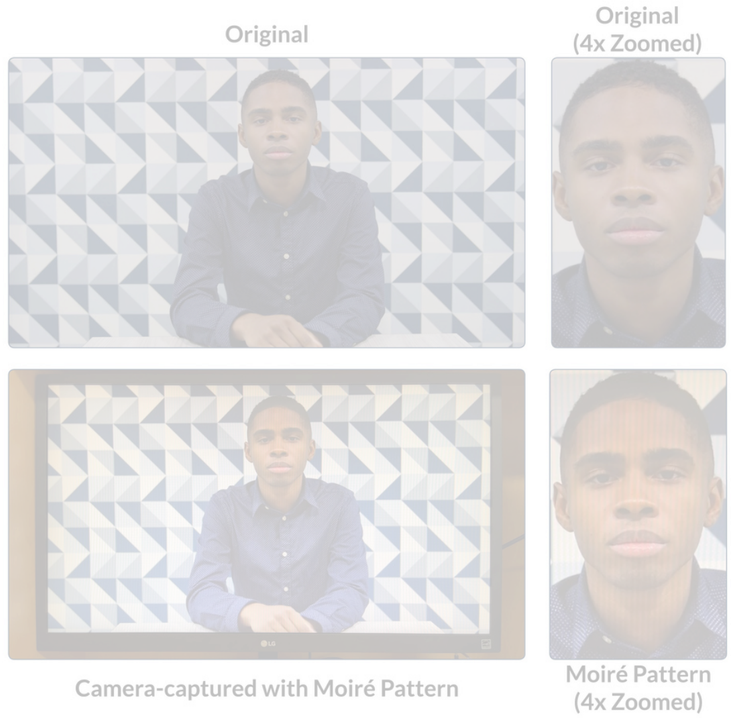




DeepMoiréFake Generation Pipeline

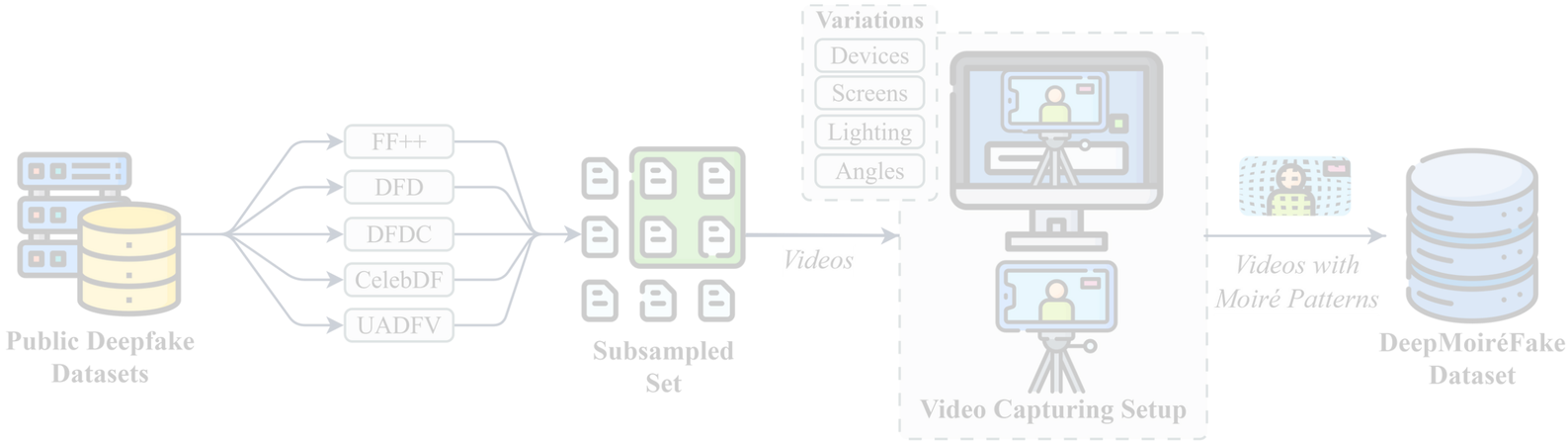
- We focus on five datasets that cover various ethnicities and genders.
- We captured the moiré pattern under 2 devices, 4 screens, 2 lightning, and 4 different camera angles.
- A 4x zoomed image of both the original and the Moiré.

