A Brief Ovewerview of Recent Advancement of Generative AI and New Challenges to Deepfake Detection

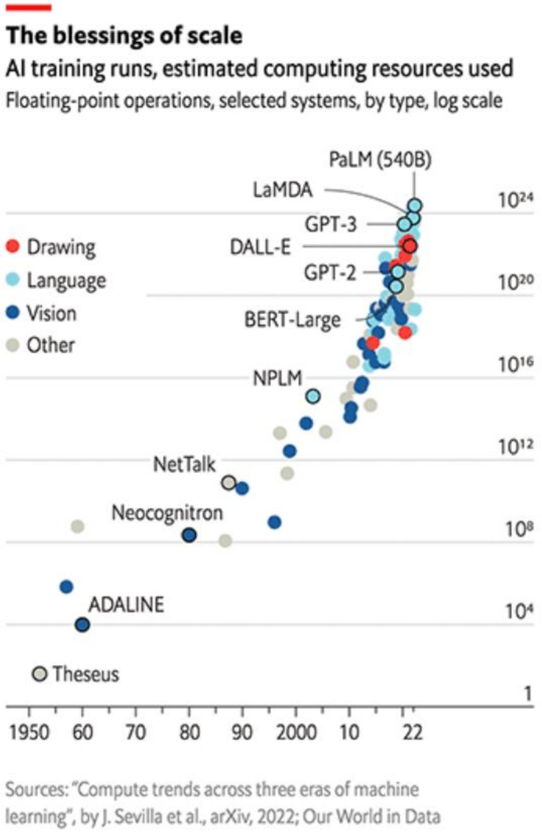
Jun-Cheng Chen

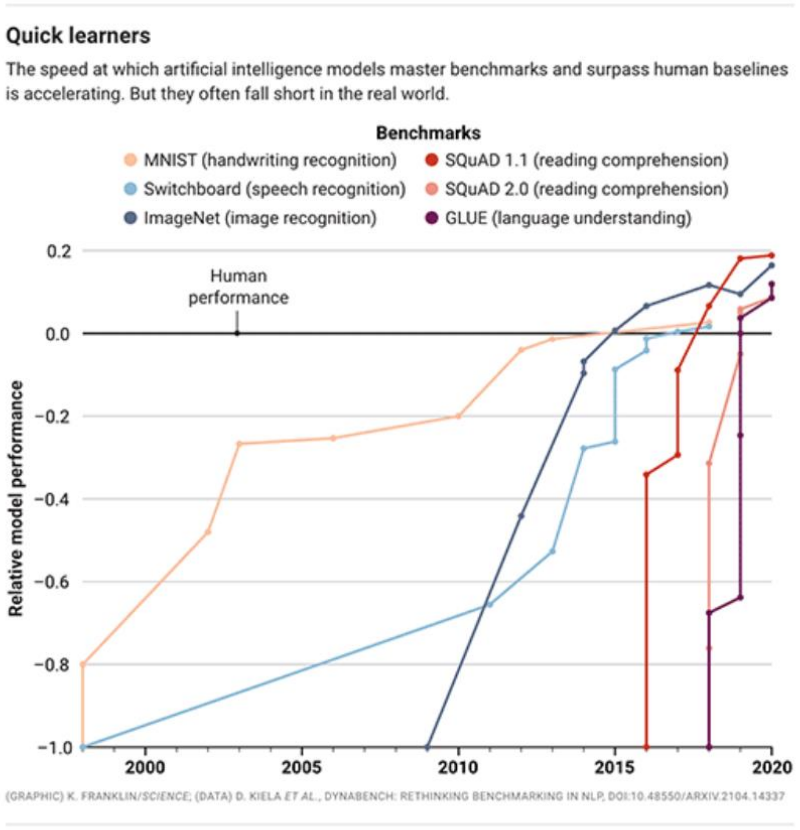
pullpull@citi.sinica.edu.tw
Artificial Intelligence and Image Understanding Lab (AIIU)
Research Center of Information Technology Innovation, Academia Sinica

2022/12/06

Recent Development of Generative AI(1/3)

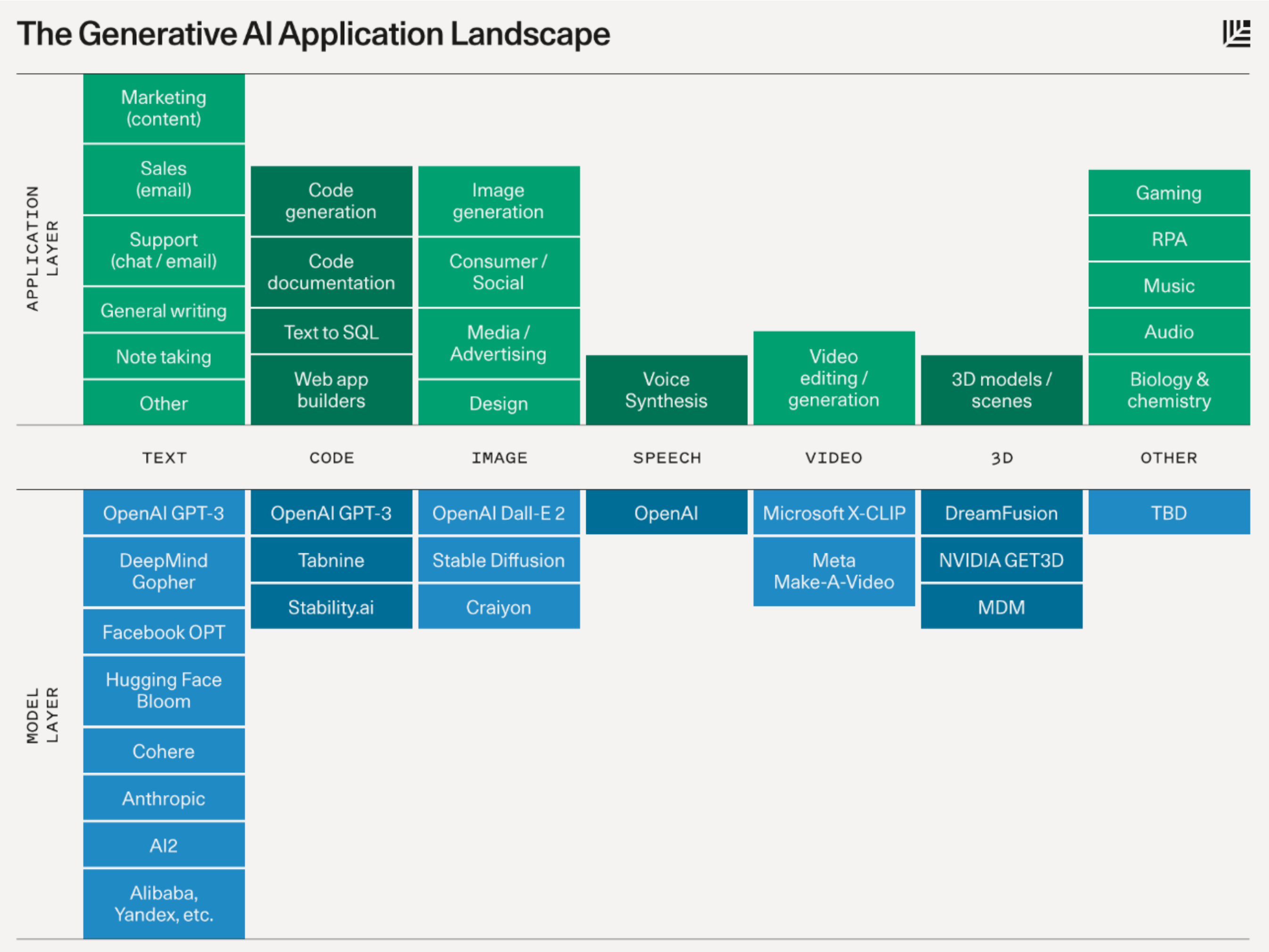
- Large generative AI models for image/video, audio, text make it possible for machines to write, draw, and create credible and amazing content.





Cited from https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/?fbclid=IwAR2GNyKNoEc_pv1TMjhUw2C7QluyIZnvnv9HlW2kD080wqwXV3L3zL-14Sk

Recent Development of Generative AI(2/3)



Cited from https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/?fbclid=IwAR2GNyKNoEc_pv1TMjhUw2C7QluyIZnvnv9HlW2kD080wqwXV3L3zL-14Sk

Recent Development of Generative AI(3/3)

- ChatGPT:

PU Could you help me write a professional article about Forensics?

Sure, I'd be happy to help. Forensics is the application of scientific techniques and methods to investigate criminal activity and answer legal questions. It is a crucial part of the criminal justice system, as it allows investigators to gather and analyze evidence in order to solve crimes and bring perpetrators to justice.

There are many different areas of forensics, including forensic anthropology, which involves the study of human remains to determine things like the cause of death and the identity of the victim; forensic pathology, which focuses on the examination of tissues and organs to determine the cause of death; and forensic psychology, which involves the study of the mind and behavior to help understand why a crime was committed.

Forensic scientists use a variety of tools and techniques to gather and analyze evidence. These can include DNA analysis, fingerprinting, and ballistics, as well as more specialized methods like trace evidence analysis, which looks at things like fibers, hairs, and paint to help link suspects to a crime scene.

PU Could you help write a rap about Forensics?

Sure, here's a short rap about forensics:

Verse 1:
I'm the master of the crime scene,
I gather evidence that's unseen.
I analyze the clues with precision,
To solve the case with expert decision.

Chorus:
Forensics, forensics,
I'm the one who cracks the case.
Forensics, forensics,
I bring the truth to light.

Verse 2:
From fingerprints to DNA,
I use the latest technology every day.
I piece together the puzzle,
To find the culprit and the motive.

Chorus:
Forensics, forensics,
I'm the one who cracks the case.
Forensics, forensics,
I bring the truth to light.

PU Could you help me debug the following code?

```
for(int i = 0; i < 10; i++) {
    printf("%d\n",i)
}
```

Sure, I'd be happy to help. It looks like you have a syntax error in your code. The issue is that you are missing a semicolon at the end of the `printf` statement.

The correct code should look like this:

Copy code

```
for(int i = 0; i < 10; i++) {
  printf("%d\n",i);
}
```

This will print out the numbers from 0 to 9, with each number on a new line.

