Is To See Still To Believe in Deepfake Era?

Jun-Cheng Chen

pullpull@citi.sinica.edu.tw

OpenMFC Workshop

Artificial Intelligence and Image Understanding Lab (AIIU)
Research Center of Information Technology Innovation, Academia Sinica

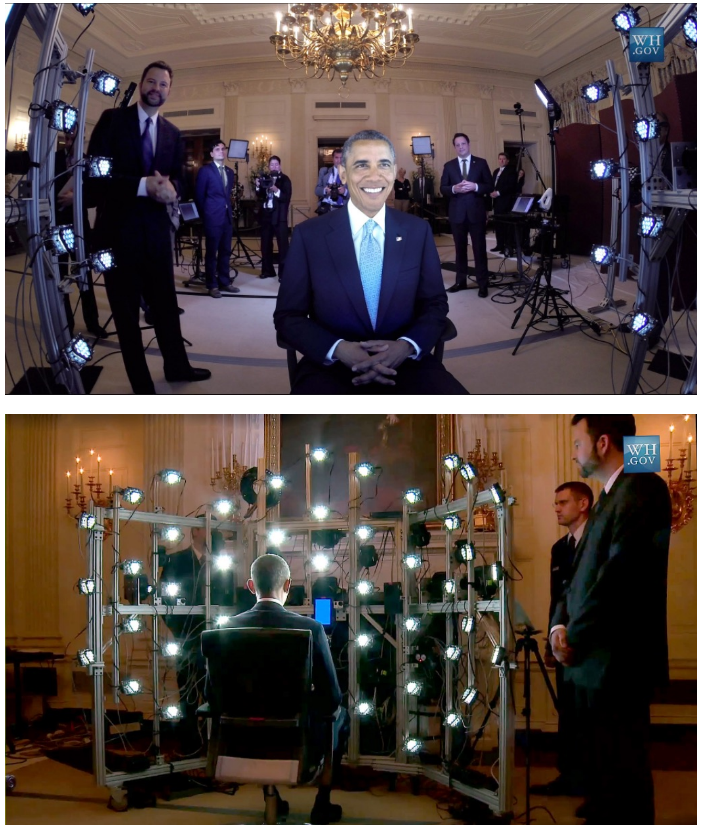
2021/12/09

Motivation

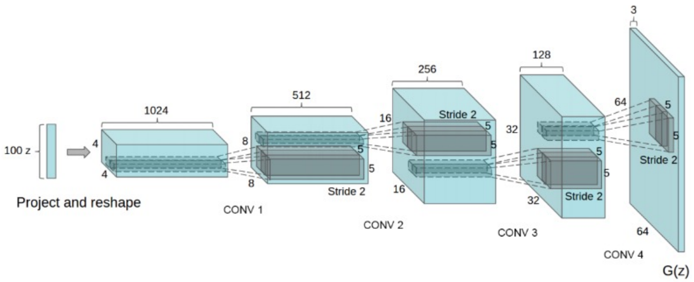
- Malicious Face Forgery Applications
  - ➢Pornography
  - ➢Politics

image source: https://technews.tw/2020/10/25/deepfake-deepnude/

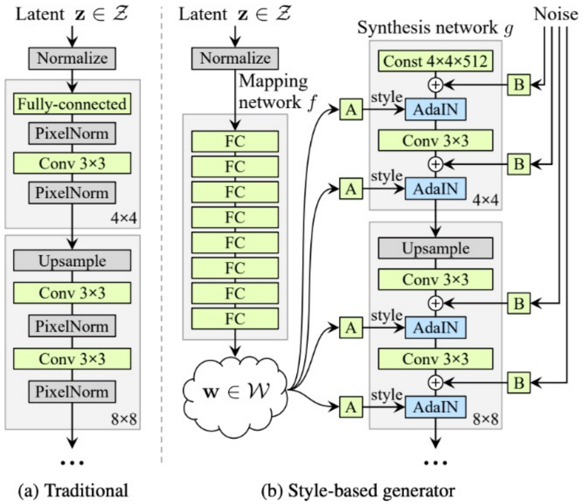
The Evolution of Content Editing(1/4)



LightStage
[USC ICT 2015]

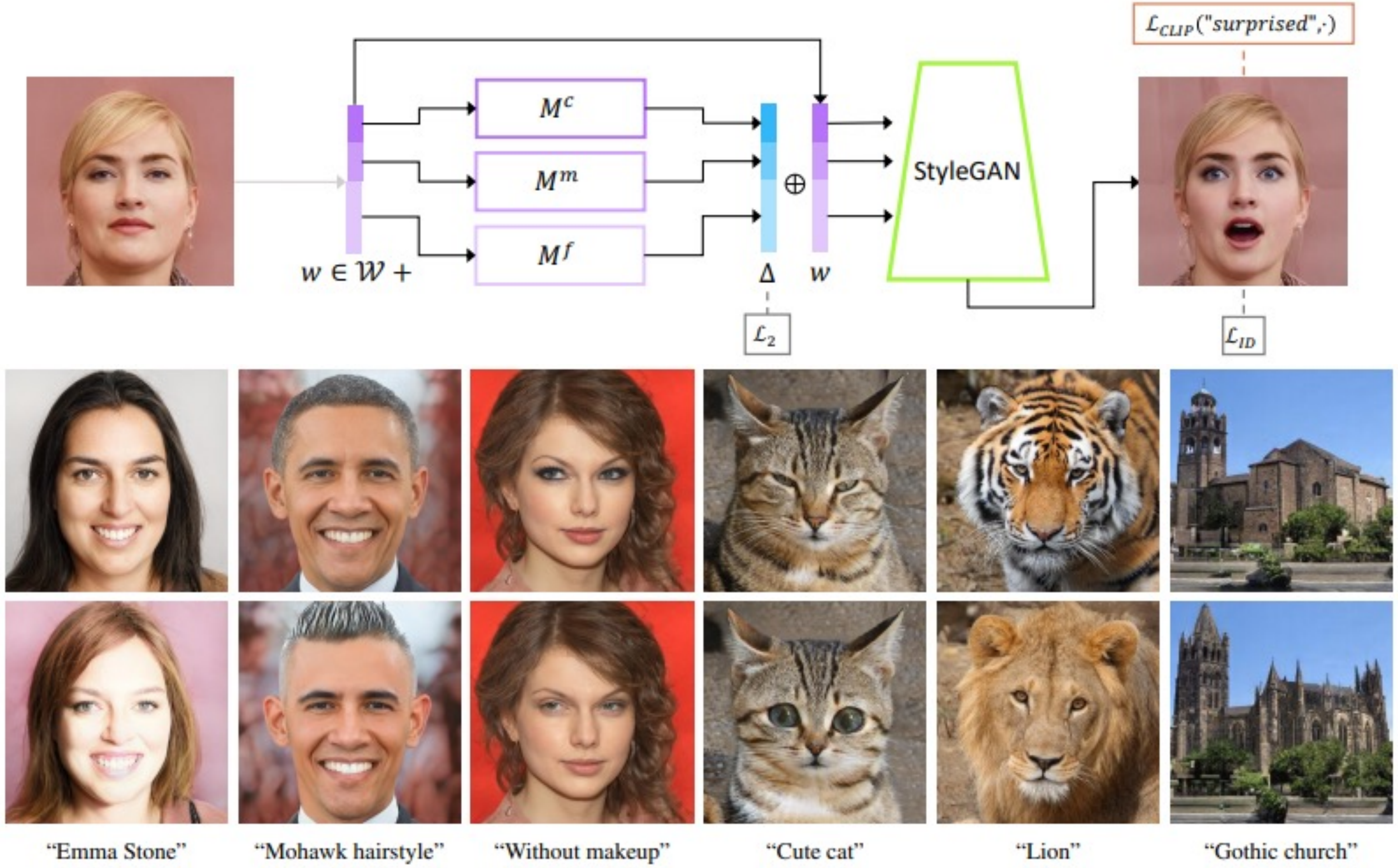


DCGAN
[Radford et al. 2016]



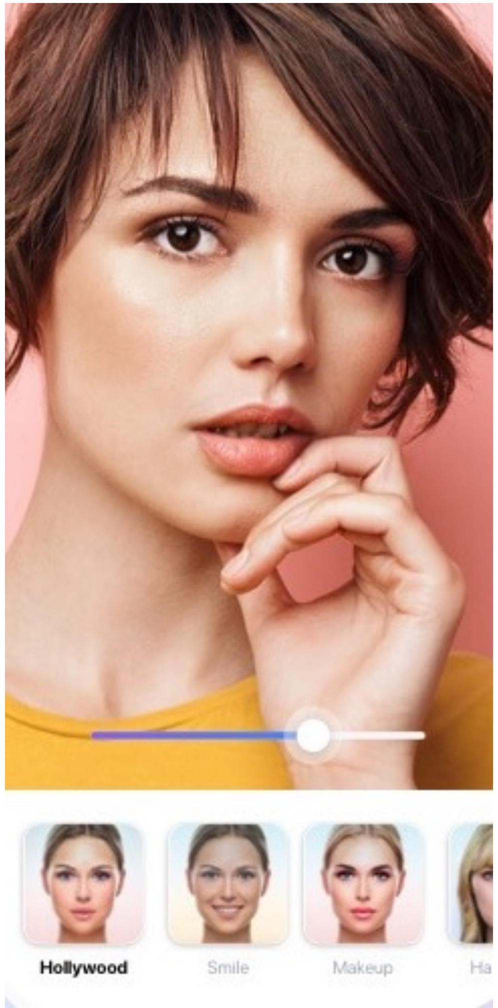
StyleGAN
[Karras et al. 2019]

The Evolution of Content Editing$^{(2/4)}$

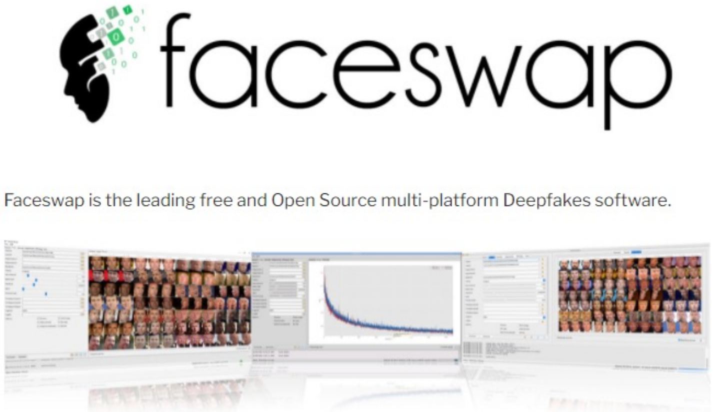


StyleCLIP [Patashnik et al. 2021]

The Evolution of Content Editing(3/4)