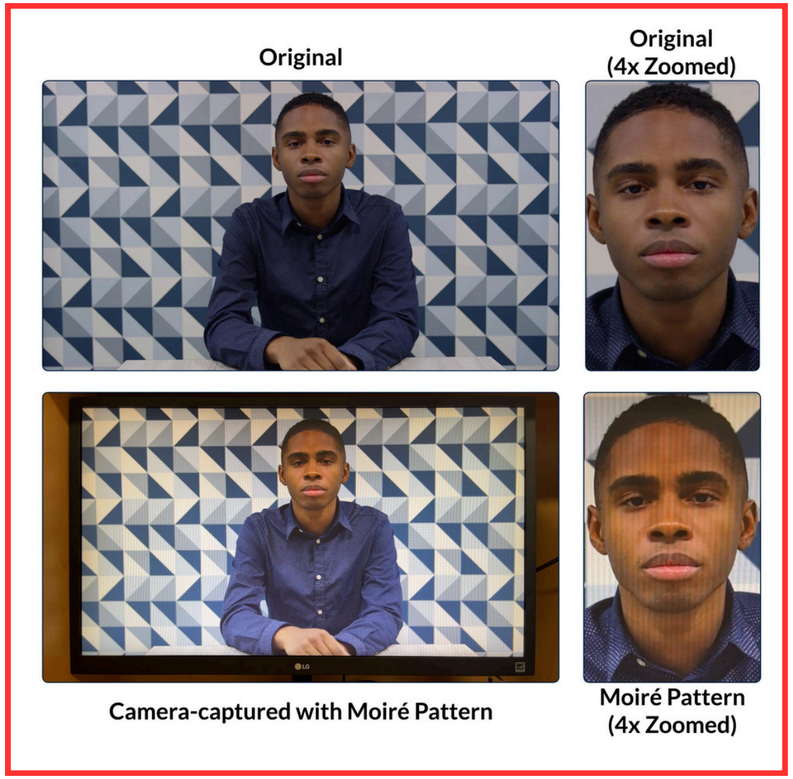




DeepMoiréFake Generation Pipeline

- We focus on five datasets that cover various ethnicities and genders.
- We captured the moiré pattern under 2 devices, 4 screens, 2 lightning, and 4 different camera angles.
- A 4x zoomed image of both the original and the Moiré.





Experiments

- We evaluate the performance of five datasets using various Spatial and Temporal methods.

	Detectors (*Type and Name*)	Original Performance	Videos captured from screens			
			LG	*BenQ*	*Lenovo*	*Samsung*
Video	LRNet BF	*61.7*	54.9	55.3	55.9	53.2
	LRNet RF	*62.2*	58.8	60.5	58.7	58.8
	FTCN	*90.2*	65.9	65.3	70.6	68.9
	LipForensics	*90.6*	80.3	80.8	**84.4**	79.8
	AltFreezing	***92.5***	**80.4**	**81.3**	83.7	**82.9**
Image	Rossler	*67.7*	56.2	54.5	59.4	56.9
	ADD	*69.7*	65.4	64.3	66.3	63.4
	Capsule	*71.3*	71.2	69.6	69.0	66.6
	ForgeryNet	*76.9*	61.5	61.8	66.5	63.6
	Rossler C40	*77.0*	67.7	66.9	67.3	67.8
	Rossler C23	*86.5*	68.6	67.4	74.5	70.9
	MAT	*87.0*	72.4	74.9	80.1	76.6
	CADDM	*87.1*	71.3	71.8	80.9	79.5
	SelfBlended	*88.8*	73.7	75.5	80.9	76.4
	CCViT	***95.0***	**81.9**	**83.7**	**86.4**	**86.0**
Avg. Performance loss (Moiré vs. Original)			-11.6	-11.4	-8.0	-10.2

Experiments under Video Deepfake Detectors

- All video detectors AUCs across all screens dropped **10.4%** on average.
- LRNet BF and LRNet RF are the most affected in every screen setting.

	Detectors (*Type and Name*)	Original Performance	Videos captured from screens			
			LG	*BenQ*	*Lenovo*	*Samsung*
Video	LRNet BF	*61.7*	54.9	55.3	55.9	53.2
Video	LRNet RF	*62.2*	58.8	60.5	58.7	58.8
Video	FTCN	*90.2*	65.9	65.3	70.6	68.9
Video	LipForensics	*90.6*	80.3	80.8	**84.4**	79.8
Video	AltFreezing	***92.5***	**80.4**	**81.3**	83.7	**82.9**
Image	Rossler	*67.7*	56.2	54.5	59.4	56.9
Image	ADD	*69.7*	65.4	64.3	66.3	63.4
Image	Capsule	*71.3*	71.2	69.6	69.0	66.6
Image	ForgeryNet	*76.9*	61.5	61.8	66.5	63.6
Image	Rossler C40	*77.0*	67.7	66.9	67.3	67.8
Image	Rossler C23	*86.5*	68.6	67.4	74.5	70.9
Image	MAT	*87.0*	72.4	74.9	80.1	76.6
Image	CADDM	*87.1*	71.3	71.8	80.9	79.5
Image	SelfBlended	*88.8*	73.7	75.5	80.9	76.4
Image	CCViT	***95.0***	**81.9**	**83.7**	**86.4**	**86.0**
	Avg. Performance loss (Moiré vs. Original)		-11.6	-11.4	-8.0	-10.2

Experiments under Image Deepfake Detectors

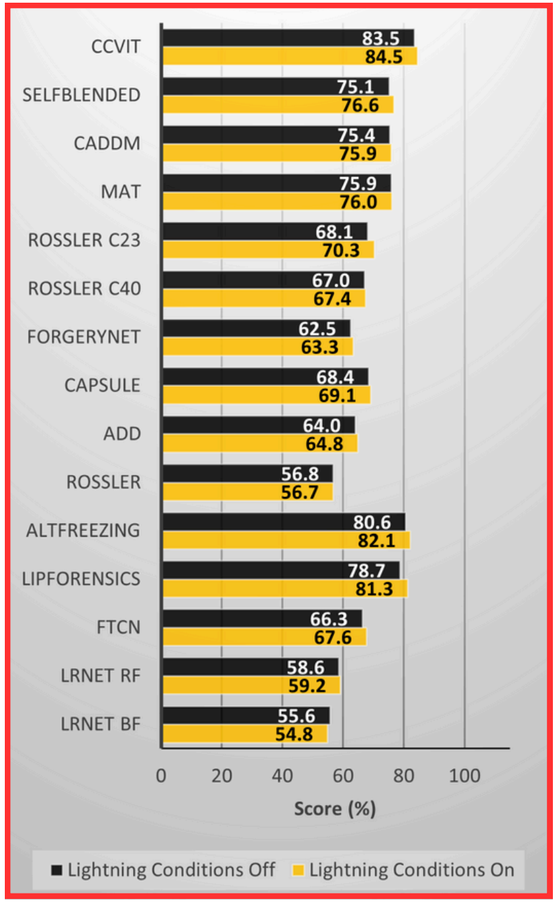
- All image detector AUCs across all screens dropped **10.2%** on average.
- The method most affected is Rossler, with an AUC loss of around **11%**.
- All detector AUCs drop under screen recapture, with average losses of **8-11.6%**, depending on the screens.