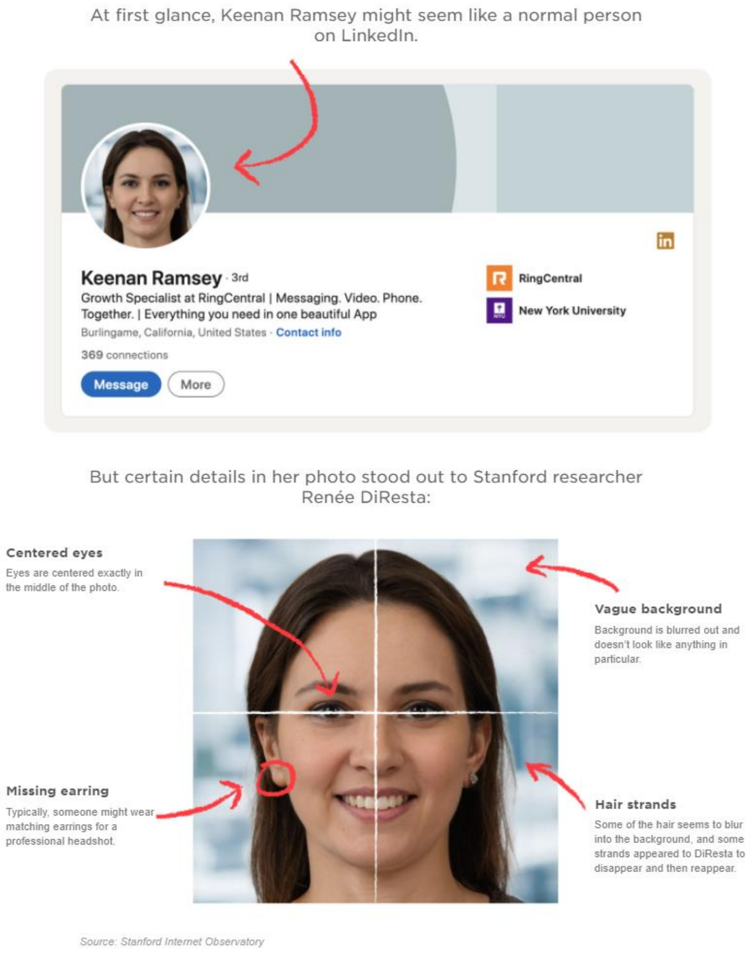
To debug this code, you can use a technique called "stepping through" the code. This means going through the code line by line, checking the values of variables and making sure that each line is executing as expected.

https://openai.com/blog/chatgpt/

Motivation

- Malicious Face Forgery Applications
  - Pornography
  - Politics

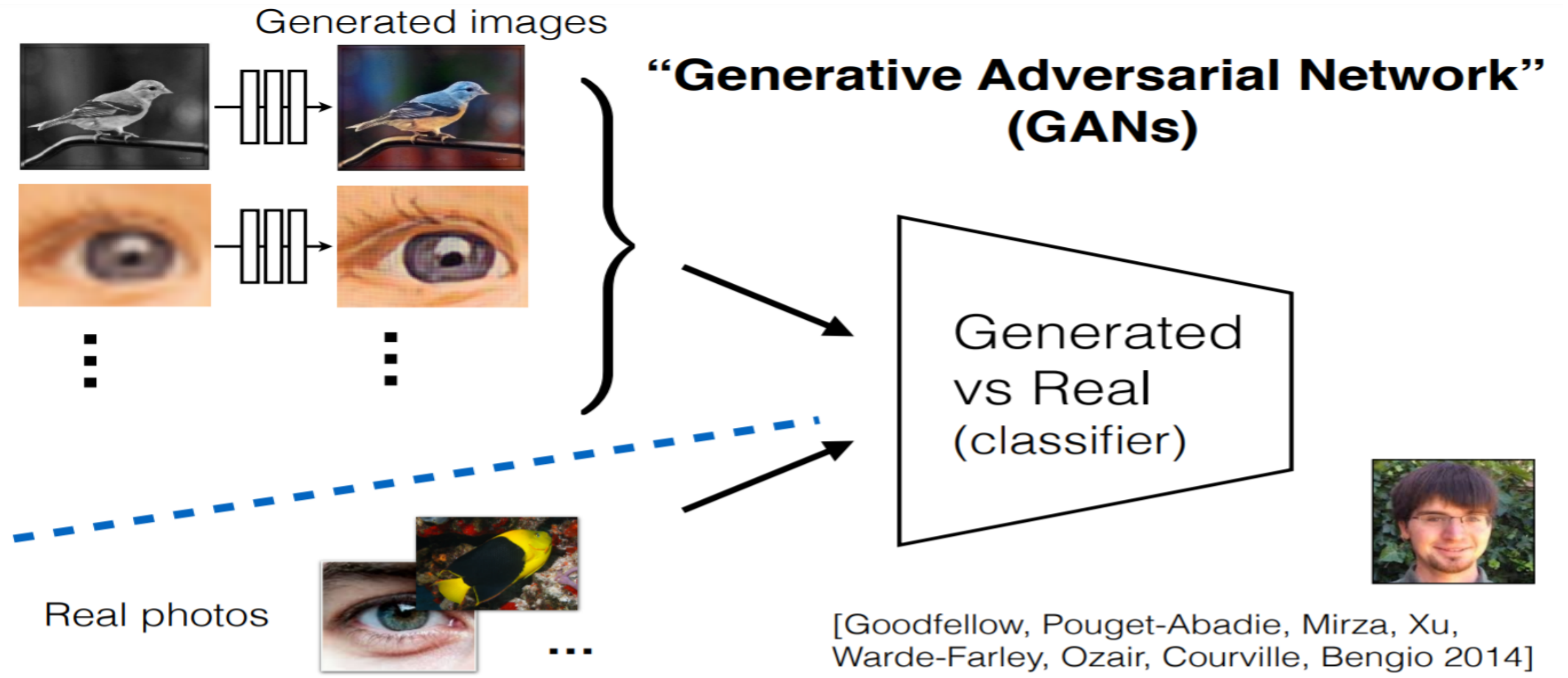
image source: https://technews.tw/2020/10/25/deepfake-deepnude/



Source: Stanford Internet Observatory

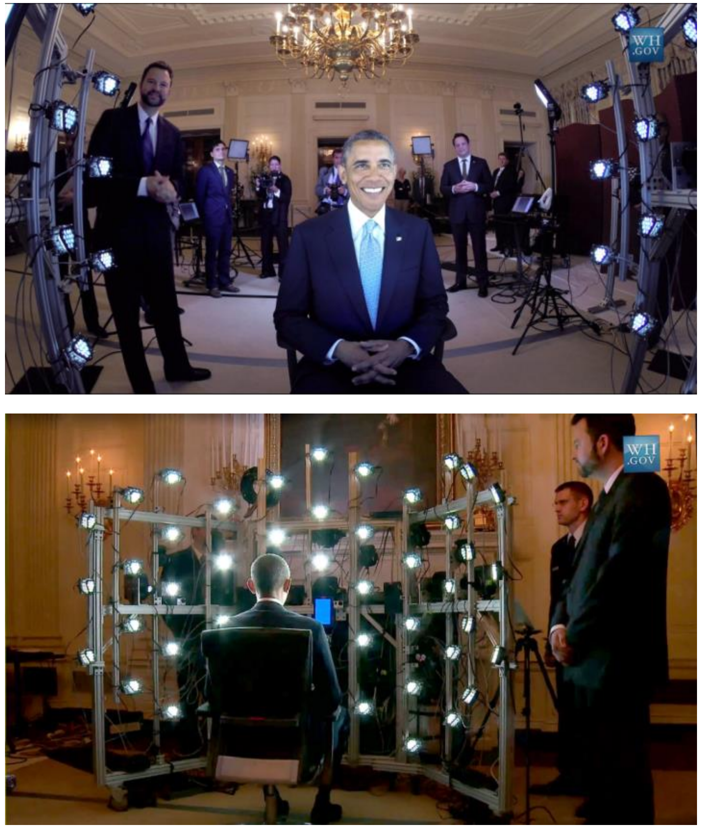
https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?fbclid=IwAR3_ubq-9niHCYj10LeqIBogoMG9ExSMjz7azLhMlteu2D6-C-shsJhAKUE

The Evolution of Content Editing

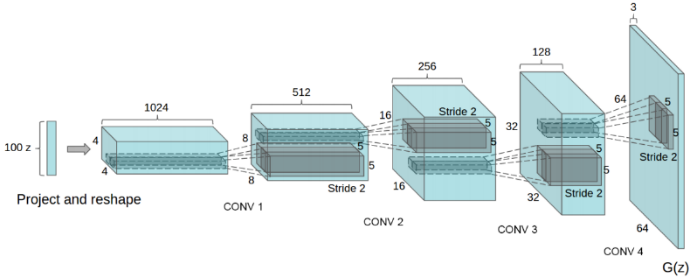


Slide Credit: Prof. Philip Isola, MIT

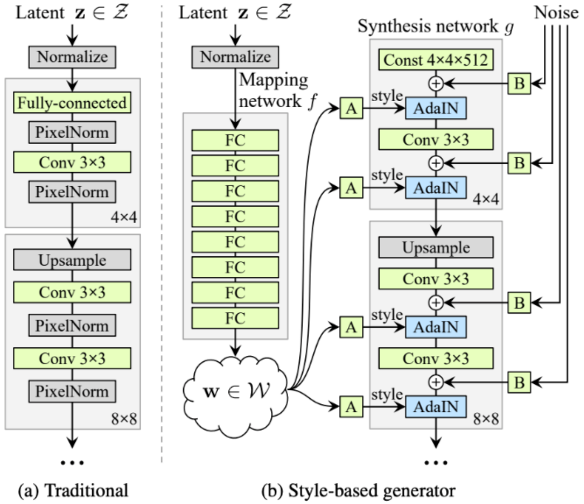
The Evolution of Content Editing



LightStage
[USC ICT 2015]