FaceApp

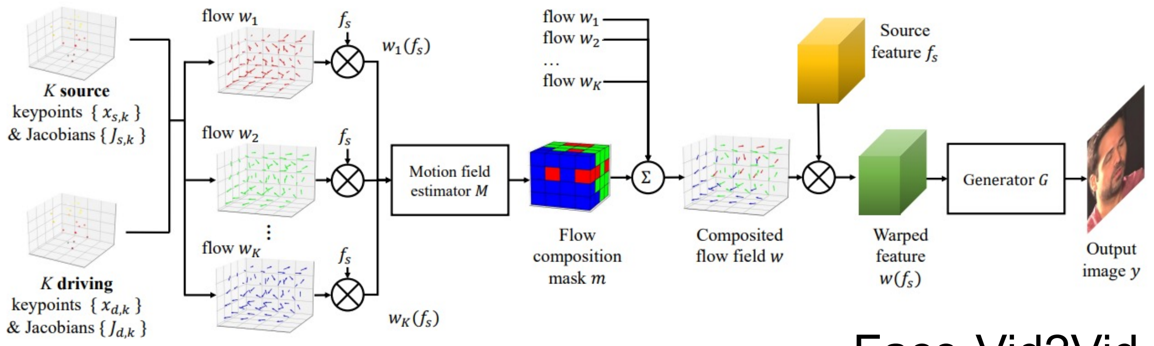


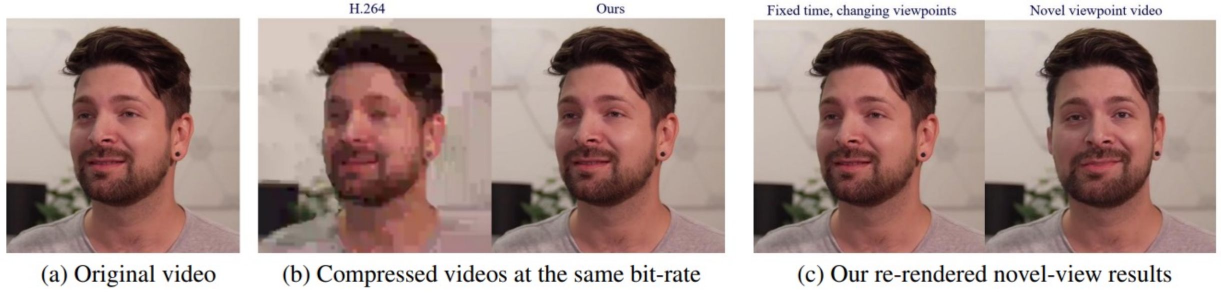
Faceswap

DeepFaceLab

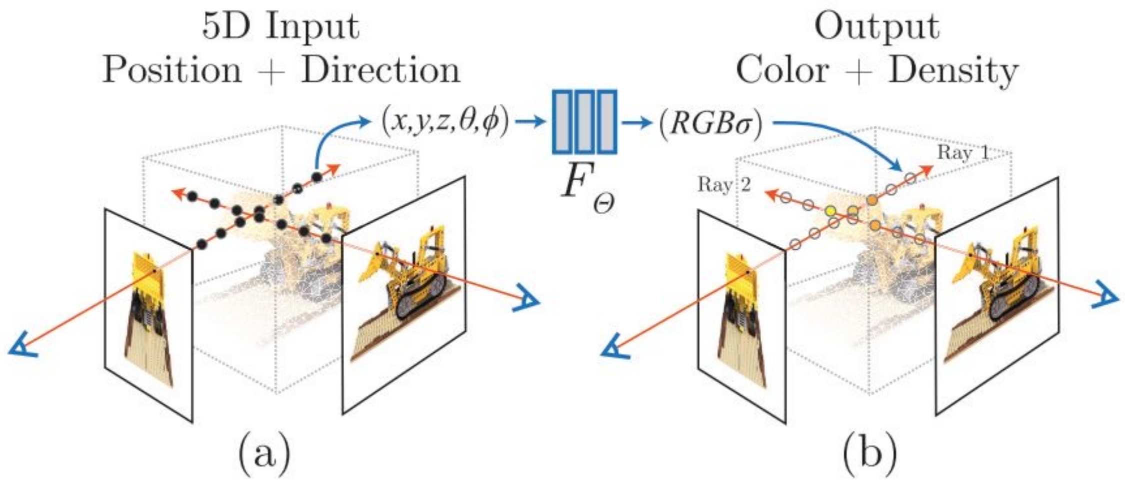
Video credit: *Chris Ume and Miles Fisher*

The Evolution of Content Editing^(4/4)

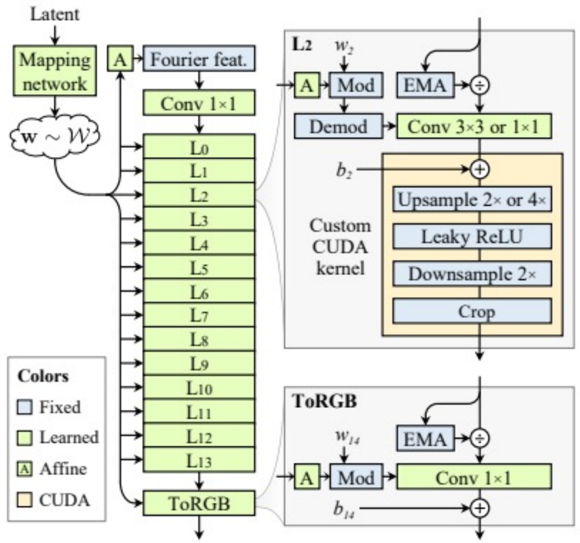




Face-Vid2Vid [Wang et al. 2021]



NeRF [Mildenhall et al. 2020]



StyleGAN3
[Karras et al. 2021]

Challenges

- The evolution of the deepfake technology is ongoing and upgrading in a very fast speed.
- The technologies are widely accessible to the public and much easier to use than before.

Possible Countermeasures

- Passive Defense
  - Deepfake Detection
  - Digital Watermark
- Active Defense
  - Adversarial Attack

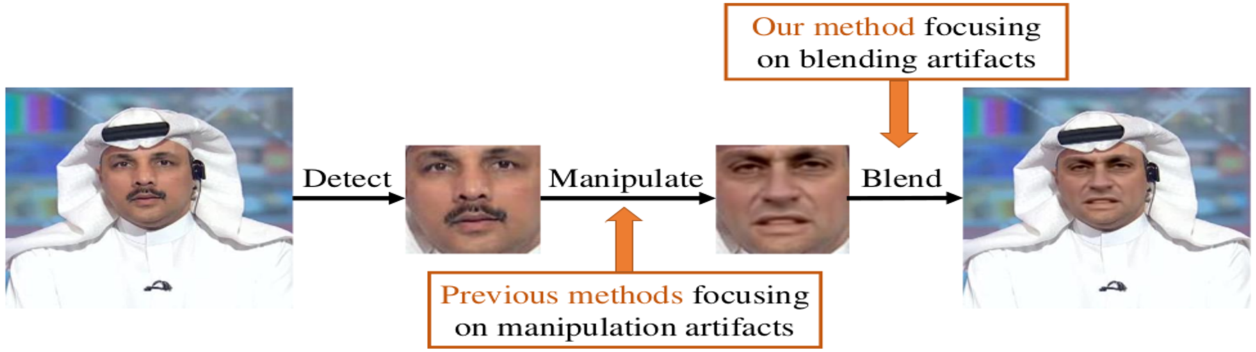
Deepfake Detection

- Sample visual cues for detection

StyleGAN
[Karras et al. 2019]

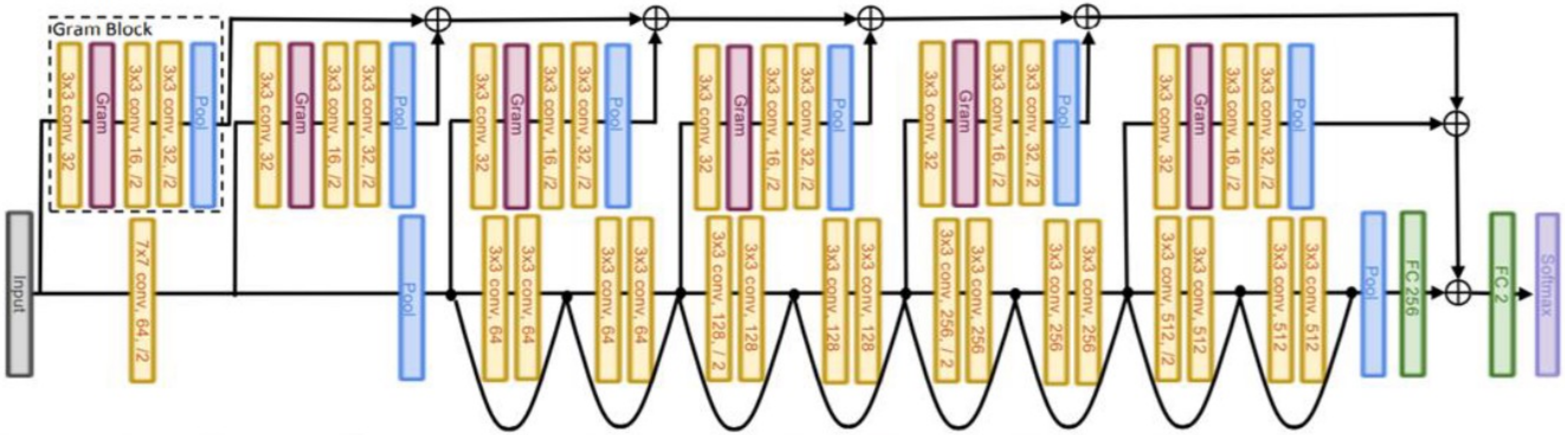
StyleGAN
[Karras et al. 2019]

Deepfake@FaceForensics++
[Rössler et al. 2019]



FaceXRay
[Li et al. 2020]

Global Texture Enhancement for Fake Face Detection In the Wild



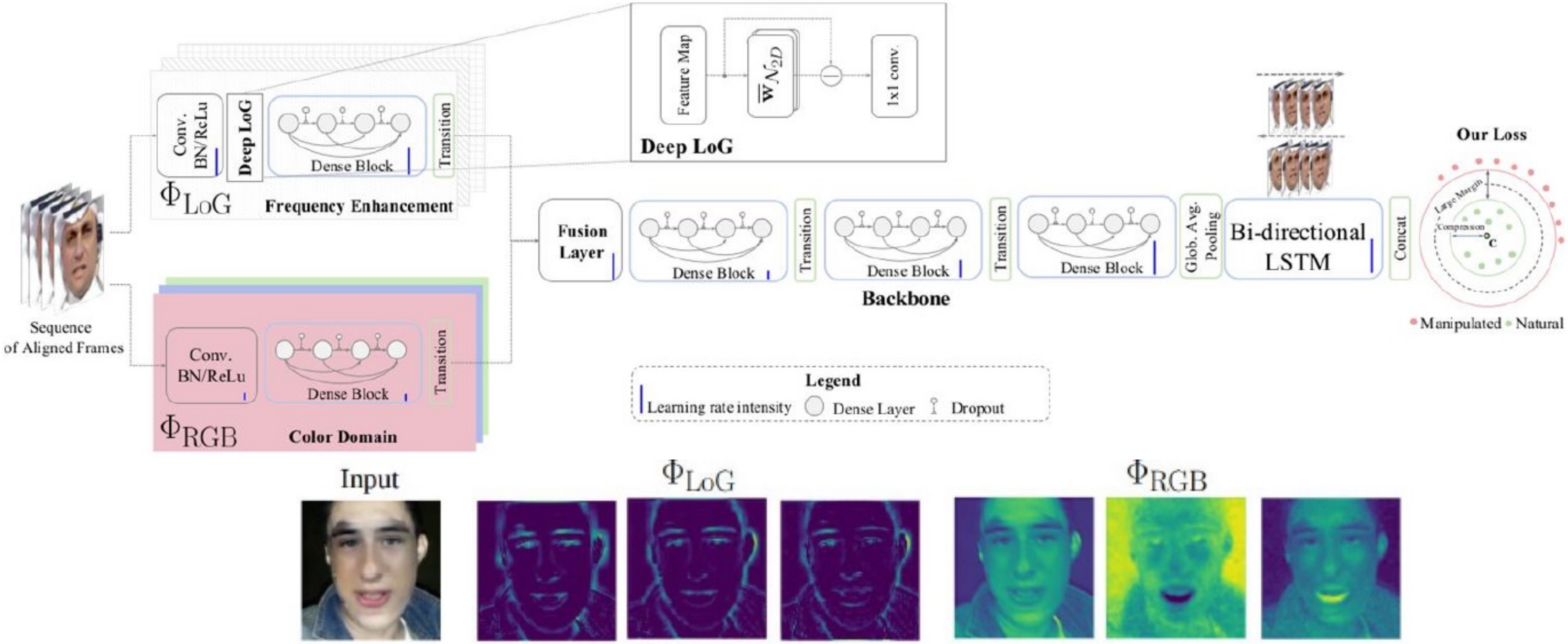
512x512 64x64 512x512 64x64 kernel size 25 std 5