	Detectors (Type and Name)	Original Performance	Videos captured from screens			
			LG	*BenQ*	*Lenovo*	*Samsung*
Video	LRNet BF	*61.7*	54.9	55.3	55.9	53.2
	LRNet RF	*62.2*	58.8	60.5	58.7	58.8
	FTCN	*90.2*	65.9	65.3	70.6	68.9
	LipForensics	*90.6*	80.3	80.8	**84.4**	79.8
	AltFreezing	***92.5***	**80.4**	**81.3**	83.7	**82.9**
Image	Rossler	*67.7*	56.2	54.5	59.4	56.9
	ADD	*69.7*	65.4	64.3	66.3	63.4
	Capsule	*71.3*	71.2	69.6	69.0	66.6
	ForgeryNet	*76.9*	61.5	61.8	66.5	63.6
	Rossler C40	*77.0*	67.7	66.9	67.3	67.8
	Rossler C23	*86.5*	68.6	67.4	74.5	70.9
	MAT	*87.0*	72.4	74.9	80.1	76.6
	CADDM	*87.1*	71.3	71.8	80.9	79.5
	SelfBlended	*88.8*	73.7	75.5	80.9	76.4
	CCViT	***95.0***	**81.9**	**83.7**	**86.4**	**86.0**
	Avg. Performance loss (Moiré vs. Original)		-11.6	-11.4	-8.0	-10.2

Experiments under Lighting Conditions

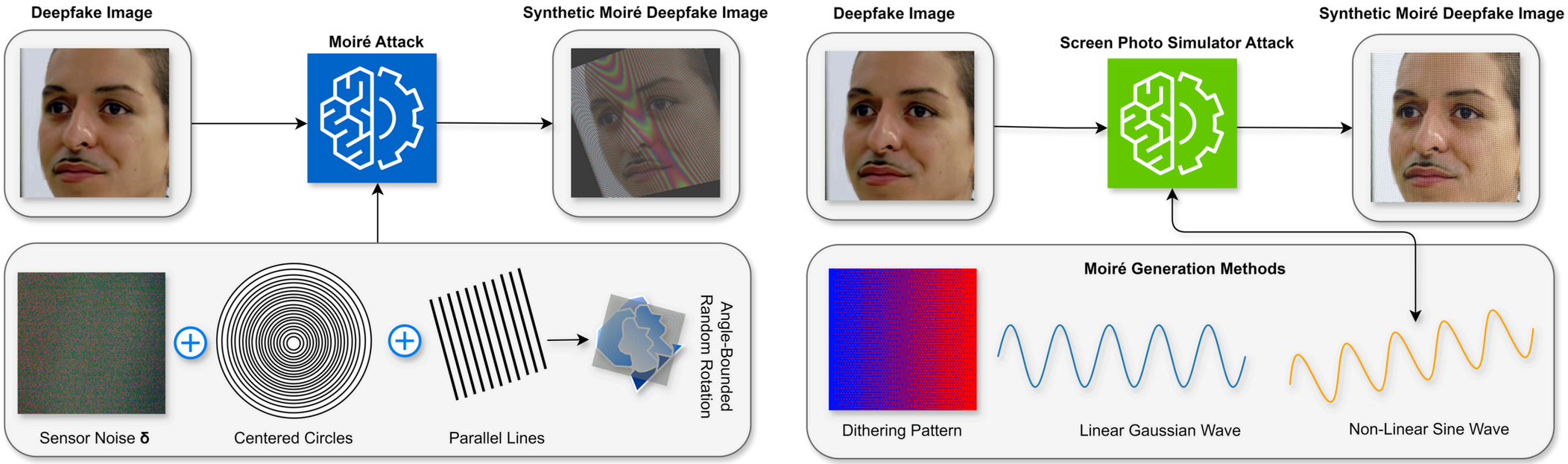
- There is an average **10.3%** AUC loss with light on and **11.2%** with lights off on average across all screens.