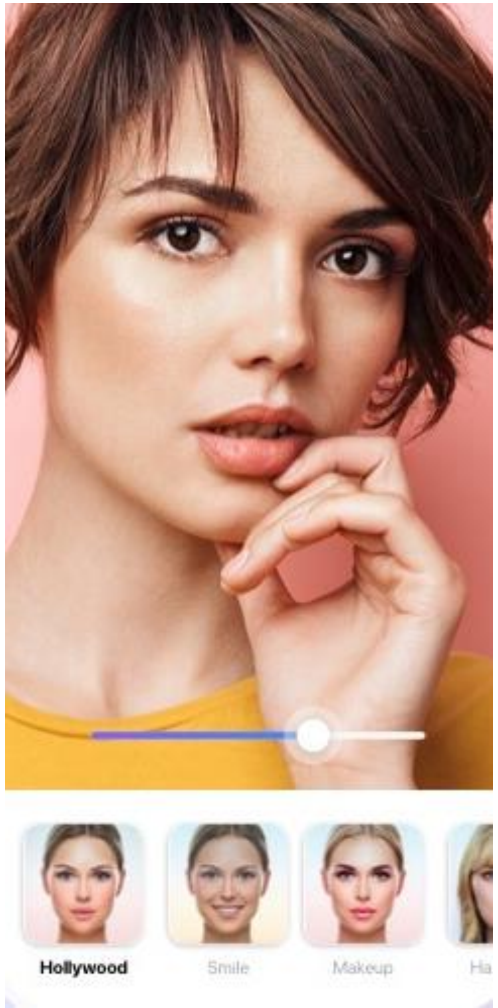


DCGAN
[Radford et al. 2016]



StyleGAN
[Karras et al. 2019]

The Evolution of Content Editing



FaceApp

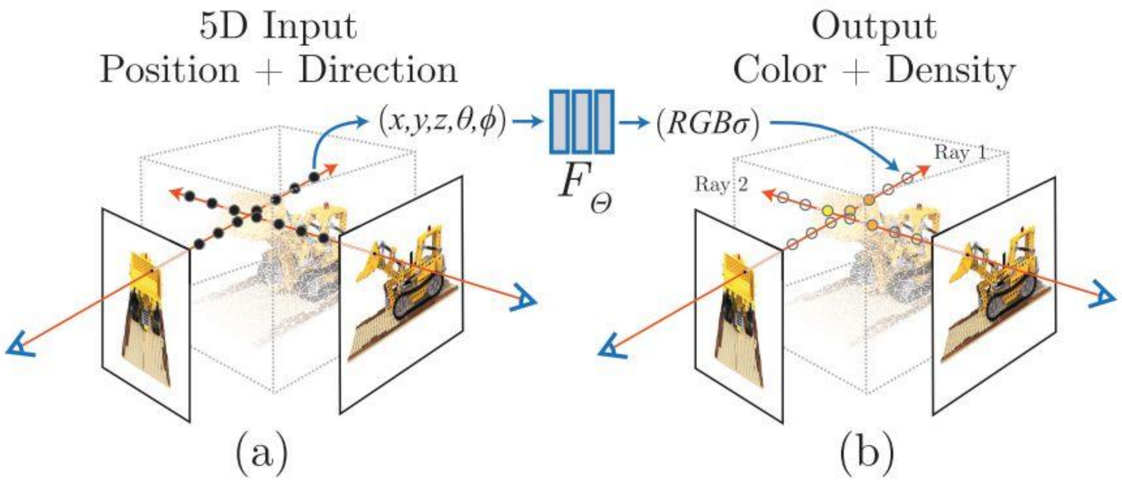
Faceswap

DeepFaceLab

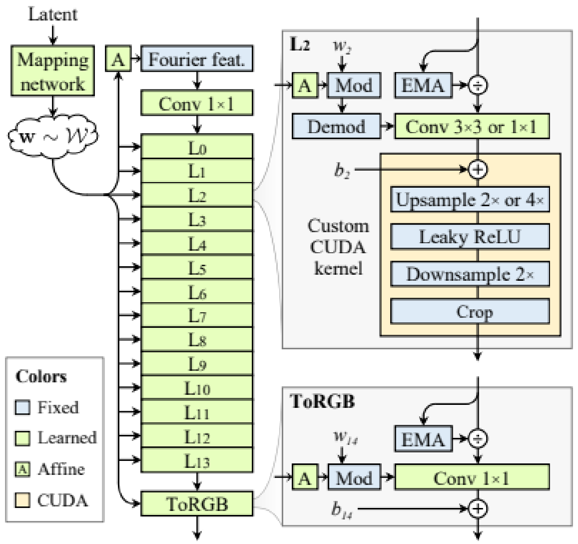
Video credit: *Chris Ume and Miles Fisher*

The Evolution of Content Editing

Face-Vid2Vid [Wang et al. 2021]



NeRF [Mildenhall et al. 2020]

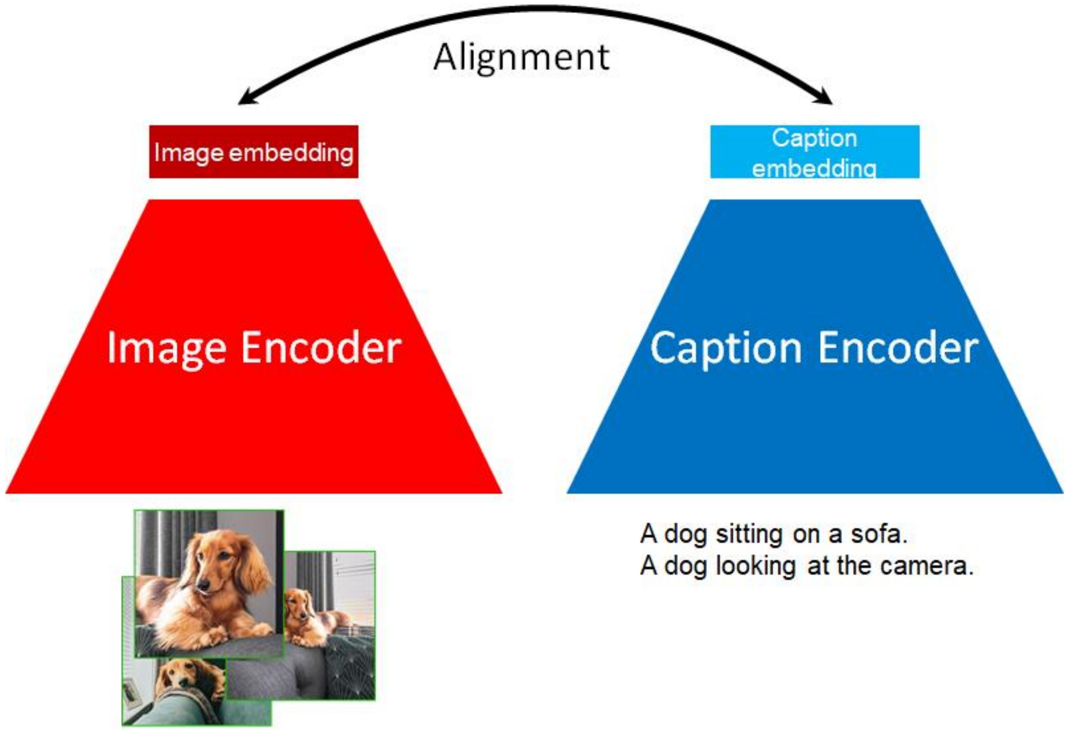


StyleGAN3
[Karras et al. 2021]

Image-Text Alignment

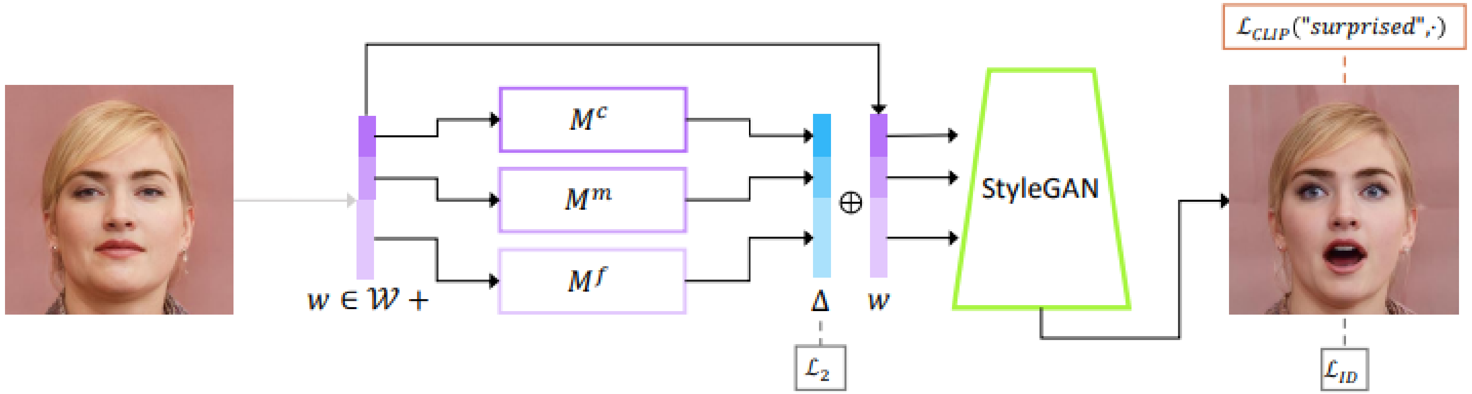
CLIP and ALIGN

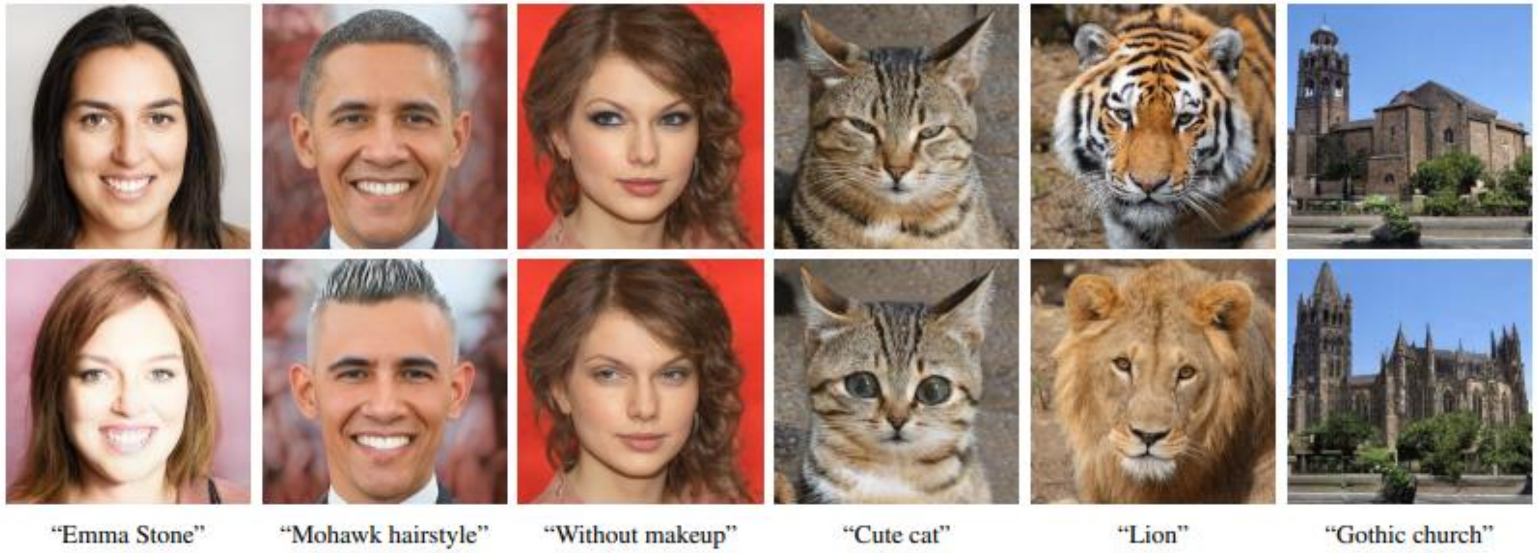
- Advantage of alignment training
  - Machine has more understanding of images through detailed captions
  - Embeddings share similar semantic are close in latent space
    - Good zero-shot performance