Training set	Testing set	Method	Original %	8x ↓ %	JPEG %	JPEG 8x ↓	Blur %	Noise %	Avg.
StyleGAN vs. CelebA-HQ	StyleGAN vs. CelebA-HQ	Co-detect	79.93 ± 1.34	71.80 ± 1.30	74.58 ± 3.25	71.25 ± 1.18	71.39 ± 1.42	54.09 ± 2.45	70.51
		ResNet	96.73 ± 3.60	85.10 ± 6.22	96.68 ± 3.50	83.33 ± 5.95	79.48 ± 8.70	87.92 ± 6.16	88.20
		Gram-Net	**99.10 ± 1.36**	**95.84 ± 1.98**	**99.05 ± 1.37**	**92.39 ± 2.66**	**94.20 ± 5.57**	**92.47 ± 4.52**	**95.51**
	PGGAN vs. CelebA-HQ	Co-detect	71.22 ± 3.76	62.02 ± 2.86	64.08 ± 1.93	61.24 ± 2.28	62.46 ± 3.31	49.96 ± 0.28	61.83
		ResNet	93.74 ± 3.03	77.75 ± 4.82	89.35 ± 1.50	69.35 ± 3.25	78.06 ± 7.57	82.65 ± 2.37	81.82
		Gram-Net	**98.54 ± 1.27**	**82.40 ± 6.30**	**94.65 ± 3.28**	**79.77 ± 6.13**	**91.96 ± 4.78**	**88.29 ± 3.44**	**89.26**
PGGAN vs. CelebA-HQ	PGGAN vs. CelebA-HQ	Co-detect	91.14 ± 0.61	82.94 ± 1.03	86.00 ± 1.70	82.46 ± 1.06	84.24 ± 0.93	54.77 ± 2.42	80.26
		ResNet	97.38 ± 0.52	90.87 ± 1.90	94.67 ± 1.15	89.93 ± 1.50	97.25 ± 0.87	66.60 ± 9.61	89.45
		Gram-Net	**98.78 ± 0.49**	**94.66 ± 3.10**	**97.29 ± 1.05**	**94.08 ± 3.22**	**98.55 ± 0.92**	**70.32 ± 12.04**	**92.28**
	StyleGAN vs. CelebA-HQ	Co-detect	57.30 ± 1.62	57.41 ± 0.85	52.90 ± 1.67	82.46 ± 1.06	57.41 ± 0.93	50.08 ± 0.10	51.47
		ResNet	97.98 ± 1.90	87.91 ± 1.01	92.03 ± 4.14	82.23 ± 1.39	94.79 ± 1.32	**60.89 ± 7.24**	85.97
		Gram-Net	**98.55 ± 0.89**	**91.57 ± 2.95**	**94.28 ± 3.67**	**83.64 ± 3.43**	**97.05 ± 1.04**	60.07 ± 7.32	**87.52**
StyleGAN vs. FFHQ	StyleGAN vs. FFHQ	Co-detect	69.73 ± 2.41	67.27 ± 1.68	67.48 ± 2.83	64.65 ± 1.67	64.55 ± 1.93	54.66 ± 3.97	64.74
		ResNet	90.27 ± 3.05	70.99 ± 1.13	89.35 ± 3.42	67.96 ± 1.13	**75.60 ± 10.75**	81.32 ± 5.06	81.50
		Gram-Net	**98.96 ± 0.51**	**89.22 ± 4.44**	**98.69 ± 0.81**	**87.86 ± 3.42**	70.99 ± 6.07	**94.27 ± 2.12**	**90.00**

$$G^l = (F_i^{lT} F_j^l)_{n\times n} = \begin{bmatrix} F_1^{lT}F_1^l & \cdots & F_1^{lT}F_n^l \\ \vdots & \ddots & \\ F_n^{lT}F_1^l & \cdots & F_n^{lT}F_n^l \end{bmatrix}$$

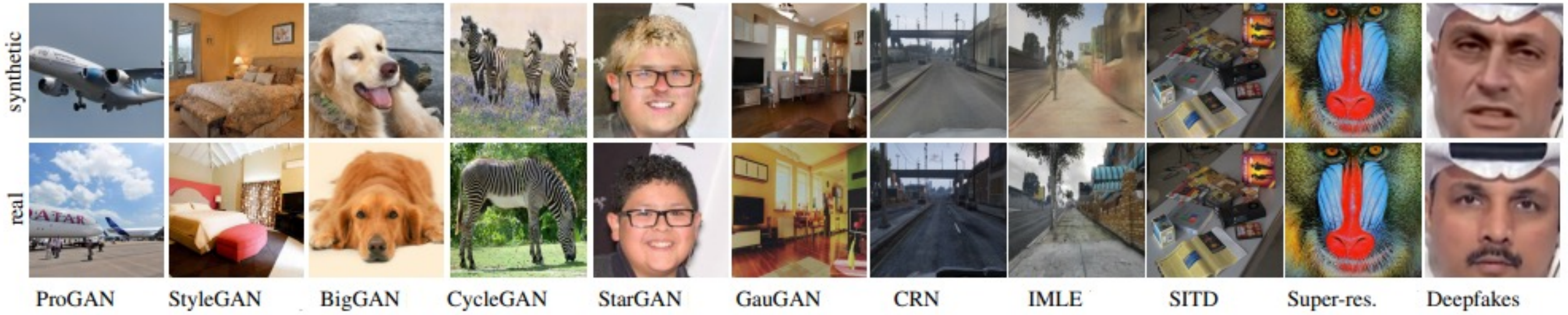
[Liu et al. 2020]

Two-branch Recurrent Network for Isolating Deepfakes in Videos



[Masi et al. 2020]

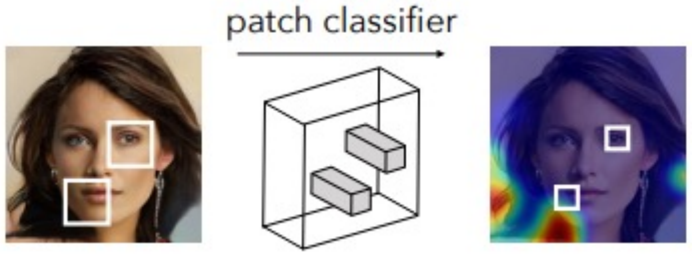
CNN-generated images are surprisingly easy to spot... for now

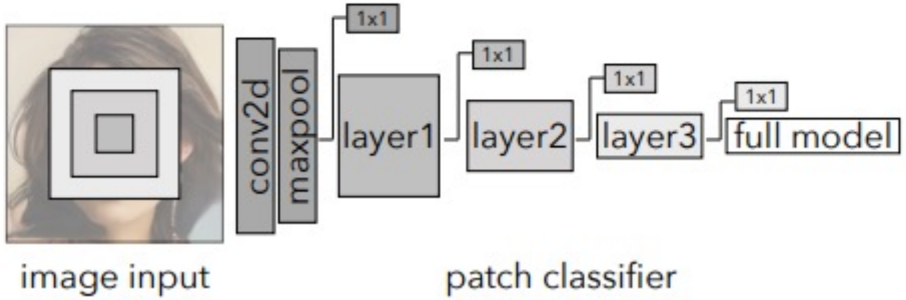


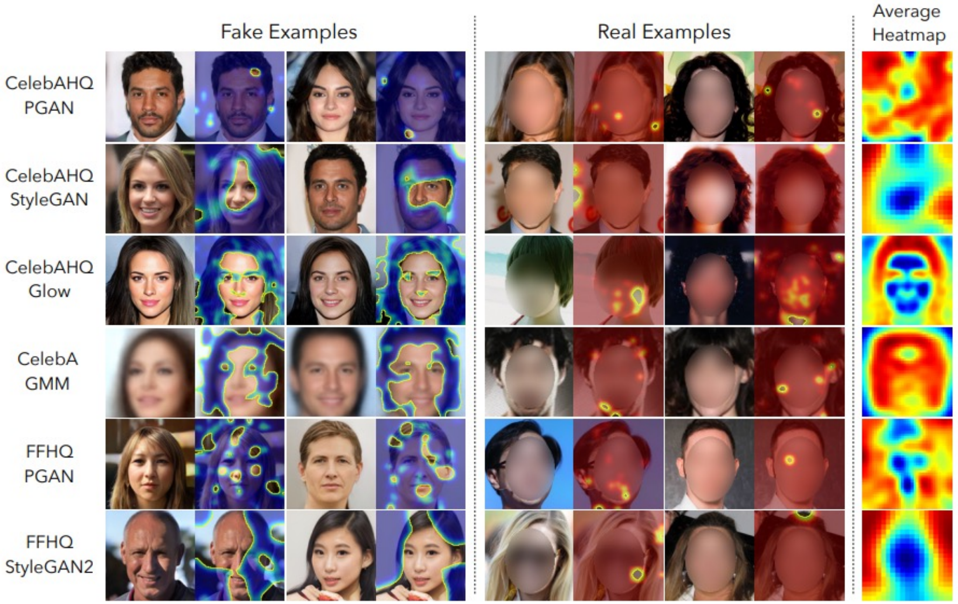
Family	Name	Training settings: Train	Input	No. Class	Augments: Blur	JPEG	Individual test generators: Pro-GAN	Style-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	CRN	IMLE	SITD	SAN	Deep-Fake	Total: mAP
Zhang et al. [50]	Cyc-Im	CycleGAN	RGB	–			84.3	65.7	55.1	100.	99.2	79.9	74.5	90.6	67.8	82.9	53.2	77.6
	Cyc-Spec	CycleGAN	Spec	–			51.4	52.7	79.6	100.	**100.**	70.8	64.7	71.3	92.2	78.5	44.5	73.2
	Auto-Im	AutoGAN	RGB	–			73.8	60.1	46.1	99.9	**100.**	49.0	82.5	71.0	80.1	86.7	80.8	75.5
	Auto-Spec	AutoGAN	Spec	–			75.6	68.6	84.9	**100.**	**100.**	61.0	80.8	75.3	89.9	66.1	39.0	76.5
Ours	2-class	ProGAN	RGB	2	✓	✓	98.8	78.3	66.4	88.7	87.3	87.4	94.0	97.3	85.2	52.9	58.1	81.3
	4-class	ProGAN	RGB	4	✓	✓	99.8	87.0	74.0	93.2	92.3	94.1	95.8	97.5	87.8	58.5	59.6	85.4
	8-class	ProGAN	RGB	8	✓	✓	99.9	94.2	78.9	94.3	91.9	95.4	98.9	99.4	91.2	58.6	63.8	87.9
	16-class	ProGAN	RGB	16	✓	✓	100.	98.2	87.7	96.4	95.5	**98.1**	**99.0**	**99.7**	95.3	63.1	71.9	91.4
	No aug	ProGAN	RGB	20			100.	96.3	72.2	84.0	**100.**	67.0	93.5	90.3	96.2	**93.6**	**98.2**	90.1
	Blur only	ProGAN	RGB	20	✓		100.	99.0	82.5	90.1	**100.**	74.7	66.6	66.7	**99.6**	53.7	95.1	84.4
	JPEG only	ProGAN	RGB	20		✓	100.	99.0	87.8	93.2	91.8	97.5	**99.0**	99.5	88.7	78.1	88.1	**93.0**
	Blur+JPEG (0.5)	ProGAN	RGB	20	✓	✓	100.	98.5	**88.2**	96.8	95.4	**98.1**	98.9	99.5	92.7	63.9	66.3	90.8
	Blur+JPEG (0.1)	ProGAN	RGB	20	†	†	100.	**99.6**	84.5	93.5	98.2	89.5	98.2	98.4	97.2	70.5	89.0	92.6