	Detectors (*Type and Name*)	Original Performance	Videos captured from screens			
			LG	*BenQ*	*Lenovo*	*Samsung*
Video	LRNet BF	*61.7*	54.9	55.3	55.9	53.2
	LRNet RF	*62.2*	58.8	60.5	58.7	58.8
	FTCN	*90.2*	65.9	65.3	70.6	68.9
	LipForensics	*90.6*	80.3	80.8	**84.4**	79.8
	AltFreezing	***92.5***	**80.4**	**81.3**	83.7	**82.9**
Image	Rossler	*67.7*	56.2	54.5	59.4	56.9
	ADD	*69.7*	65.4	64.3	66.3	63.4
	Capsule	*71.3*	71.2	69.6	69.0	66.6
	ForgeryNet	*76.9*	61.5	61.8	66.5	63.6
	Rossler C40	*77.0*	67.7	66.9	67.3	67.8
	Rossler C23	*86.5*	68.6	67.4	74.5	70.9
	MAT	*87.0*	72.4	74.9	80.1	76.6
	CADDM	*87.1*	71.3	71.8	80.9	79.5
	SelfBlended	*88.8*	73.7	75.5	80.9	76.4
	CCViT	***95.0***	**81.9**	**83.7**	**86.4**	**86.0**
Avg. Performance loss (Moiré vs. Original)			-11.6	-11.4	-8.0	-10.2



Synthetic Moiré Generation pipeline

- We simulated a scenario where an attacker could use synthetic Moiré.
- We utilize two methods to generate synthetic Moiré patterns.
- The performance of the deepfake detectors went down from **78-76%** to **75%-55%**, respectively.

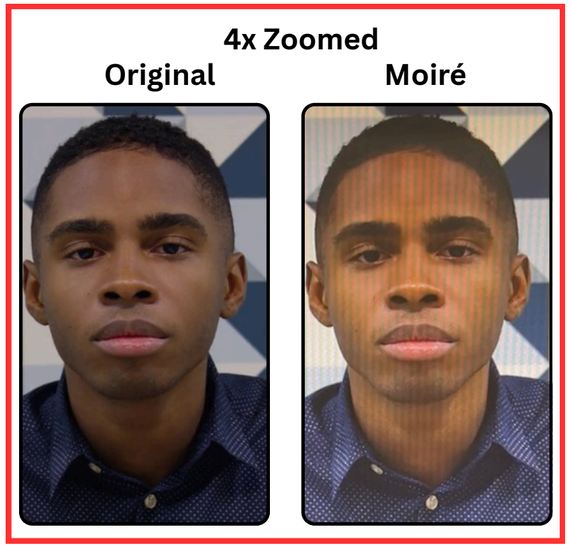


Performance after Moiré Pattern

- We choose image deepfake detectors for comparison with image-based demoiréing methods.

Detectors	AUC on OG	AUC on Moiré	Demoiréing Methods Performance					
			ESDNet (FHDMi)	*ESDNet (UHDM)*	*MBCNN*	*DMCNN*	*DDA*	*Average*
Rossler	*67.7*	*58.5*	58.1	53.7	55.9	57.1	54.4	55.8
ADD	*69.7*	*67.6*	68.5	65.5	66.1	65.7	64.8	66.1
Capsule	*71.3*	*70.4*	69.4	60.7	60.5	62.0	59.9	62.5
ForgeryNet	*76.9*	*64.3*	64.4	60.4	54.6	61.3	51.2	58.4
Rossler C40	*77.0*	*68.2*	69.1	66.6	64.2	66.5	63.3	65.9
Rossler C23	*86.5*	*72.8*	76.9	69.2	67.3	71.5	66.5	70.3
MAT	*87.0*	*75.2*	75.5	66.0	63.6	65.9	63.4	66.9
CADDM	*87.1*	*78.5*	79.5	73.4	72.7	75.0	72.3	74.6
SelfBlended	*88.8*	*78.4*	73.6	60.7	70.1	70.6	69.3	68.9
LipForensics	*90.6*	*83.3*	67.2	66.1	66.1	71.6	65.9	67.3
CCViT	***95.0***	***85.8***	**84.5**	**75.3**	**76.2**	**82.2**	**77.9**	**79.2**
Avg. AUC loss (DeMoiré vs. Moiré)			-1.5↓	-7.8↓	-7.8↓	4.9↓	8.6↓	6.1↓
Avg. AUC loss (DeMoiré vs. OG)			-10.1↓	-16.4↓	-16.4↓	13.5↓	17.2↓	14.7↓

Detector	Original	Moiré Video	VD-Moiré (Demoiré)	FPANet (Demoiré)
AltFreezing	**100.0**	84.4	74.7	**92.9**
FTCN	56.3	43.8	68.8	40.6
LipForensics	**100.0**	**87.5**	**90.6**	**90.6**



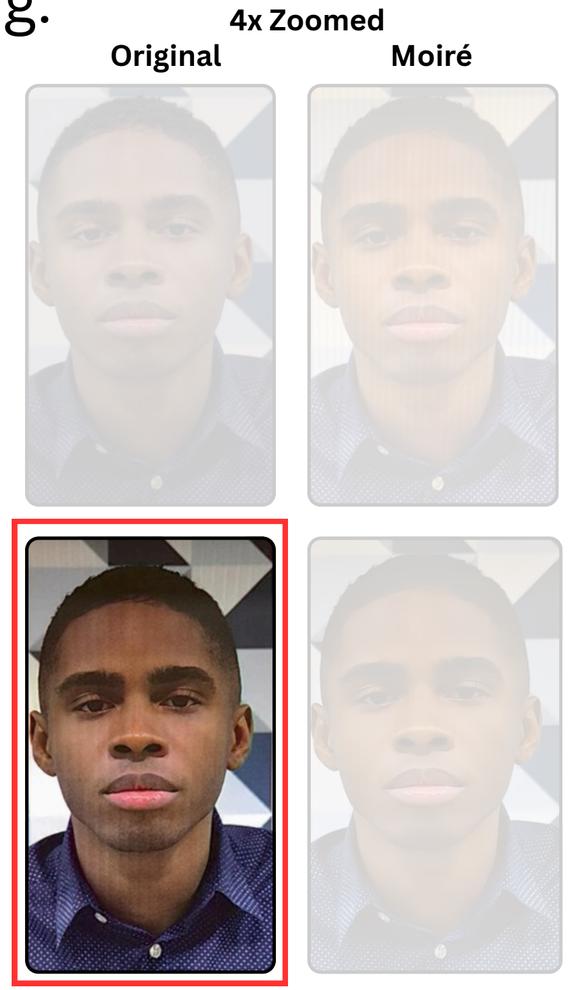
Demoiréd (4x Zoomed)