Radford et al. Learning Transferable Visual Models From Natural Language Supervision, OpenAI, 2021
Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, Google, 2021

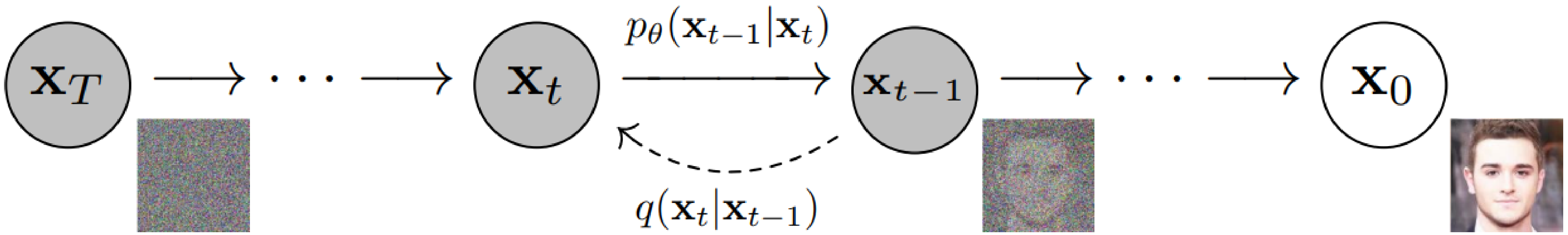
The Evolution of Content Editing





StyleCLIP [Patashnik et al. 2021]

Denoising Diffusion Probabilistic Model



$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}) \Rightarrow q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boxed{\mu_\theta(\mathbf{x}_t, t)}, \sigma_t^2 \mathbf{I}) \Rightarrow p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

Learned Model (U-Net)

Diffusion model:
1. Better mode coverage/diversity
2. Higher quality samples
3. Slower sampling

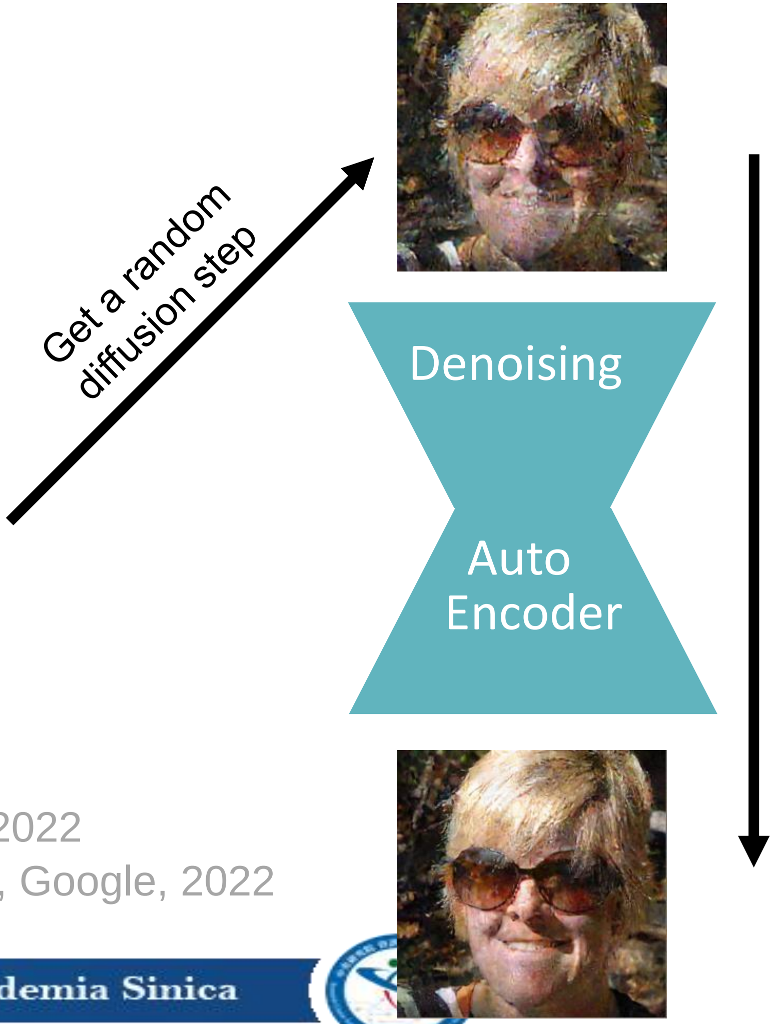
Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020): 6840-6851.

https://cvpr2022-tutorial-diffusion-models.github.io/

Text-to-image Models

DALL·E 2, Imagen, Stable Diffusion

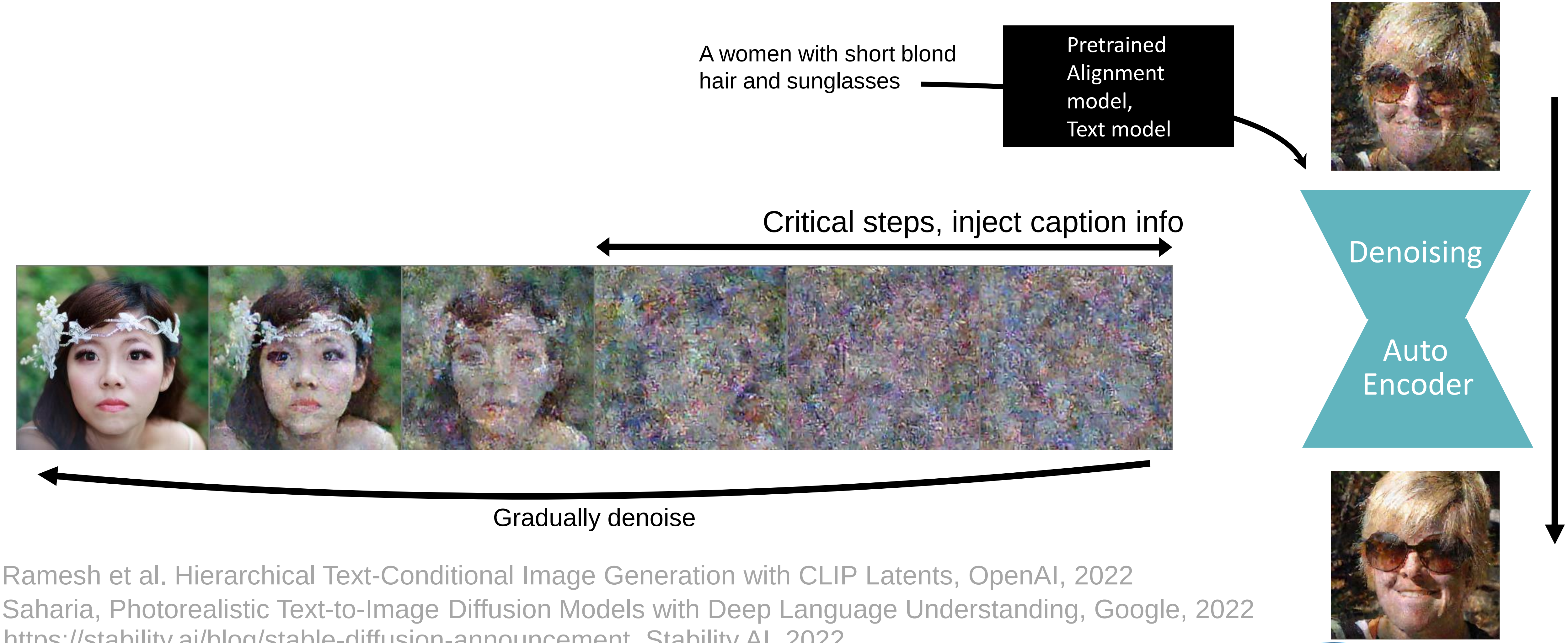
Gradually diffuse



Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, OpenAI, 2022
Saharia, Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, Google, 2022
https://stability.ai/blog/stable-diffusion-announcement, Stability AI, 2022