[Wang et al. 2020]

What makes fake images detectable?
Understanding properties that generalize



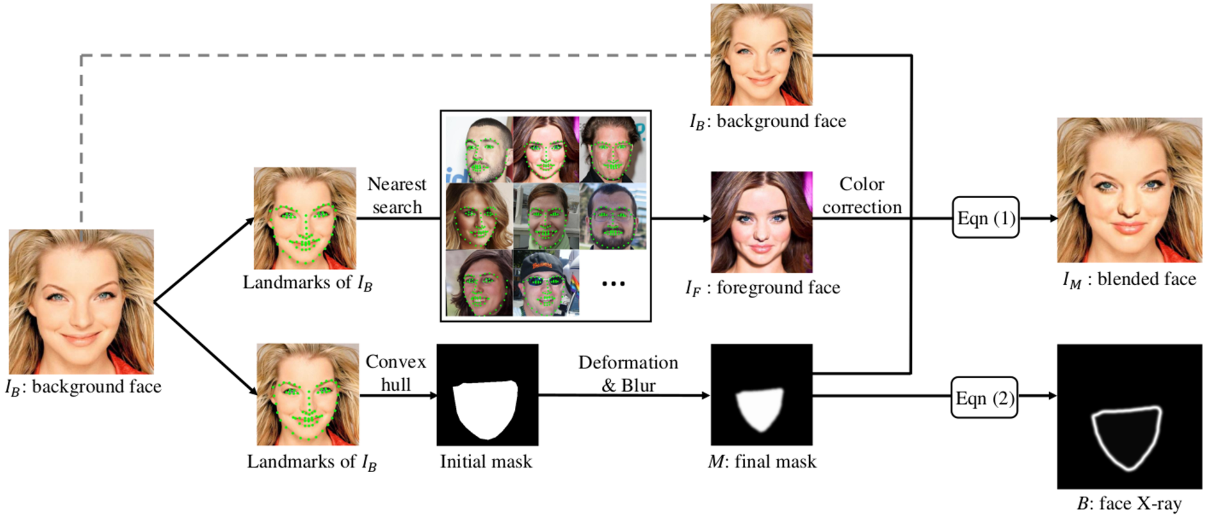


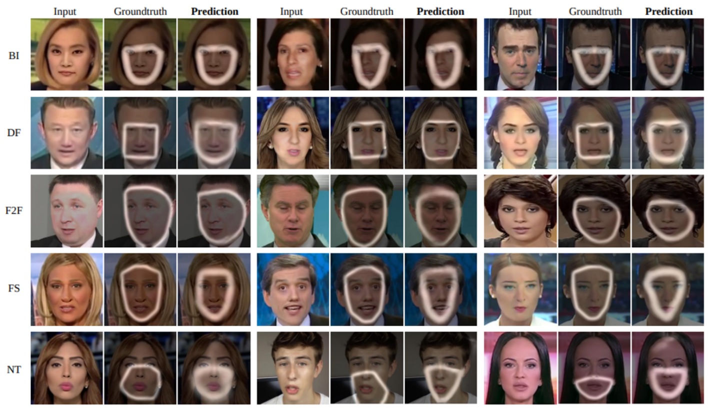


	Architectures				FFHQ dataset		
Model	**PGAN**	**SGAN**	**Glow***	**GMM**	**PGAN**	**SGAN**	**SGAN2**
Resnet Layer 1	100.0	97.22	72.80	80.69	99.81	72.91	71.81
Xception Block 1	100.0	98.68	95.48	76.21	99.68	81.35	77.40
Xception Block 2	100.0	99.99	67.49	**91.38**	**100.0**	90.12	90.85
Xception Block 3	100.0	**100.0**	74.98	80.96	100.0	92.91	**91.45**
Xception Block 4	100.0	99.99	66.79	42.82	100.0	**95.85**	90.62
Xception Block 5	100.0	100.0	60.44	48.92	100.0	93.09	89.08
[2] MesoInception4	100.0	97.90	49.72	45.98	98.71	80.57	71.27
[13] Resnet-18	100.0	64.80	47.06	54.69	79.20	51.15	52.37
[6] Xception	100.0	99.75	55.85	40.98	99.94	85.69	74.33
[33] CNN (p=0.1)	100.0	98.41	90.46	50.65	99.95	90.48	85.27
[33] CNN (p=0.5)	100.0	97.34	**97.32**	73.33	99.93	88.98	84.58

[Chai et al. 2020]

Face X-ray for More General Face Forgery Detection

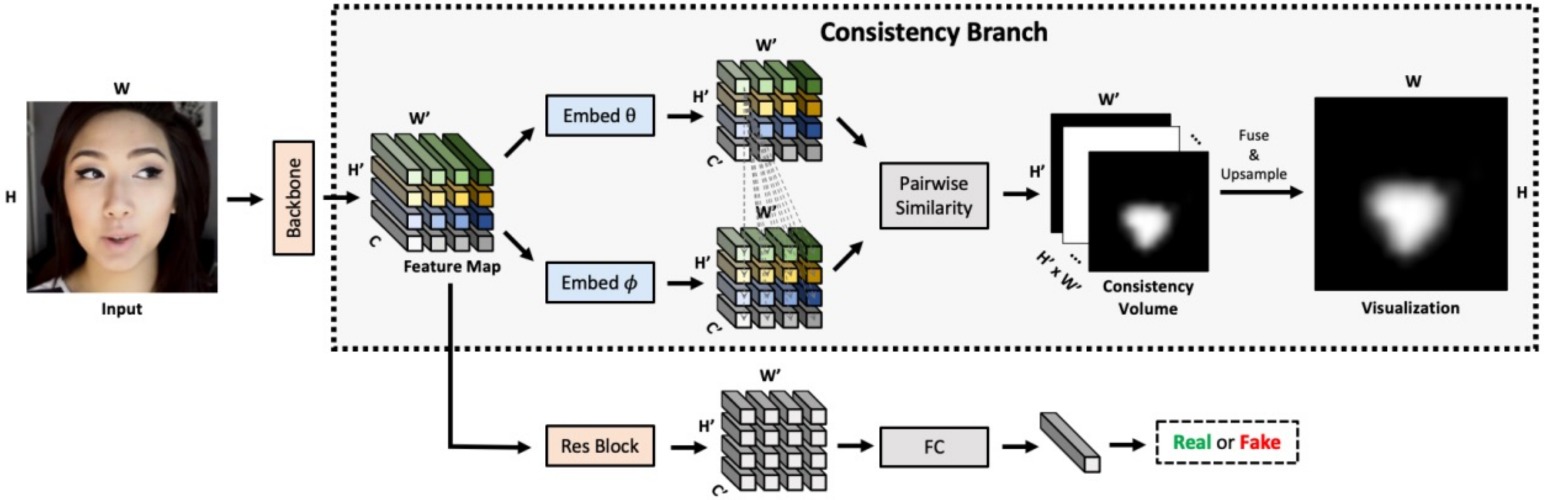




Model	Training set		Test set AUC				
	DF	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	99.38	75.05	49.13	80.39	76.34
HRNet	✓	–	99.26	68.25	39.15	71.39	69.51
Face X-ray	✓	–	99.17	94.14	75.34	93.85	90.62
	✓	✓	99.12	**97.64**	**98.00**	**97.77**	**97.97**
	F2F	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	87.56	99.53	65.23	65.90	79.55
HRNet	✓	–	83.64	99.50	56.60	61.26	74.71
Face X-ray	✓	–	98.52	99.06	72.69	91.49	93.41
	✓	✓	**99.03**	99.31	**98.64**	**98.14**	**98.78**
	FS	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	70.12	61.70	99.36	68.71	74.91
HRNet	✓	–	63.59	64.12	99.24	68.89	73.96
Face X-ray	✓	–	93.77	92.29	99.20	86.63	93.13
	✓	✓	**99.10**	**98.16**	99.09	**96.66**	**98.25**
	NT	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	93.09	84.82	47.98	99.50	83.42
HRNet	✓	–	94.05	87.26	64.10	98.61	86.01
Face X-ray	✓	–	99.14	98.43	70.56	98.93	91.76
	✓	✓	**99.27**	**98.43**	**97.85**	99.27	**98.71**
	FF++	BI	DF	F2F	FS	NT	FF++
Xception [36]	–	✓	98.95	97.86	89.29	97.29	95.85
HRNet	–	✓	99.11	97.42	83.15	**98.17**	94.46
Face X-ray	–	✓	**99.17**	**98.57**	**98.21**	98.13	**98.52**

[Li et al. 2020]

Learning Self-Consistency for Deepfake Detection

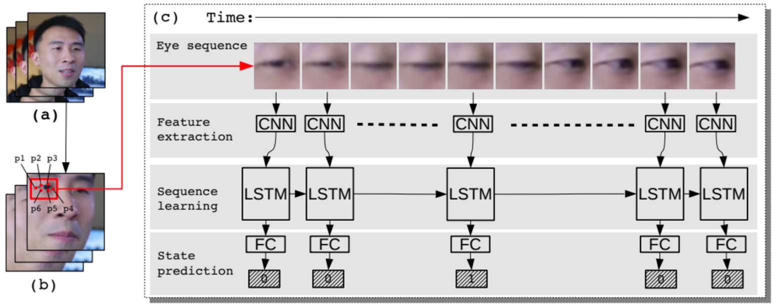


Method	Backbone	Train Set	Test Set (AUC (%))				
			DF	F2F	FS	NT	FF++
MIL [59]	Xception	FF++	99.51	98.59	94.86	97.96	97.73
Fakespotter [56]	ResNet-50	FF++, CD2, DFDC	-	-	-	-	98.50
XN-avg [45]	Xception	FF++	99.38	99.53	99.36	97.29	98.89
Face X-ray [25]	HRNet	FF++	99.12	99.31	99.09	99.27	99.20
S-MIL-T [27]	Xception	FF++	99.84	99.34	99.61	98.85	99.41
PCL + I2G	ResNet-34	FF++	**100.00**	**99.57**	**100.00**	**99.58**	**99.79**