Performance after Image Demoiréing Methods

- Image demoiréing techniques yield inconsistent results; overall, most detectors show performance degradation after demoiréing.

Detectors	AUC on OG	AUC on Moiré	Demoiréing Methods Performance					
			ESDNet (FHDMi)	*ESDNet (UHDM)*	*MBCNN*	*DMCNN*	*DDA*	*Average*
Rossler	*67.7*	*58.5*	58.1	53.7	55.9	57.1	54.4	55.8
ADD	*69.7*	*67.6*	68.5	65.5	66.1	65.7	64.8	66.1
Capsule	*71.3*	*70.4*	69.4	60.7	60.5	62.0	59.9	62.5
ForgeryNet	*76.9*	*64.3*	64.4	60.4	54.6	61.3	51.2	58.4
Rossler C40	*77.0*	*68.2*	69.1	66.6	64.2	66.5	63.3	65.9
Rossler C23	*86.5*	*72.8*	76.9	69.2	67.3	71.5	66.5	70.3
MAT	*87.0*	*75.2*	75.5	66.0	63.6	65.9	63.4	66.9
CADDM	*87.1*	*78.5*	79.5	73.4	72.7	75.0	72.3	74.6
SelfBlended	*88.8*	*78.4*	73.6	60.7	70.1	70.6	69.3	68.9
LipForensics	*90.6*	*83.3*	67.2	66.1	66.1	71.6	65.9	67.3
CCViT	***95.0***	*85.8*	**84.5**	**75.3**	**76.2**	**82.2**	**77.9**	**79.2**
Avg. AUC loss (DeMoiré vs. Moiré)			-1.5↓	-7.8↓	-7.8↓	4.9↓	8.6↓	6.1↓
Avg. AUC loss (DeMoiré vs. OG)			-10.1↓	-16.4↓	-16.4↓	13.5↓	17.2↓	14.7↓

Detector	Original	Moiré Video	VD-Moiré (Demoiré)	FPANet (Demoiré)
AltFreezing	**100.0**	84.4	74.7	**92.9**
FTCN	56.3	43.8	68.8	40.6
LipForensics	**100.0**	**87.5**	**90.6**	**90.6**

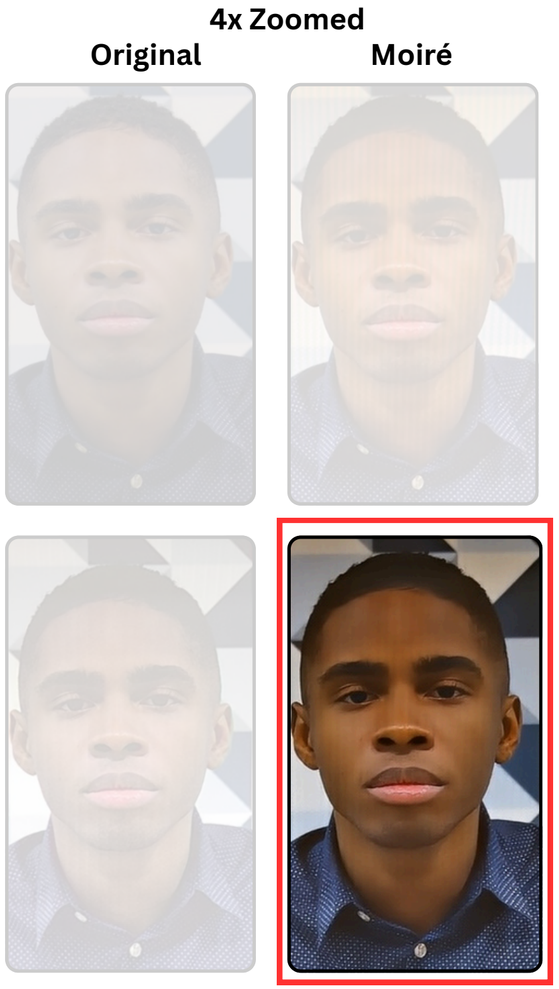


Performance after Video Demoiréing Methods

- Video demoiréing lowers the AUC for deepfake detectors like FTCN (FPANet technique) but performs better with other methods.