Text-to-image Models

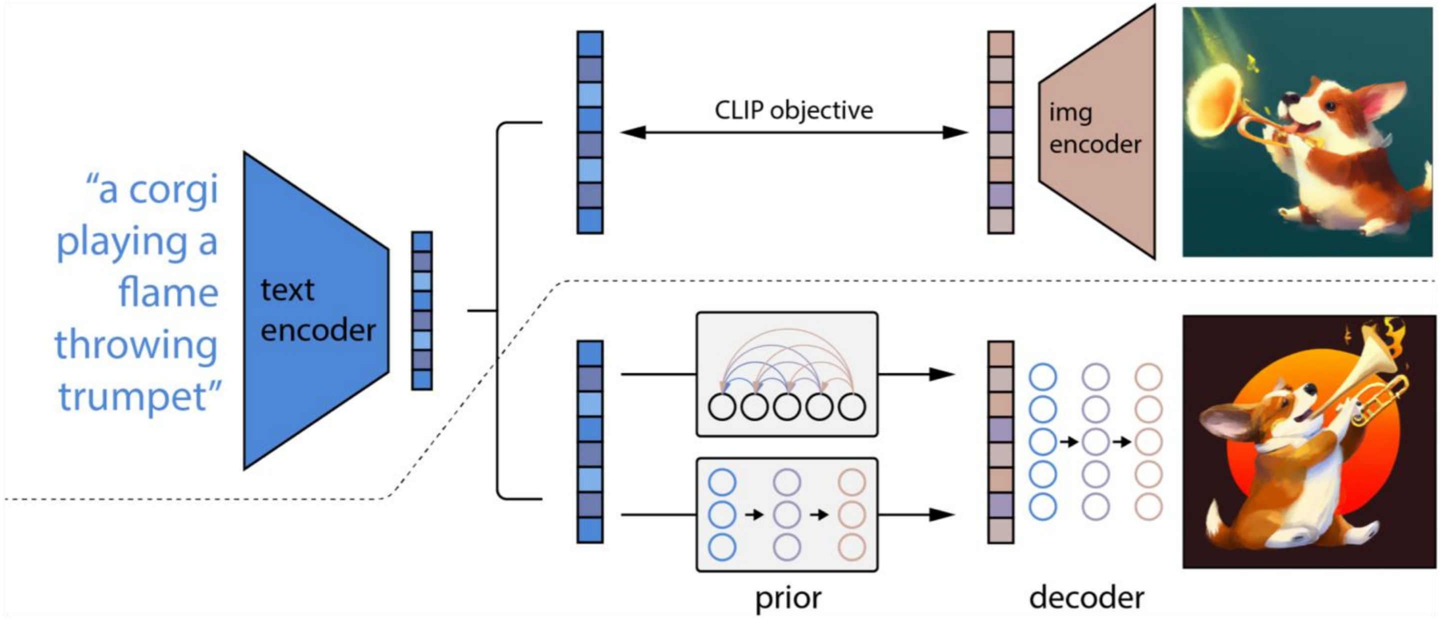
DALL·E 2, Imagen, Stable Diffusion

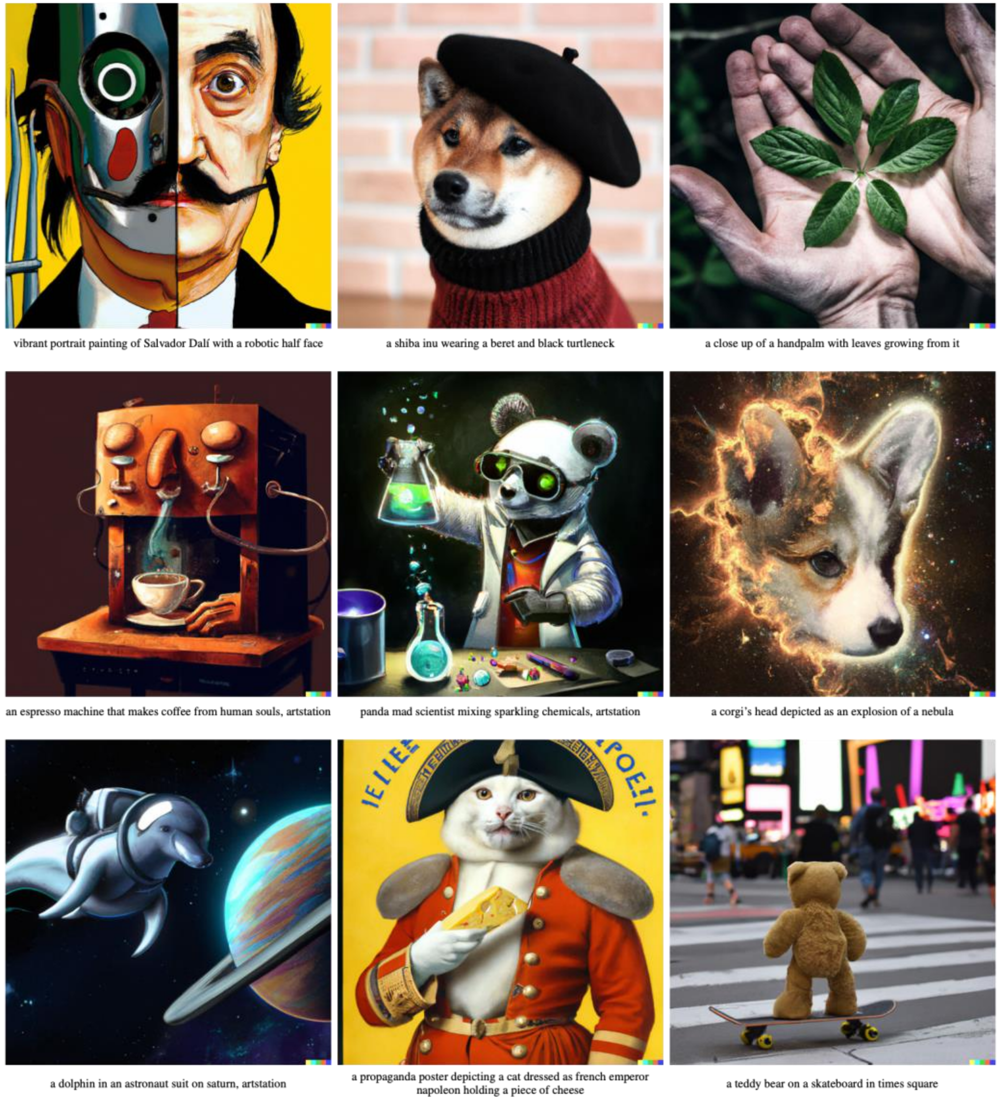


Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, OpenAI, 2022
Saharia, Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, Google, 2022
https://stability.ai/blog/stable-diffusion-announcement, Stability AI, 2022

DALL·E 2

From OpenAI

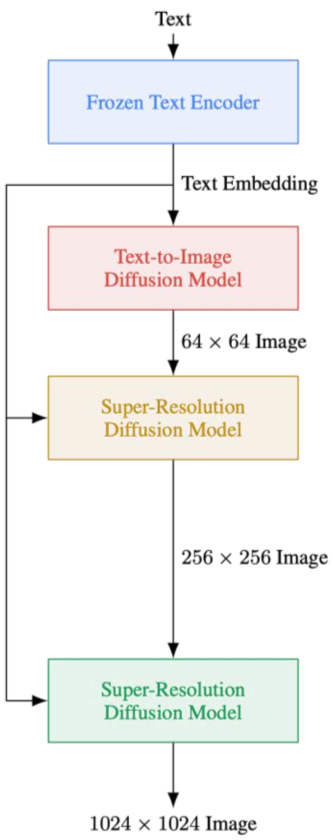


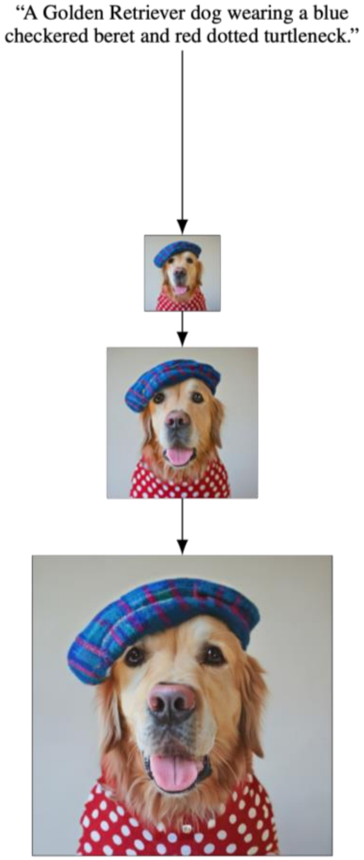


Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, OpenAI, 2022

Imagen

From Google





A chromeplated cat sculpture placed on a Persian rug.

Android Mascot made from bamboo.

Intricate origami of a fox and a unicorn in a snowy forest.

A transparent sculpture of a duck made out of glass.

A raccoon wearing cowboy hat and black leather jacket is behind the backyard window. Rain droplets on the window.

A bucket bag made of blue suede. The bag is decorated with intricate golden paisley patterns. The handle of the bag is made of rubies and pearls.

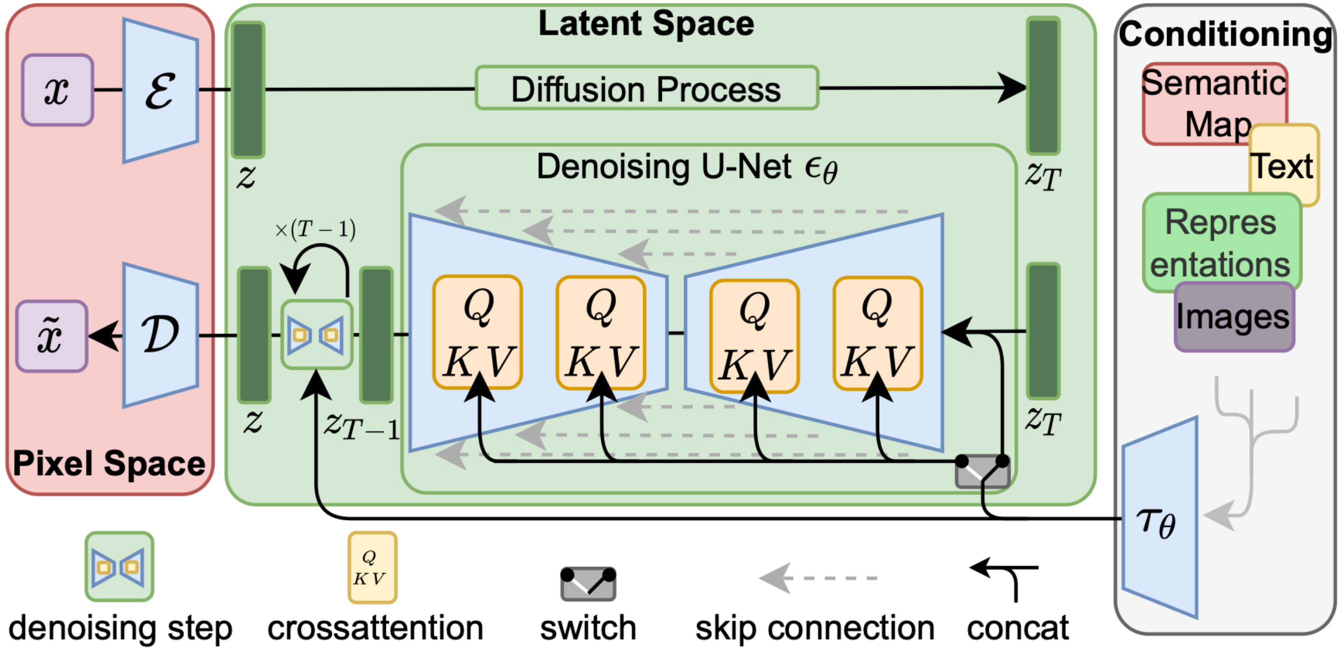
Three spheres made of glass falling into ocean. Water is splashing. Sun is setting.