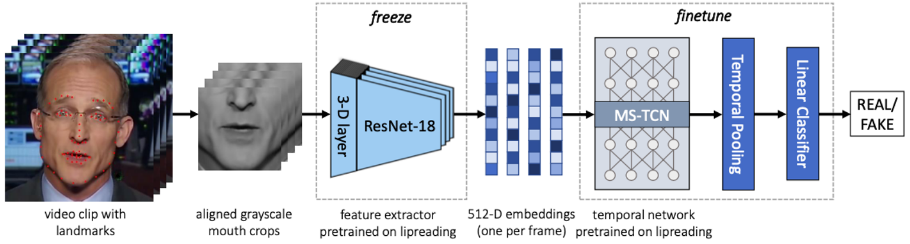
[Zhao et al. 2021]

Temporal Consistency

- Video Inconsistency between frames

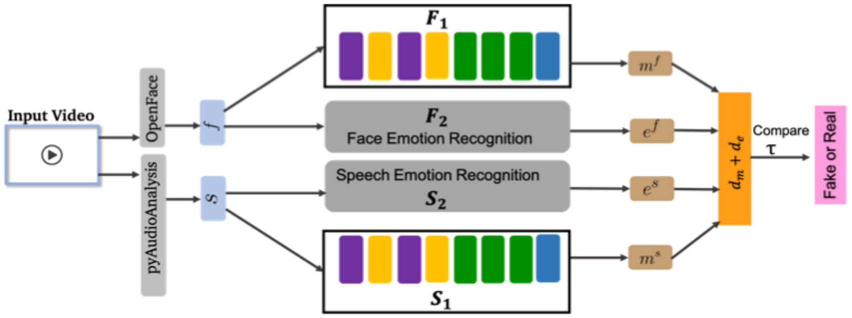


In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, WIFS 2018

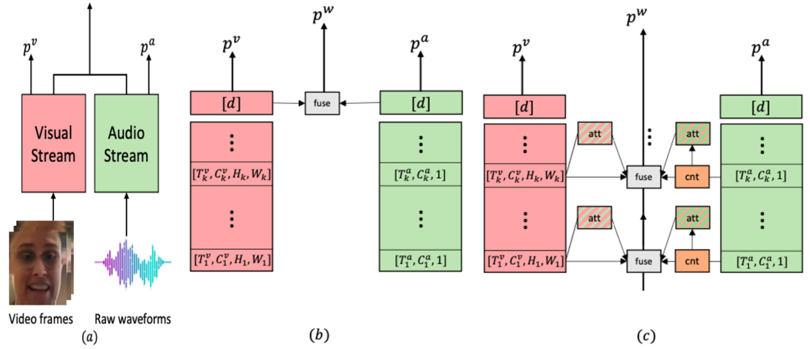


Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection, CVPR 2021

- Audio-visual inconsistency



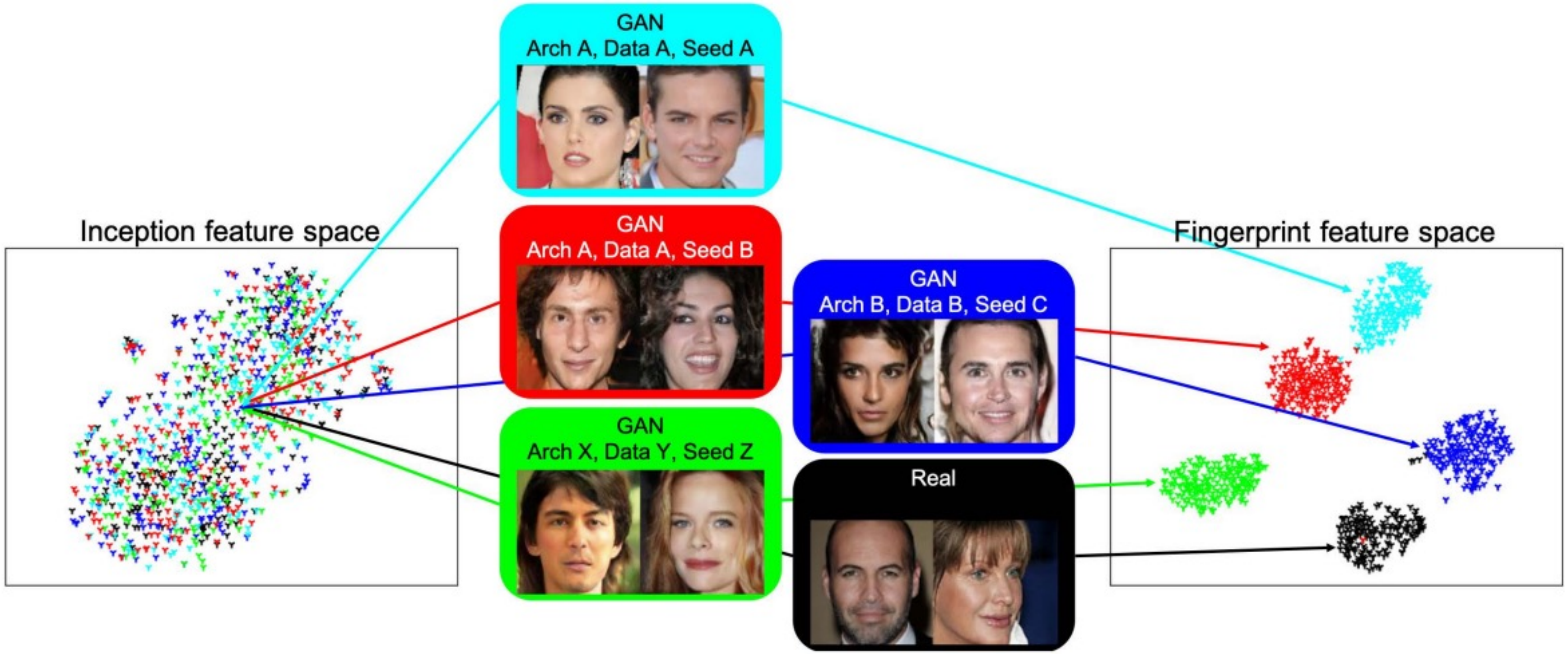
Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues, ACMMM 2020



Joint Audio-Visual Deepfake Detection, ICCV 2021

Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

- Every GAN has its fingerprint.

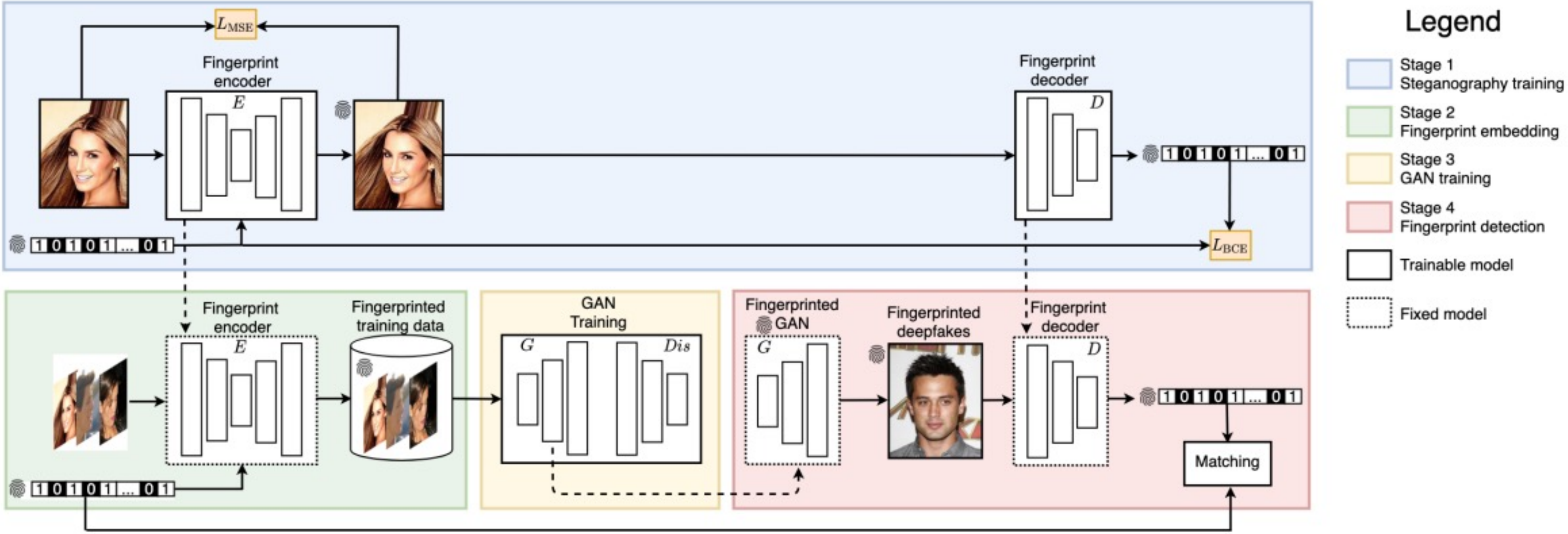


[Ning et al. 2019]

Possible Countermeasures

- Passive Defense
  - Deepfake Detection
  - Digital Watermark
- Active Defense
  - Adversarial Attack

Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data

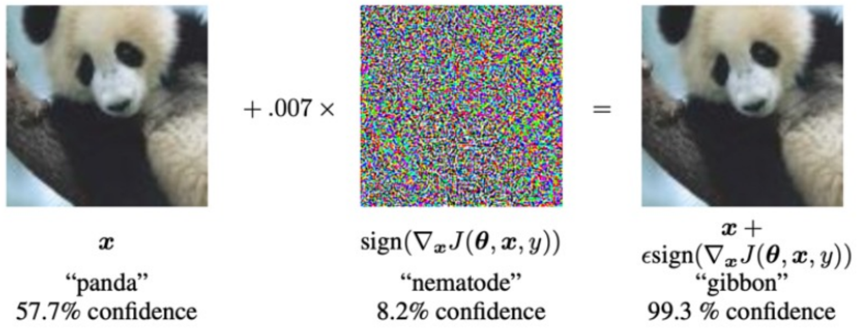


[Ning et al. 2021]