Detectors	AUC on OG	AUC on Moiré	Demoiréing Methods Performance					
			ESDNet (FHDMi)	*ESDNet (UHDM)*	*MBCNN*	*DMCNN*	*DDA*	*Average*
Rossler	*67.7*	*58.5*	58.1	53.7	55.9	57.1	54.4	55.8
ADD	*69.7*	*67.6*	68.5	65.5	66.1	65.7	64.8	66.1
Capsule	*71.3*	*70.4*	69.4	60.7	60.5	62.0	59.9	62.5
ForgeryNet	*76.9*	*64.3*	64.4	60.4	54.6	61.3	51.2	58.4
Rossler C40	*77.0*	*68.2*	69.1	66.6	64.2	66.5	63.3	65.9
Rossler C23	*86.5*	*72.8*	76.9	69.2	67.3	71.5	66.5	70.3
MAT	*87.0*	*75.2*	75.5	66.0	63.6	65.9	63.4	66.9
CADDM	*87.1*	*78.5*	79.5	73.4	72.7	75.0	72.3	74.6
SelfBlended	*88.8*	*78.4*	73.6	60.7	70.1	70.6	69.3	68.9
LipForensics	*90.6*	*83.3*	67.2	66.1	66.1	71.6	65.9	67.3
CCViT	***95.0***	***85.8***	**84.5**	**75.3**	**76.2**	**82.2**	**77.9**	**79.2**
Avg. AUC loss (DeMoiré vs. Moiré)			-1.5↓	-7.8↓	-7.8↓	4.9↓	8.6↓	6.1↓
Avg. AUC loss (DeMoiré vs. OG)			-10.1↓	-16.4↓	-16.4↓	13.5↓	17.2↓	14.7↓

Detector	Original	Moiré Video	VD-Moiré (Demoiré)	FPANet (Demoiré)
AltFreezing	**100.0**	84.4	74.7	**92.9**
FTCN	56.3	43.8	68.8	40.6
LipForensics	**100.0**	**87.5**	**90.6**	**90.6**

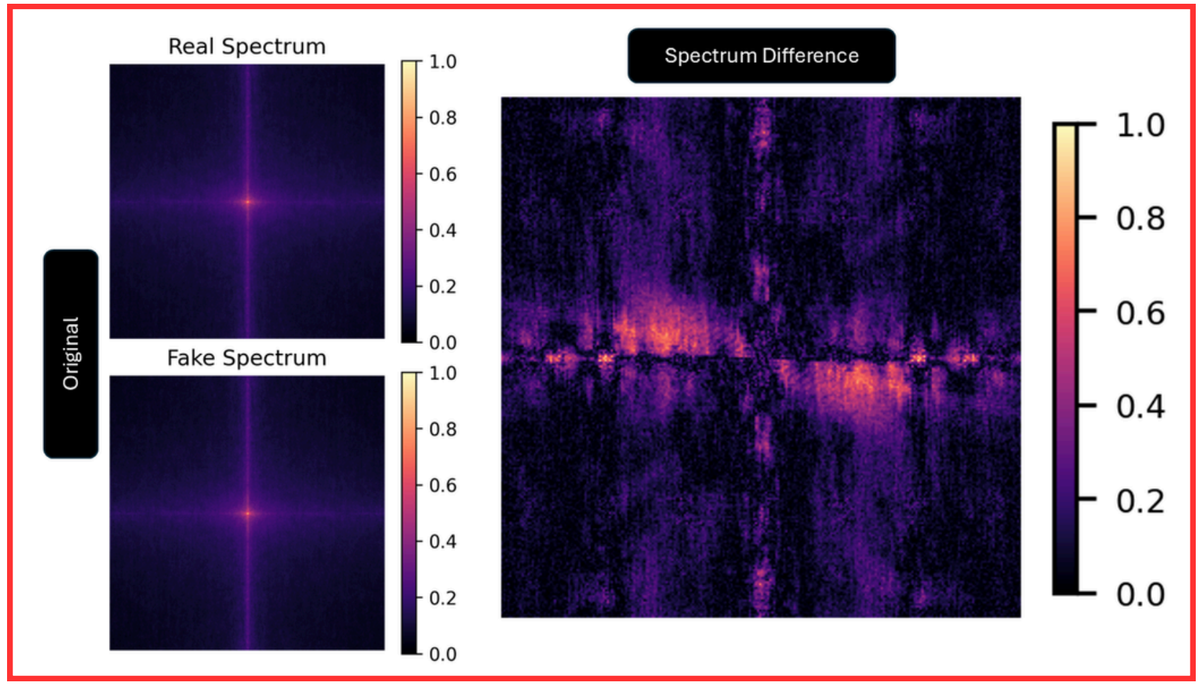


Moiré pattern under Frequency Spectrum

- Moire pattern distorts the original frequency pattern of a deepfake video.
- This degradation explains the drops in performance in all detectors.