Vines in the shape of text ’Imagen’ with flowers and butterflies bursting out of an old TV.

A strawberry splashing in the coffee in a mug under the starry sky.

Saharia et al., Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding, Google, 2022

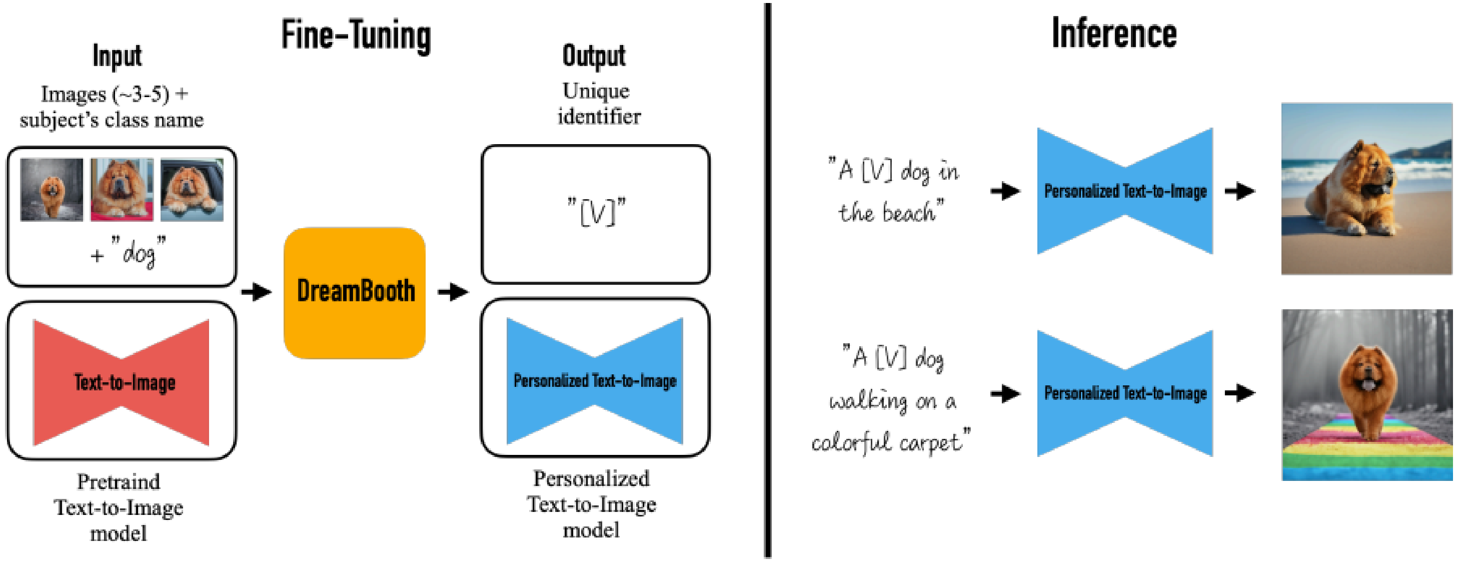
Stable Diffusion



Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684-10695. 2022.

https://stability.ai/blog/stable-diffusion-announcement, Stability AI, 2022

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation



Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *arXiv preprint arXiv:2208.12242* (2022).

Text-to-Video

Imagen Video (Google), Phenaki (Google), Make-a-Video (Meta)

A bunch of autumn leaves falling on a calm lake to form the text 'Imagen Video' Smooth

A clear wine glass with turquoise-colored waves inside it.

A giraffe underneath a microwave.

A panda taking a selfie

Video credit: https://imagen.research.google/video/

Ho, Jonathan, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma et al. "Imagen video: High definition video generation with diffusion models." *arXiv preprint arXiv:2210.02303* (2022).

Villegas, Ruben, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. "Phenaki: Variable length video generation from open domain textual description." *arXiv preprint arXiv:2210.02399* (2022).

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O. and Parikh, D., 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

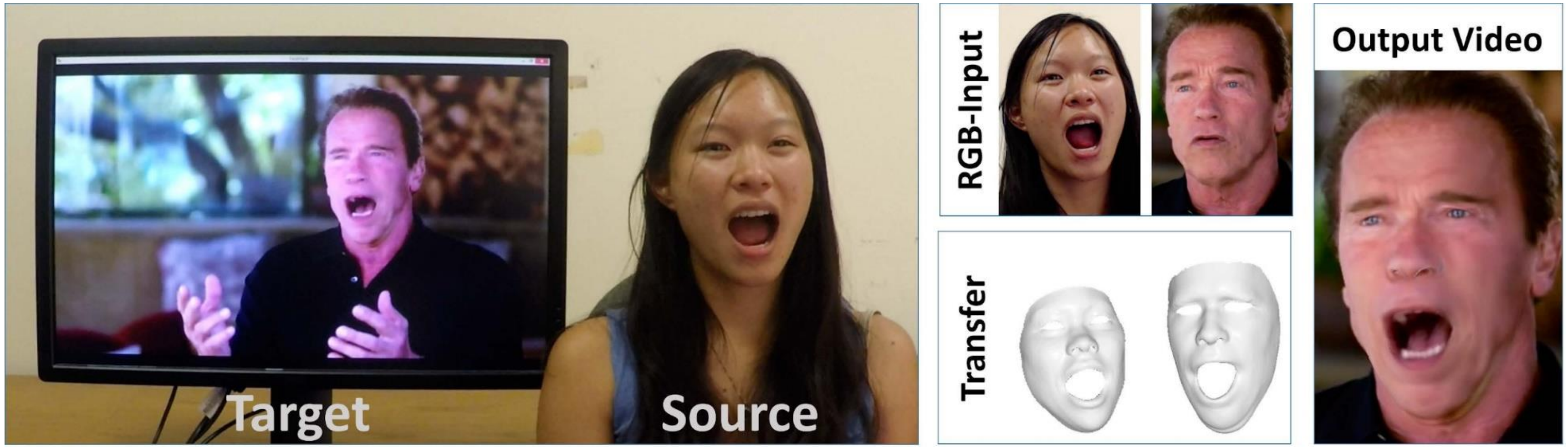
Challenges

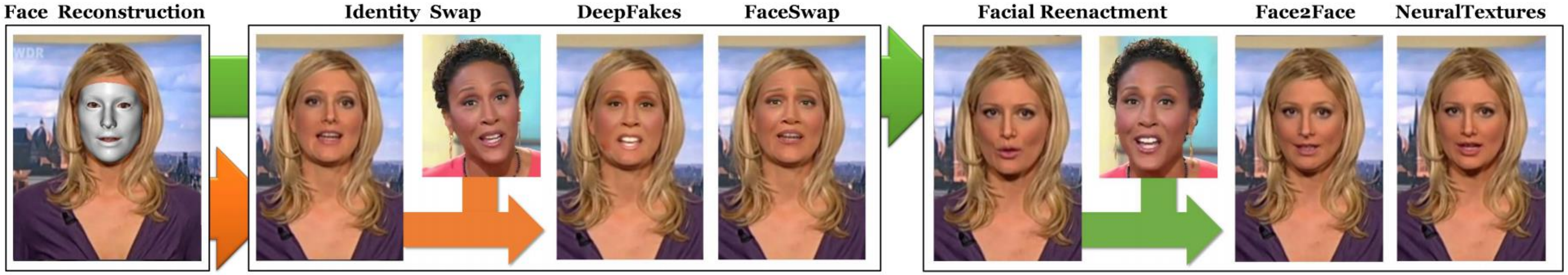
- The evolution of the deepfake technology is ongoing and upgrading in a very fast speed.
- The technologies are widely accessible to the public and much easier to use than before.

Possible Countermeasures

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

Common Deepfake Manipulation

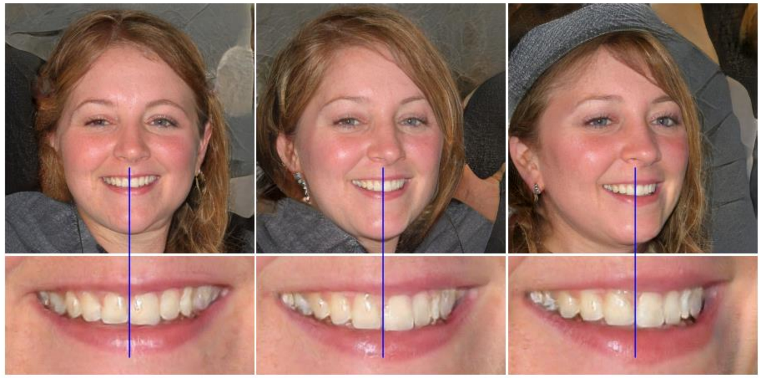




Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1-11).
Thies, Justus, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. "Face2face: Real-time face capture and reenactment of rgb videos." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387-2395. 2016.