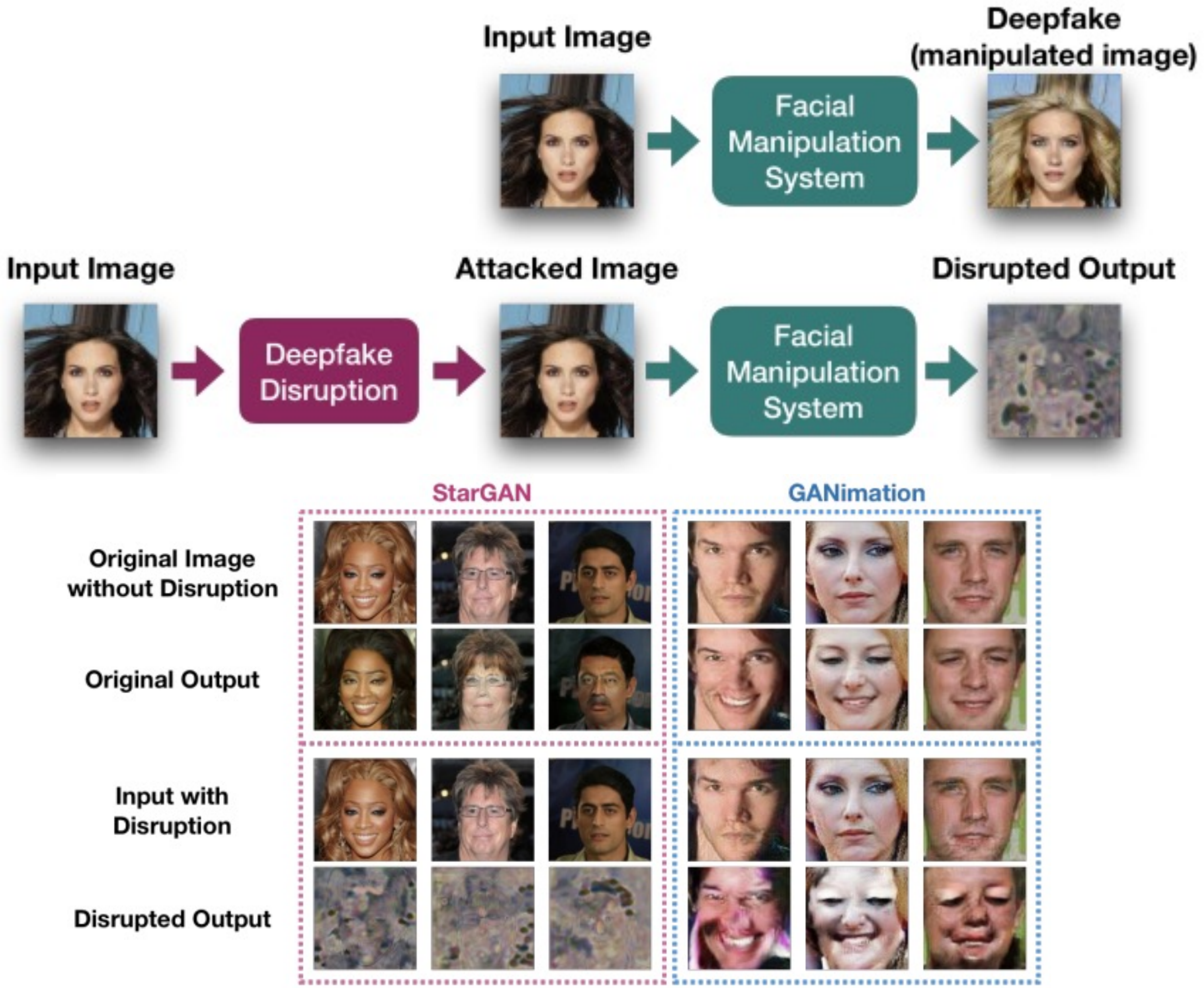
Possible Countermeasures

- Passive Defense
  - Deepfake Detection
  - Digital Watermark
- Active Defense
  - Adversarial Attack

Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems

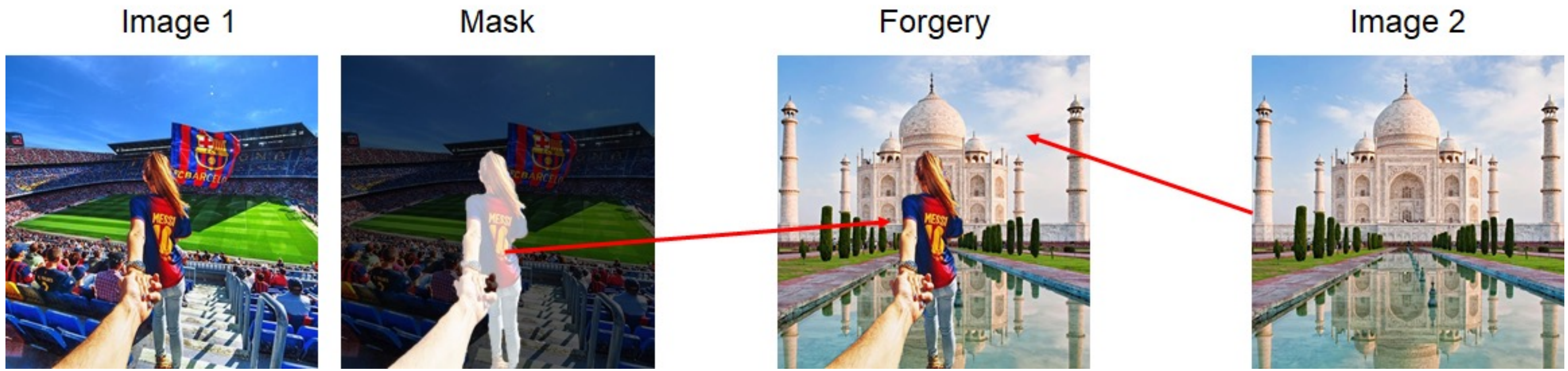


[Goodfellow et al. 2015]



[Ruiz et al. 2020]

Making Forgeries



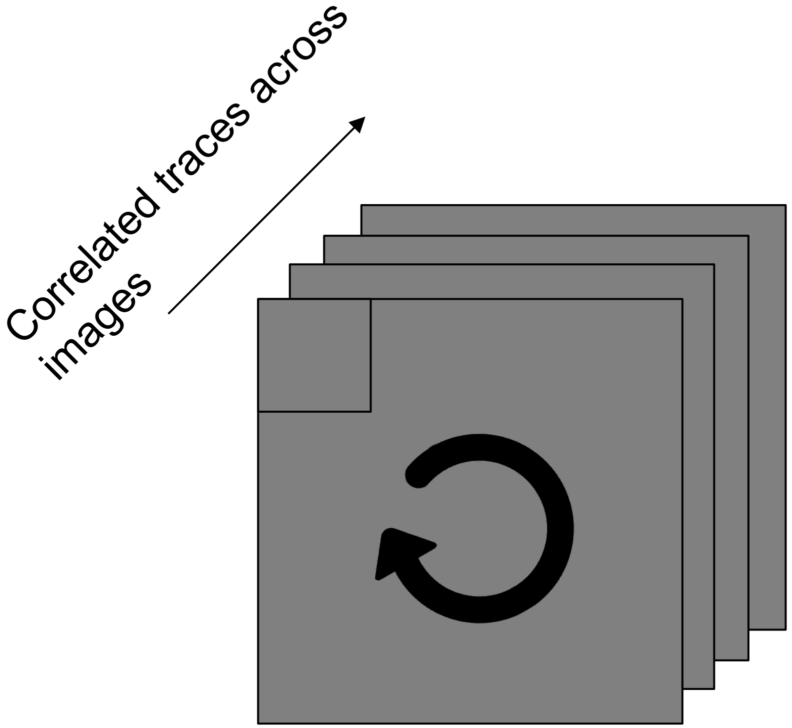
Traces in images allow us to detect forgery

Correlated traces across images

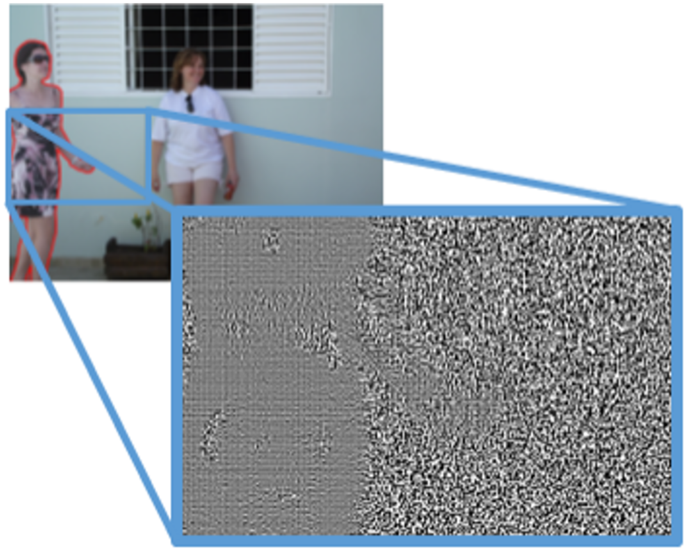
- Photo-response non uniformity noise (PRNU)

Correlated traces within images (usually periodic)