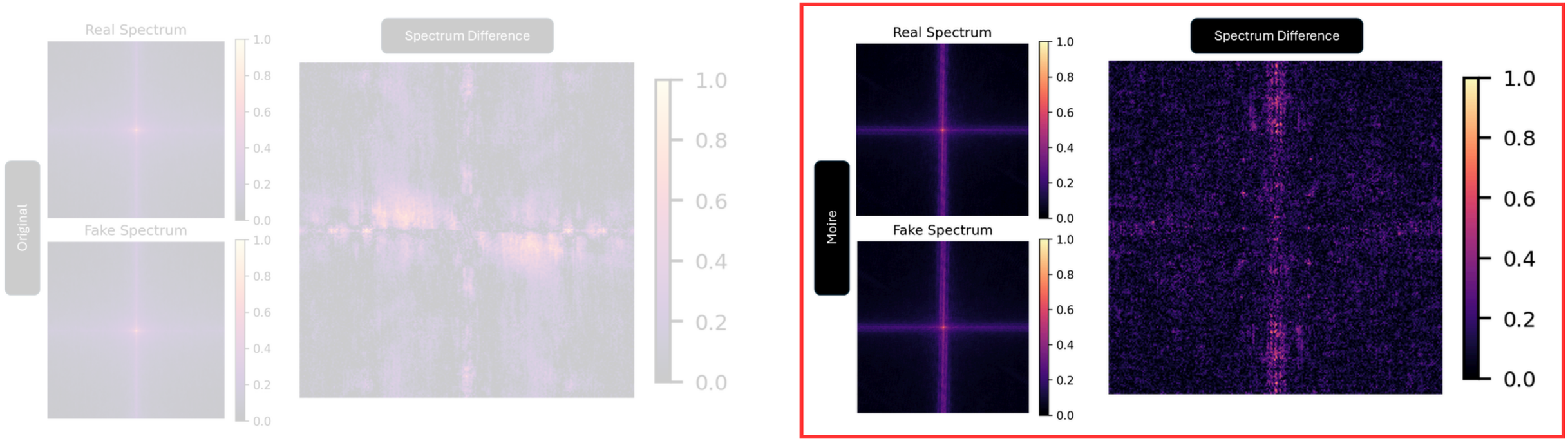
Moiré pattern under Frequency Spectrum

- The spectrum difference of a deepfake video exhibits a distinct characteristic.



Moiré pattern under Frequency Spectrum

- Moire pattern distorts the original frequency pattern of a deepfake video.
- This degradation explains the drops in performance in all detectors.

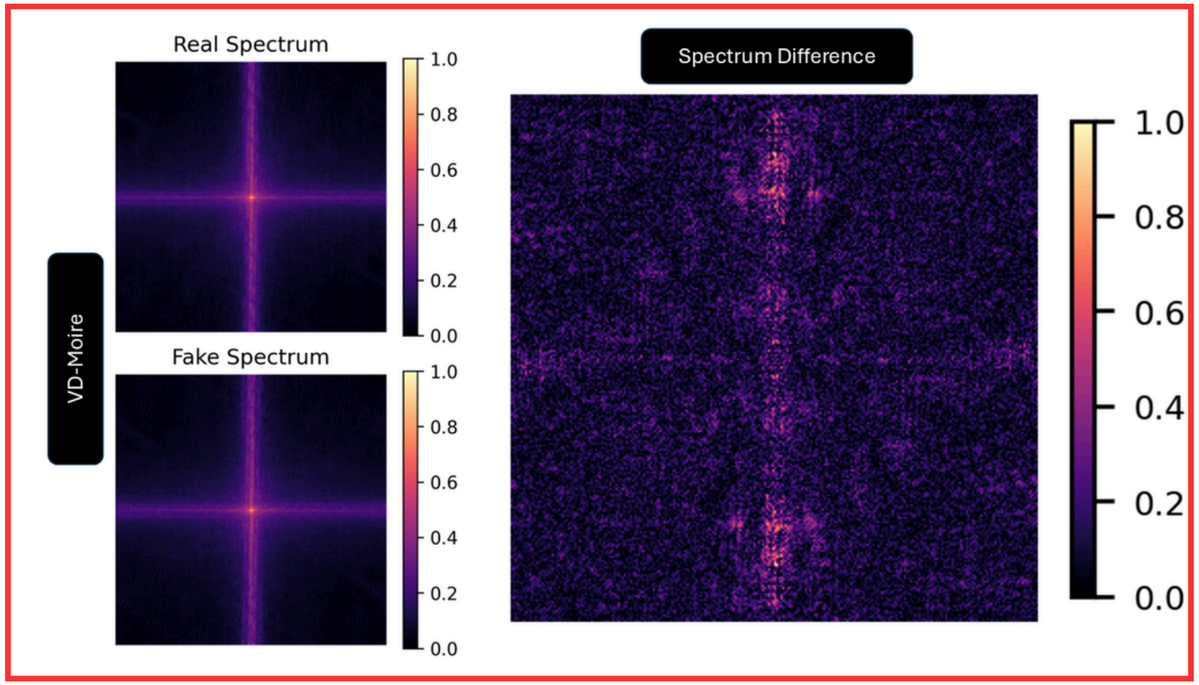


Moiré pattern under Frequency Spectrum—Demoiréing

- Using Demoiréing to recover the original frequency pattern is not enough.
- The original pattern is lost, leading to a significant loss in prediction.