Deepfake Detection

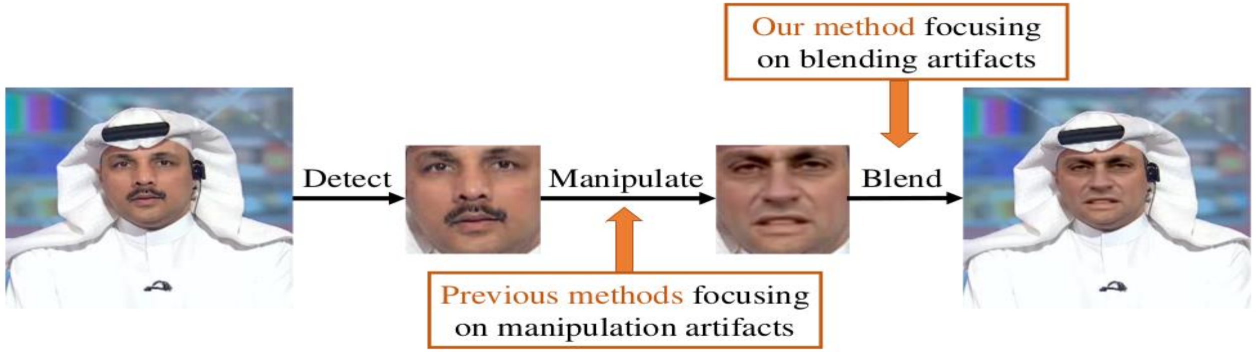
- Sample visual cues for detection

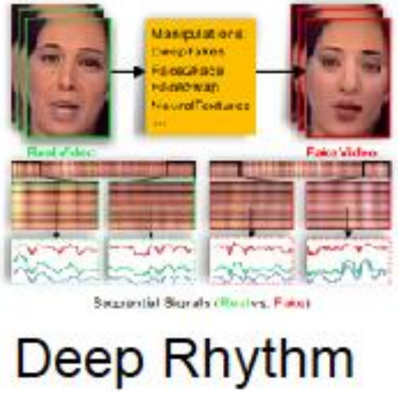


StyleGAN
[Karras et al. 2019]

[Hui et al. 2022]

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



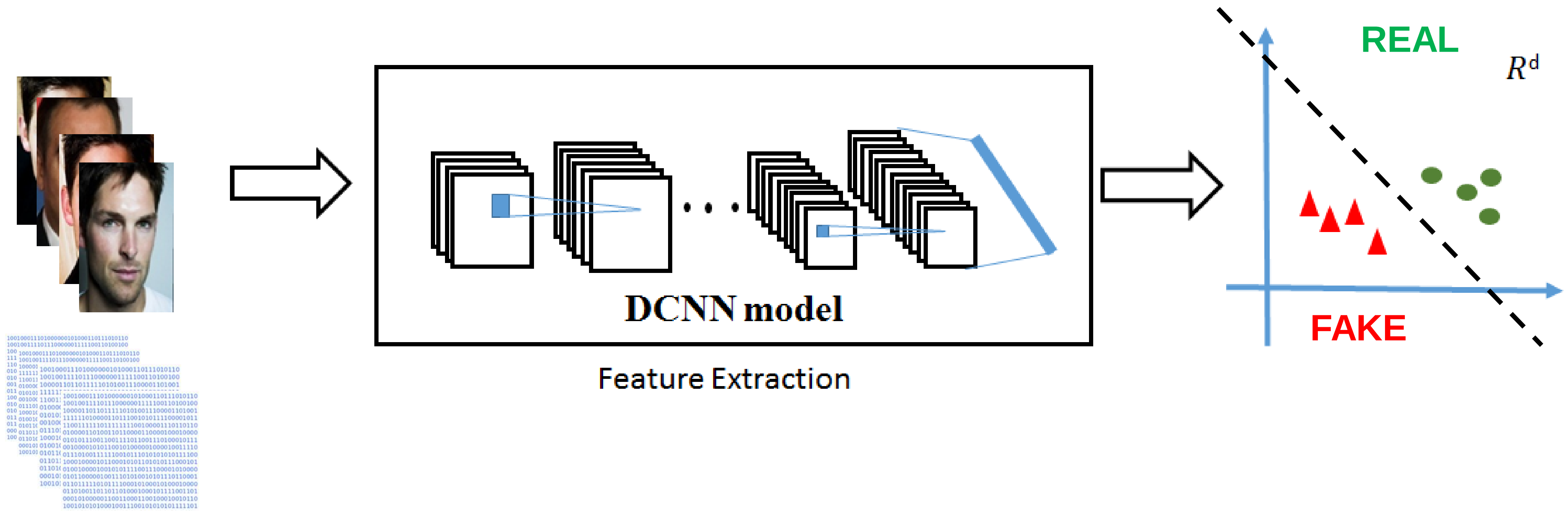


StyleGAN
[Karras et al. 2019]

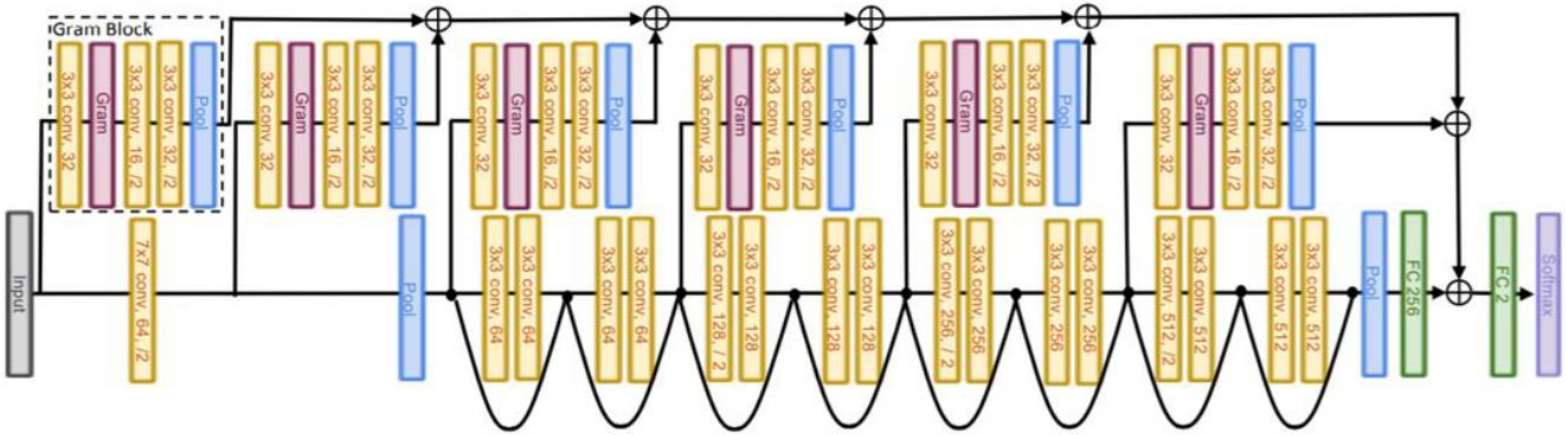
FaceXRay
[Li et al. 2020]

Deepfake Detection

- Train a binary classifier to distinguish real images from fake ones.



Global Texture Enhancement for Fake Face Detection In the Wild

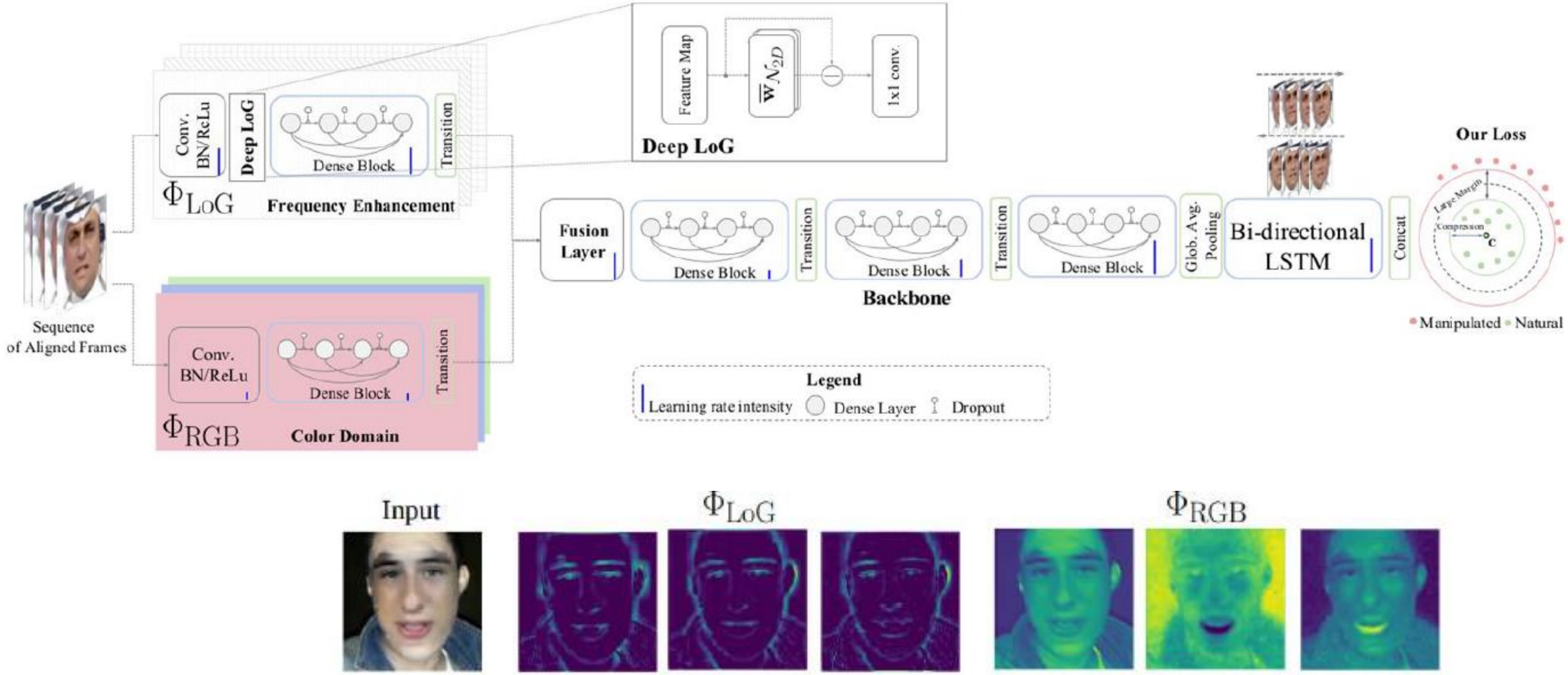


			512x512	64x64	512x512	64x64	kernel size 25	std 5	
Training set	Testing set	Method	Original %	8x ↓ %	JPEG %	JPEG 8x ↓	Blur %	Noise %	Avg.
StyleGAN vs. CelebA-HQ	StyleGAN vs. CelebA-HQ	Co-detect	79.93 ± 1.34	71.80 ± 1.30	74.58 ± 3.25	71.25 ±1.18	71.39 ±1.42	54.09 ± 2.45	70.51
		ResNet	96.73 ± 3.60	85.10 ± 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