- Compression (e.g. blocking)
- Resampling
- Demosaicing

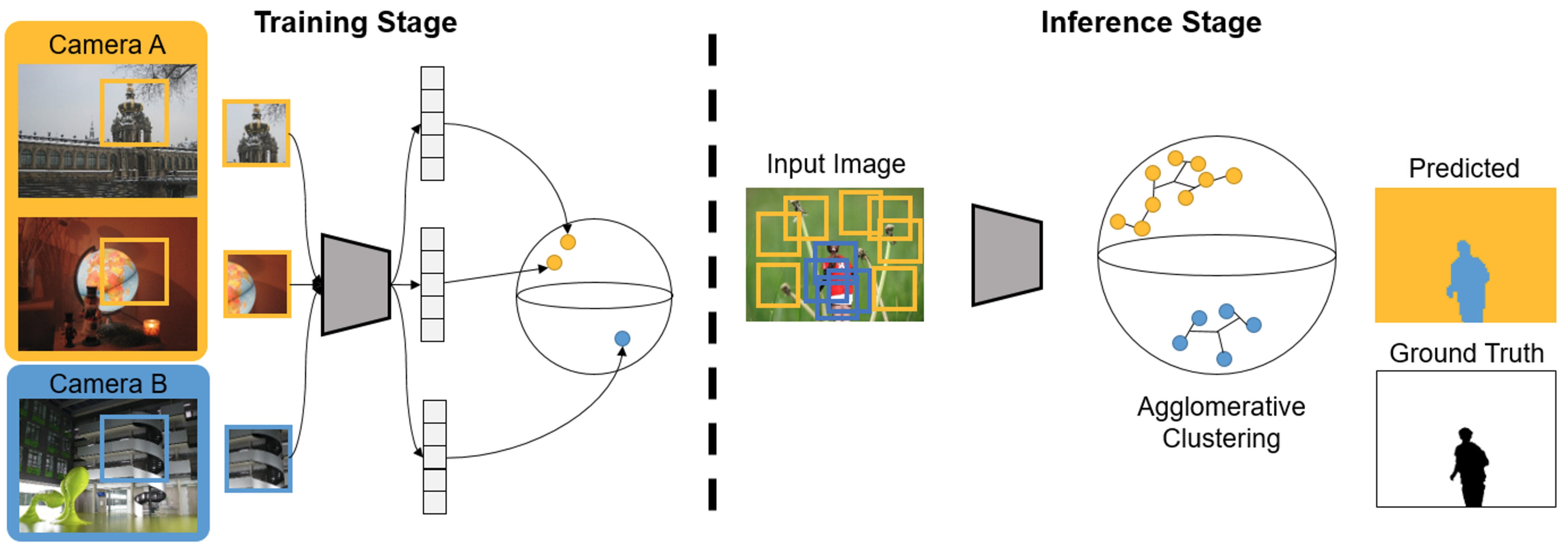


Correlated traces within images



PRNU Differences

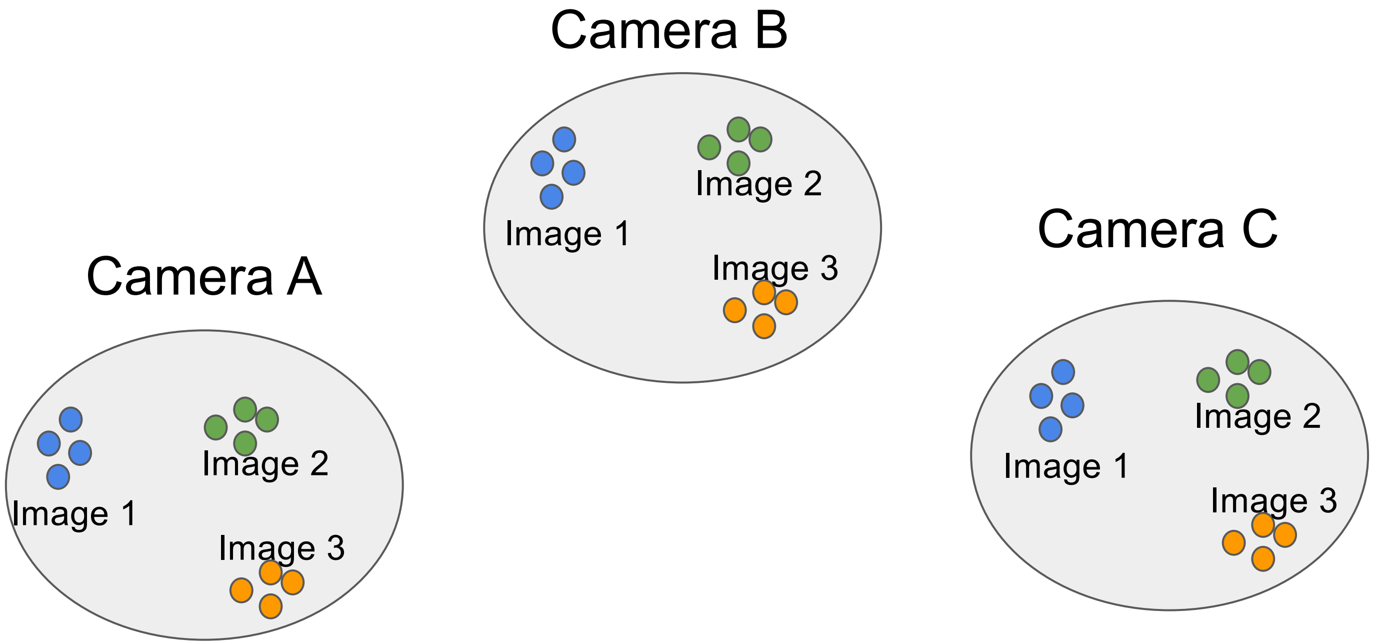
Patch Contrastive Learning



- Under preparation for submission

Learned Patch Embeddings

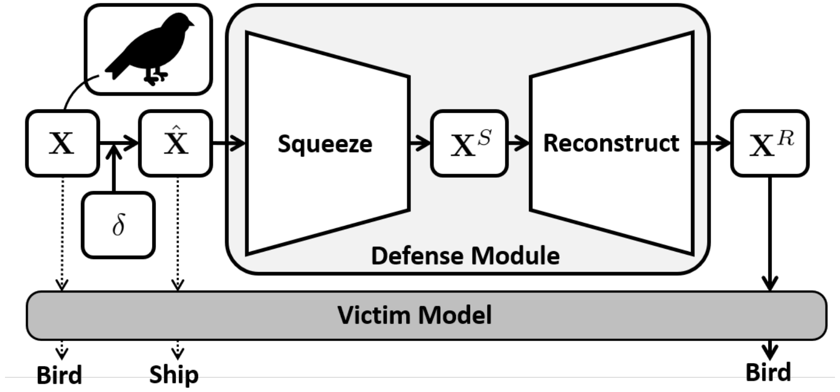
- We want the patch embeddings to be able to discriminate between images taken from different cameras as well as differentiate patches belonging to the same image.

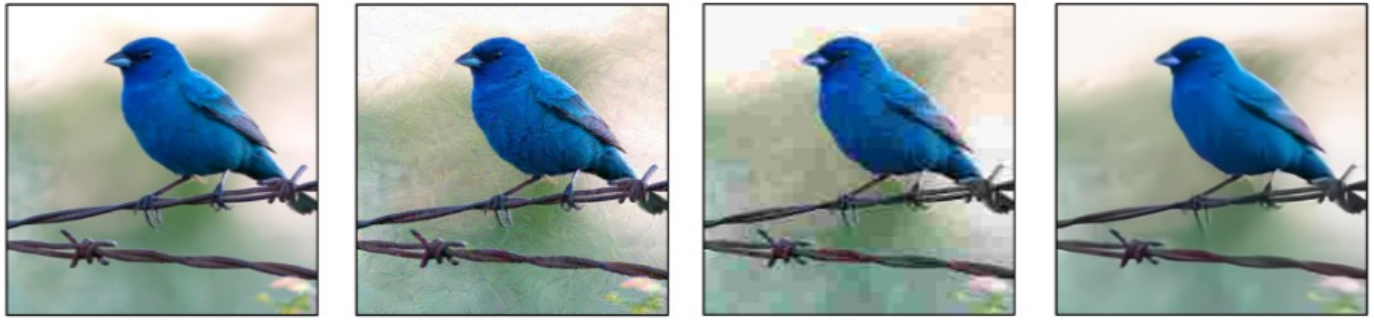


Some Results

Adversarial Defense for Image Classifier

- Non-robust features reconstruction.
- Pre-processing based defense.
- Outperform SOTA comparable methods.





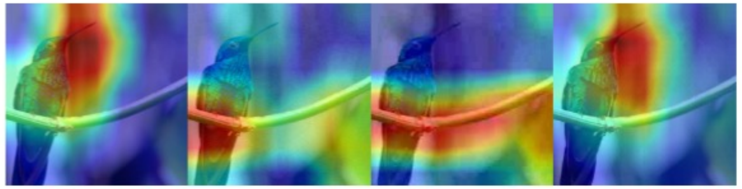
Raw Image
Indigo bunting=99.75%

Attacked Image
Knot=99.99%

JPG Image
Indigo bunting=84.18%

Recovered Image
Indigo bunting=99.72%

(a) Great gray owl = 99.9 (b) Sea anemone = 99.9 (c) Great gray owl = 48.7 (d) Great gray owl = 98.8

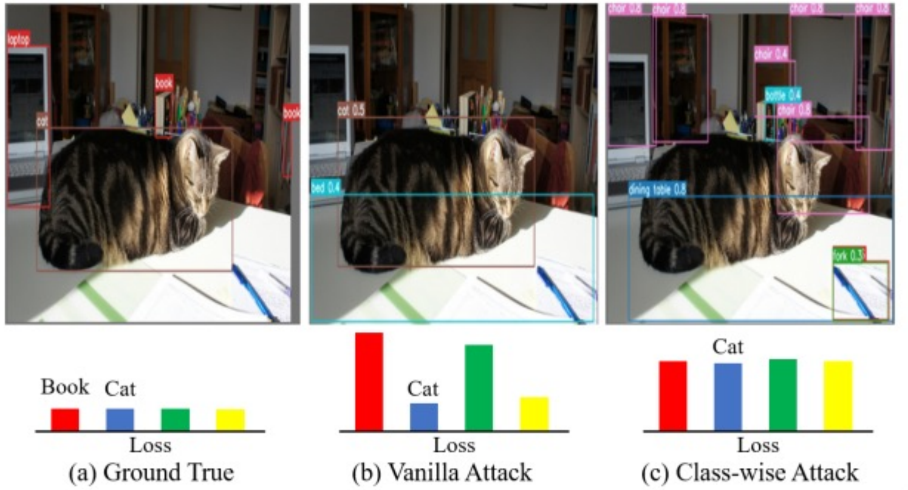


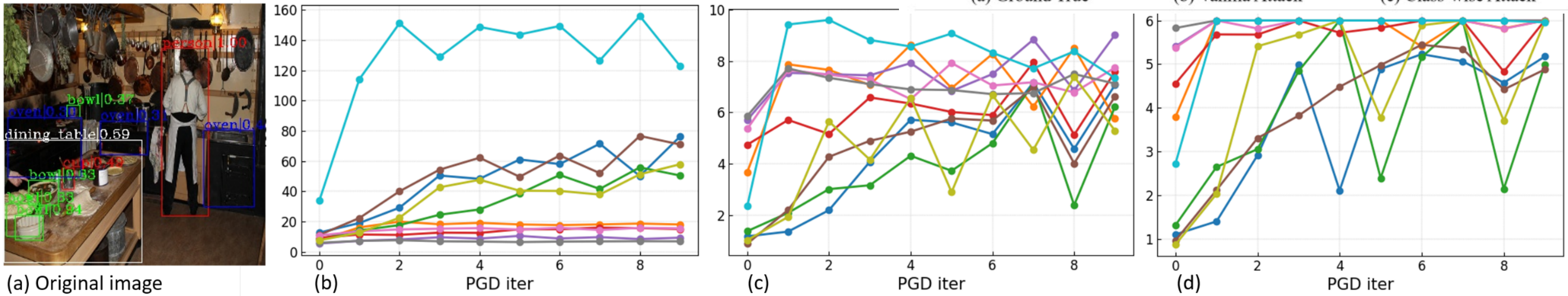
(e) Hummingbird = 85.14 (f) Green snake = 97.7 (g) Vine snake = 72.3 (h) Hummingbird = 80.43

Bo-Han Kung, Pin-Chun Chen, Yu-Cheng Liu, Jun-Cheng Chen, "Squeeze and Reconstruct: Improved Practical Adversarial Defense using Paired Image Compression and Reconstruction," *IEEE International Conference on Image Processing*, September 2021.

Adversarial Defense for Object Detector

- Vanilla PGD training:
  - Imbalance attack (loss dominates)
  - Overfitting
- Proposed method:
  - Balance the loss of each object

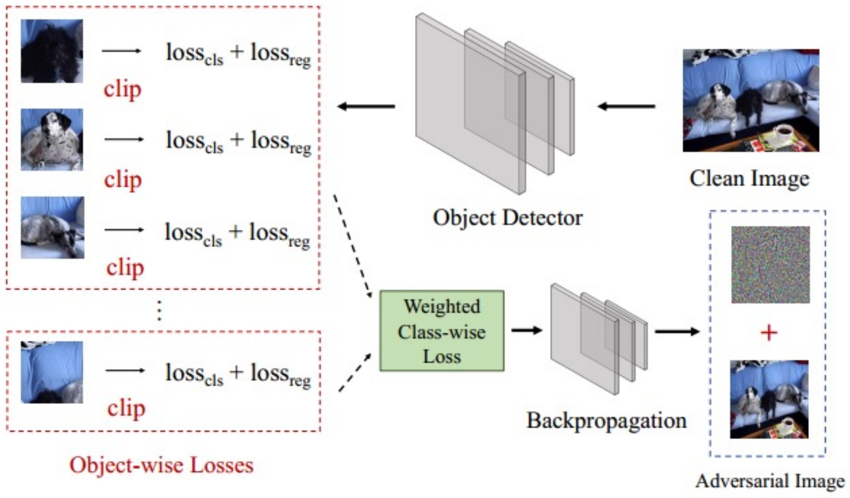


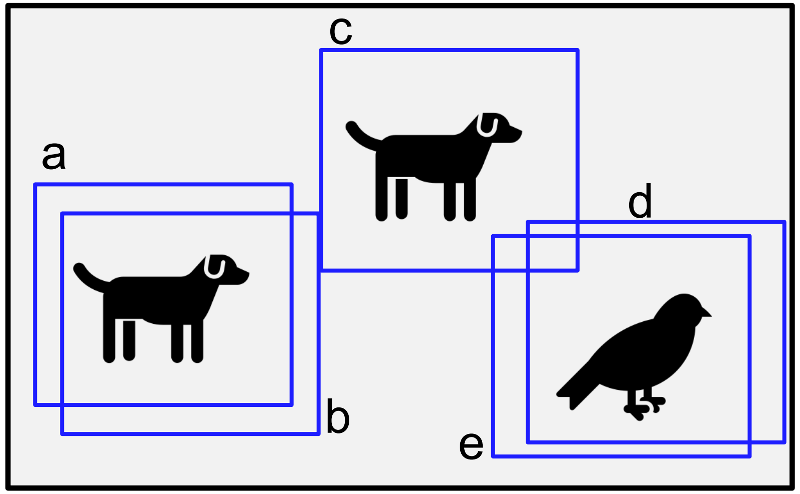


- Pin-Chun Chen, Bo-Han Kung, Jun-Cheng Chen, "Class-Aware Robust Adversarial Training for Object Detection," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Class-aware Adversarial Training