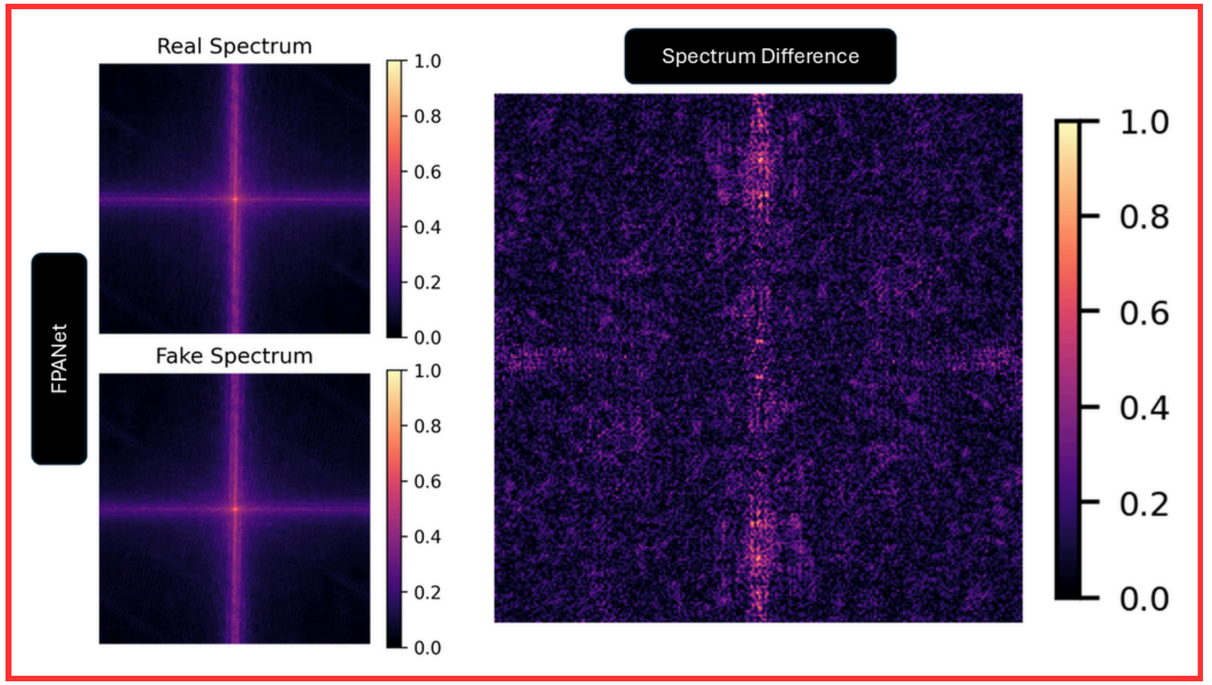
Moiré pattern under Frequency Spectrum—Demoiréing

- Using Demoiréing to recover the original frequency pattern is not enough.
- The original pattern is lost, leading to a significant loss in prediction.



Moiré pattern under Frequency Spectrum—Demoiréing

- Using Demoiréing to recover the original frequency pattern is not enough.
- The original pattern is lost, leading to a significant loss in prediction.



Performance after Synthetic and Compression Moiré Attacks

- Although some methods showed slight improvement, we consider them to be methods that can handle these distortions.
- MAT showed the most severe performance drop of **21.4%**.
- This is mainly due to the change of artifacts introduced at inference time.

Detectors	Without Attack	Moiré Attack		
		CMPA	*SMPA-MA*	*SMPA-SPS*
Rossler C23	**78.1**	**81.9**	83.1	75.4
MAT	76.8	68.8	55.4	61.8
CADDM	73.0	73.1	**86.8**	**80.7**

Performance after Synthetic and Compression Moiré Attacks

- Although some methods showed slight improvement, we consider them to be methods that can handle these distortions.
- MAT showed the most severe performance drop of **21.4%**.
- This is mainly due to the change of artifacts introduced at inference time.

Detectors	Without Attack	Moiré Attack		
		CMPA	*SMPA-MA*	*SMPA-SPS*
Rossler C23	**78.1**	**81.9**	83.1	75.4
MAT	76.8	68.8	55.4	61.8
CADDM	73.0	73.1	**86.8**	**80.7**

Performance after Fine-Tuning and Retraining

- Fine-tuning and retraining improved all models overall, with MAT and CADDM showing uniform gains across domains.
- Rossler exhibited small **(1–2%)** drops in a few subsets but remained largely improved.

Detectors	Fine-tune				Retrain			
	OG	*CMPA*	*SMPA-MA*	*SMPA-SPS*	*OG*	*CMPA*	*SMPA-MA*	*SMPA-SPS*
Rossler C23	77.0	80.6	94.4	81.1	87.9	84.7	**94.9**	79.5
MAT	**94.5**	**85.4**	70.3	**95.6**	**97.9**	**89.0**	71.3	**96.5**
CADDM	86.3	84.6	**94.4**	95.0	85.1	81.9	92.9	95.4

Performance after Fine-Tuning and Retraining

- Fine-tuning and retraining improved all models overall, with MAT and CADDM showing uniform gains across domains.
- Rossler exhibited small **(1–2%)** drops in a few subsets but remained largely improved.

Detectors	Fine-tune				Retrain			
	OG	*CMPA*	*SMPA-MA*	*SMPA-SPS*	*OG*	*CMPA*	*SMPA-MA*	*SMPA-SPS*
Rossler C23	77.0	80.6	94.4	81.1	87.9	84.7	**94.9**	79.5
MAT	**94.5**	**85.4**	70.3	**95.6**	**97.9**	**89.0**	71.3	**96.5**
CADDM	86.3	84.6	**94.4**	95.0	85.1	81.9	92.9	95.4

Conclusion

- We provide a Novel Real-World Moiré Benchmark dataset; **DeepMoiréFake (DMF)**.
- We highlight Authentic Moiré vs. Synthetic Moiré results and how they affect deepfake detectors.
- We show how demoiréing methods can be effective, but they may essentially remove deepfake features.
- We also perform retraining, which enhances model performance.
- We conclude by calling for Distortion-Aware training for future researchers by incorporating our DMF dataset in their pipeline.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!!!

SUNG KYUN KWAN UNIVERSITY

DASH-LAB Data-driven AI Security HCI SUNGKYUNKWAN UNIVERSITY

CSIRO DATA 61

DeepMoiréFake
Dataset

DeepMoiréFake
GitHub