- TOAT $\min_\theta \max_{\|\delta\|_p \le \epsilon} \mathcal{L}(\theta,\delta) = \hat{l}_{cls}(x+\delta,\{y\},\theta) + \hat{l}_{reg}(x+\delta,\{b\},\theta)$
- OWAT $\min_\theta \max_{\|\delta\|_p \le \epsilon} \mathcal{L}(\theta,\delta) = \sum_{i=1}^{N_o} \hat{l}^o_{cls}(O_i+\delta,\{y_i\},\theta) + \sum_{i=1}^{N_o} \hat{l}^o_{reg}(O_i+\delta,\{b_i\},\theta)$
- CWAT $\min_\theta \max_{\|\delta\|_p \le \epsilon} \mathcal{L}_{\mathcal{C}}{}' = \frac{1}{C}\sum_{i=1}^{C} \frac{1}{n_i}\sum_{j=1}^{n_i} \hat{l}^o_{cls}(O_j,\{y_j\},\theta) + \hat{l}^o_{reg}(O_j,\{b_j\},\theta)$





Class-aware Adversarial Training

- Adopt “forfree” method.
- 7-30 times faster than vanilla methods.
- Better performance on COCO and PASCAL datasets.

attack	clean	FGSM		PGD-10		CWA
		A_{cls}	A_{reg}	A_{cls}	A_{reg}	
STD	**0.451**	0.133	0.167	0.030	0.029	0.003
MTD[1]	0.190	0.127	0.146	0.110	0.135	0.082
MTD-fast	0.242	0.167	0.182	0.130	0.134	0.077
TOAT-6	0.182	0.120	0.148	0.098	0.123	0.074
OWAT	0.211	0.129	0.169	0.100	0.140	0.074
CWAT	0.237	**0.168**	**0.189**	**0.142**	**0.155**	**0.092**

Table 1: MS-COCO test set.

Algorithm 1 Fast Class-wise Adversarial Training

Require: dataset D, training epoch N_{ep}, perturbation bound ϵ, learning rate γ

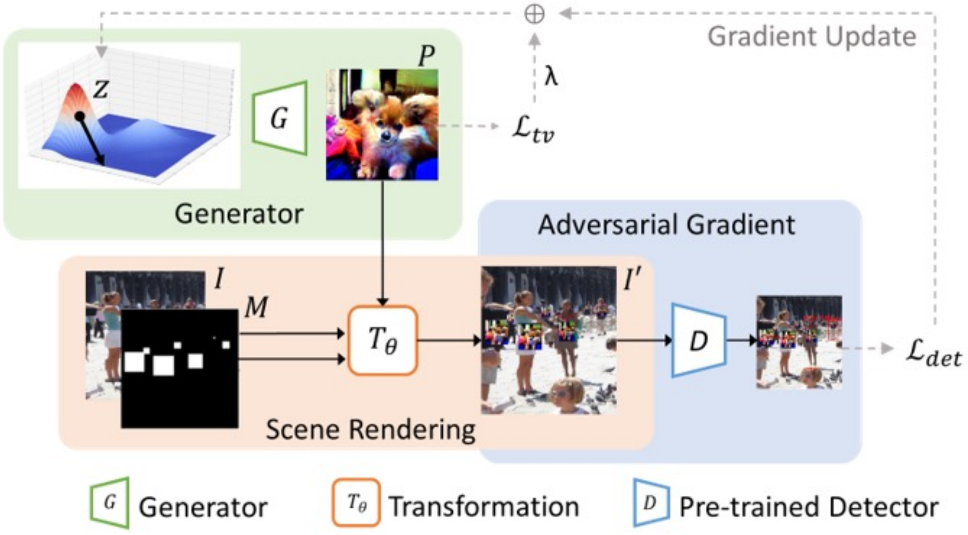
for epoch = 1, ..., N_{ep}/m **do**
  for minibatch $B \sim D$ **do**
    for $iter = 1$ to m **do**
      Compute gradient of loss with respect to δ
      $d_\delta \leftarrow \mathbb{E}_{x \in B}\left[\nabla_\delta \mathcal{L}_C{}'(\theta, x+\delta)\right]$
      Update θ with momentum stochastic gradient
      $g_\theta \leftarrow \mu g_\theta - \mathbb{E}_{x \in B}\left[\nabla_\theta \mathcal{L}(\theta, x+\delta)\right]$
      $\theta \leftarrow \theta + \gamma g_\theta$
      Update perturbation δ with gradient
      $\delta \leftarrow \delta + \epsilon sign(d_\delta)$
      Project δ to ℓ_p-ball
    end for
  end for
end for

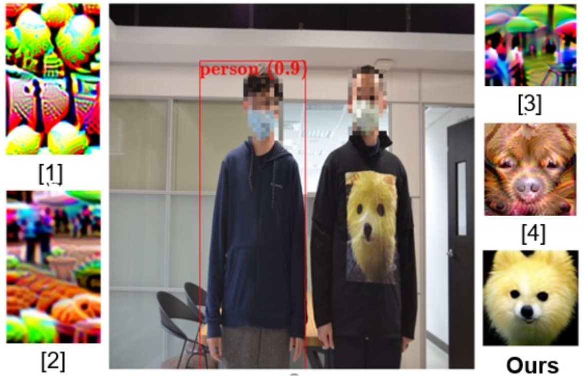
Class-aware Adversarial Training

(a) Clean Image Result (b) Vanilla Adversarial Attack (c) Multi-task domain attack (d) Object-wise attack (e) Class-wise Attack

(e) No attack; Model: **STD** (f) Attack: **CWA**; Model: **STD** (g) Attack: **CWA**; Model: **CWAT** (h) Attack: **DAG**; Model: **CWAT**

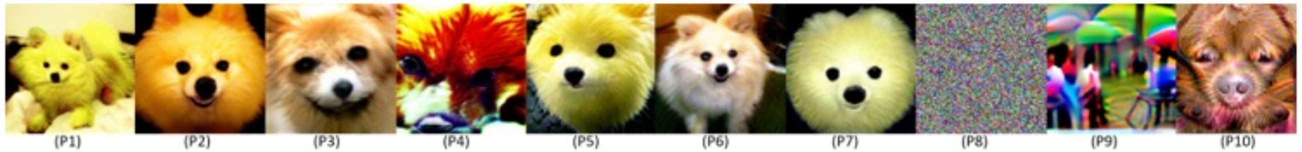
Naturalistic Physical Adversarial Patch for Object Detectors





Trained on \ Victim	YOLOv2	YOLOv3	YOLO3tiny	YOLOv4	YOLOv4tiny	FasterRCNN
(P1)Ours-YOLOv2	**12.06**	43.50	32.12	50.56	24.89	52.54
(P2)Ours-YOLOv3	56.67	**34.93**	41.46	56.29	37.46	61.78
(P3)Ours-YOLOv3tiny	31.61	28.81	**10.02**	65.13	18.61	55.08
(P4)Ours-YOLOv4	44.27	56.59	56.61	**22.63**	50.04	59.42
(P5)Ours-YOLOv4tiny	34.68	37.79	21.69	46.80	**8.67**	59.97
(P6)Ours-FasterRCNN	28.26	39.05	37.06	51.46	29.06	**42.47**
(P7)Ours-ensemble†	49.42	35.46	25.29	51.71	18.51	61.28
Gray	72.66	74.17	67.52	66.52	64.74	61.54
(P8)Random	75.03	73.75	78.91	76.71	75.74	73.00
White	69.63	74.93	66.45	72.48	59.66	65.40
(P9)Adversarial Patches* [42]	2.13	22.51	8.74	12.89	3.25	39.41
(P10)UPC** [19]	48.62	54.40	63.82	64.21	57.93	61.87

†trained on YOLOv2+YOLOv3+YOLOv4tiny * trained on YOLO ** trained on FasterRCNN



Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, Wen-Huang Cheng, "Naturally Physical Adversarial Patch for Object Detectors," *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

Takeways

- The evolution of the deepfake technologies is fast and requires more ethical consideration for it.
- Educate the public to less rely on the videos as the evidence.

References

- [USC ICT 2015] https://vgl.ict.usc.edu/Research//PresidentialPortrait/
- [Radford et al. 2016] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." ICLR 2016.
- [Karras et al. 2019] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401-4410. 2019.
- [Patashnik et al. 2021] Patashnik, Or, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. "Styleclip: Text-driven manipulation of stylegan imagery." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2085-2094. 2021.
- [Wang et al. 2021] Wang, Ting-Chun, Arun Mallya, and Ming-Yu Liu. "One-shot free-view neural talking-head synthesis for video conferencing." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10039-10049. 2021.
- [Mildenhall et al. 2020] Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. "Nerf: Representing scenes as neural radiance fields for view synthesis." In European conference on computer vision, pp. 405-421. Springer, Cham, 2020.
- [Karras et al. 2021] Karras, Tero, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Alias-free generative adversarial networks." In Thirty-Fifth Conference on Neural Information Processing Systems. 2021.

References

- [Li et al. 2020] Li, Lingzhi, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. "Face x-ray for more general face forgery detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001-5010. 2020.
- [Rössler et al. 2019] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11. 2019.
- [Liu et al. 2020] Liu, Zhengzhe, Xiaojuan Qi, and Philip HS Torr. "Global texture enhancement for fake face detection in the wild." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8060-8069. 2020.
- [Masi et al. 2020] Masi, Iacopo, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. "Two-branch recurrent network for isolating deepfakes in videos." In European Conference on Computer Vision, pp. 667-684. Springer, Cham, 2020.
- [Wang et al. 2020] Wang, Sheng-Yu, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. "CNN-generated images are surprisingly easy to spot... for now." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8695-8704. 2020.

References

- [Chai et al. 2020] Chai, Lucy, David Bau, Ser-Nam Lim, and Phillip Isola. "What makes fake images detectable? understanding properties that generalize." In European Conference on Computer Vision, pp. 103-120. Springer, Cham, 2020.
- [Zhao et al. 2021] Zhao, Tianchen, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. "Learning Self-Consistency for Deepfake Detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15023-15033. 2021.
- [Ning et al. 2019] Yu, Ning, Larry S. Davis, and Mario Fritz. "Attributing fake images to gans: Learning and analyzing gan fingerprints." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7556-7566. 2019.
- [Ning et al. 2021] Yu, Ning, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14448-14457. 2021.
- [Goodfellow et al. 2015] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." ICLR 2015.
- [Ruiz et al. 2020] Ruiz, Nataniel, Sarah Adel Bargal, and Stan Sclaroff. "Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems." In *European Conference on Computer Vision*, pp. 236-251. Springer, Cham, 2020.