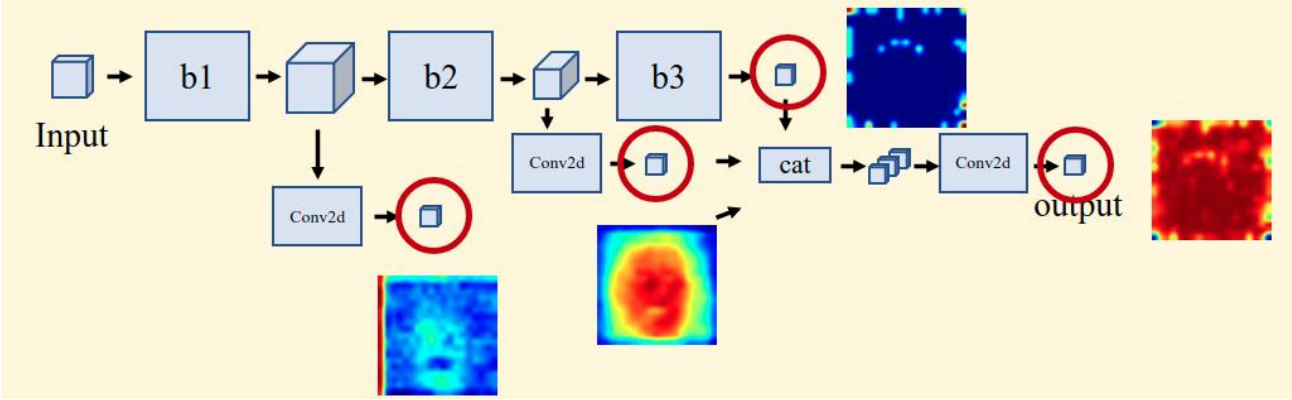


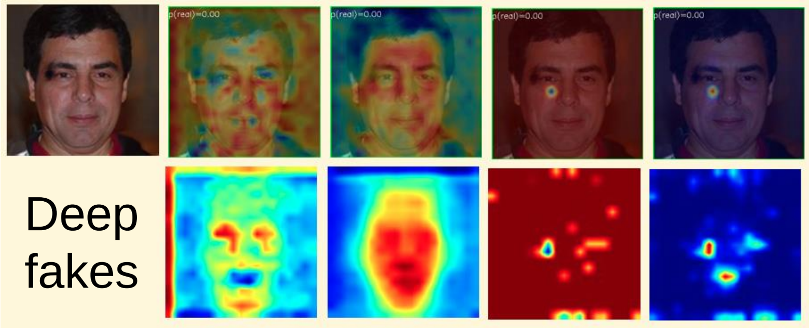
[Masi et al. 2020]

What makes fake images detectable? Understanding properties that generalize

- Handling 2D Deepfakes (Patch-Forensics)

Train\Test	DF	FF	FS	NT
Deepfake (DF)	0.990	0.698	0.524	0.738
Face2Face (FF)	0.627	0.991	0.547	0.964
FaceSwap (FS)	0.595	0.575	0.953	0.496
NeuralTextures (NT)	0.623	0.938	0.533	0.982

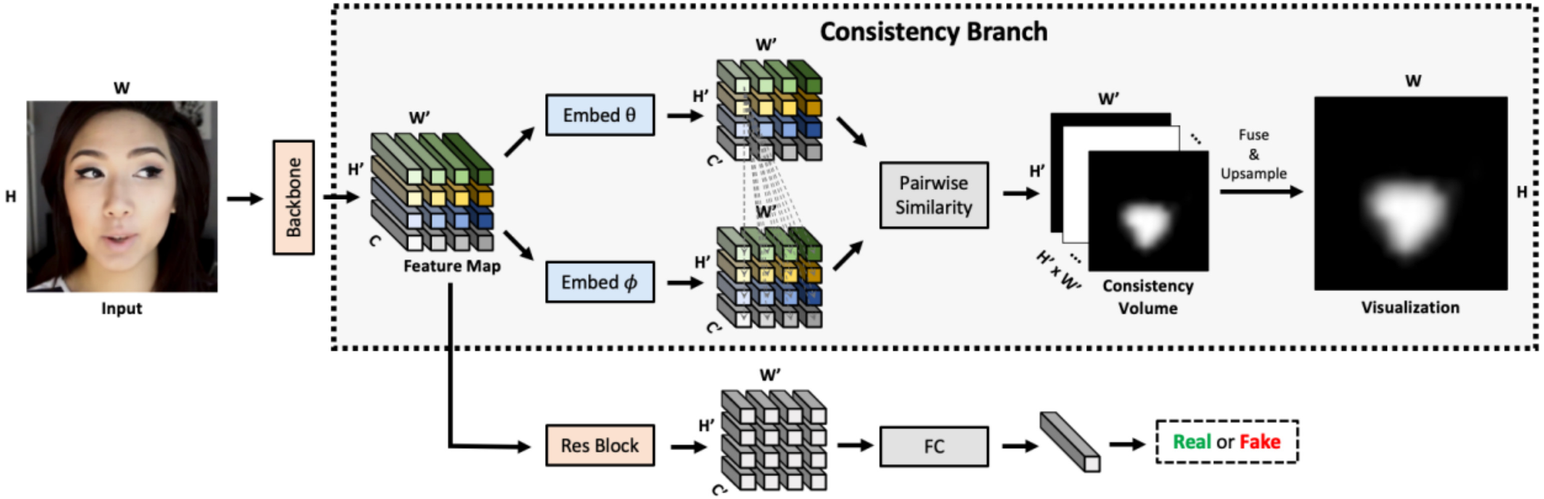


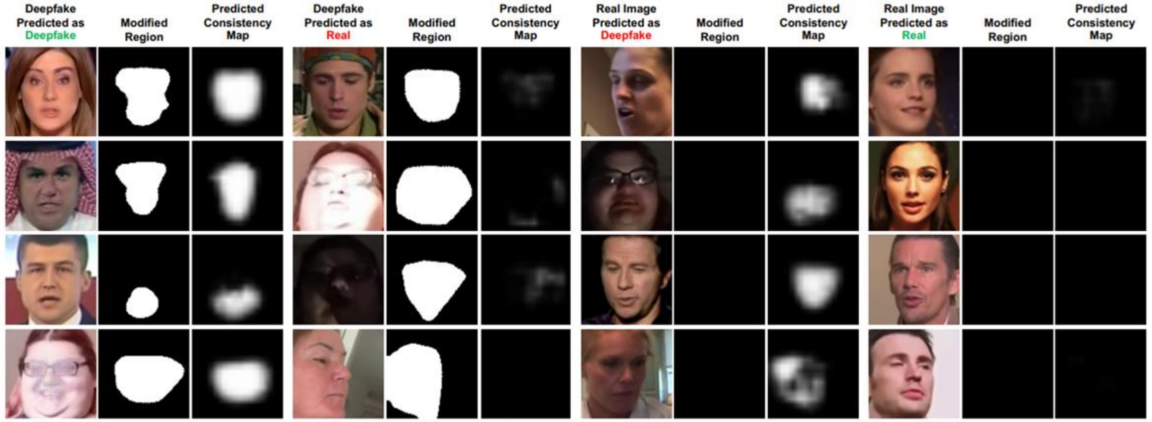


block 1 block 2 block 3 output

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.

Learning Self-Consistency for Deepfake Detection

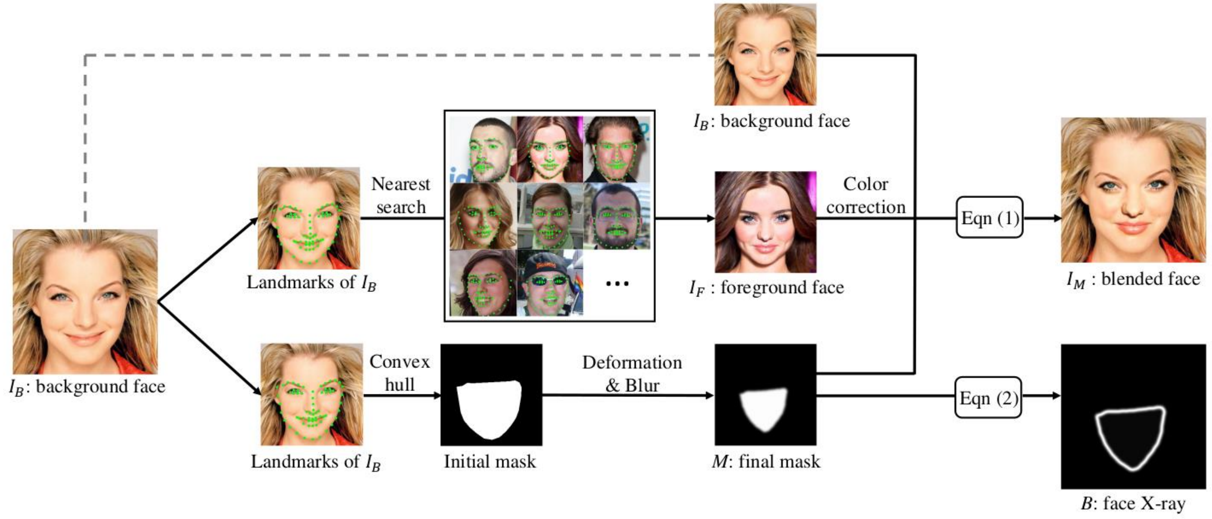


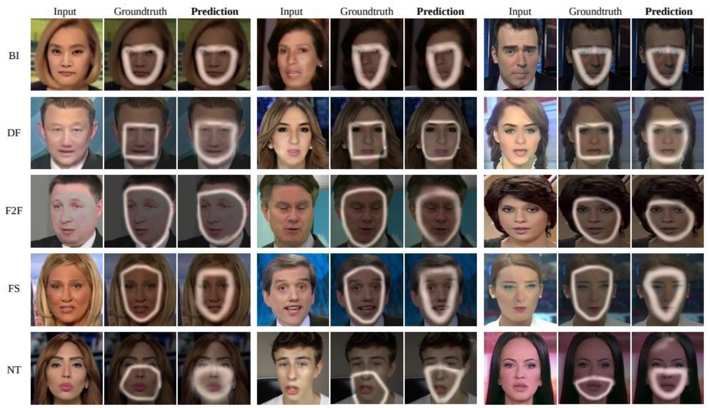


Method	Backbone	Train Set	Test Set (AUC (%))				
			DF	F2F	FS	NT	FF++
MIL [59]	Xception	FF++	99.51	98.59	94.86	97.96	97.73
Fakespotter [56]	ResNet-50	FF++, CD2, DFDC	-	-	-	-	98.50
XN-avg [45]	Xception	FF++	99.38	99.53	99.36	97.29	98.89
Face X-ray [25]	HRNet	FF++	99.12	99.31	99.09	99.27	99.20
S-MIL-T [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]

Face X-ray for More General Face Forgery Detection



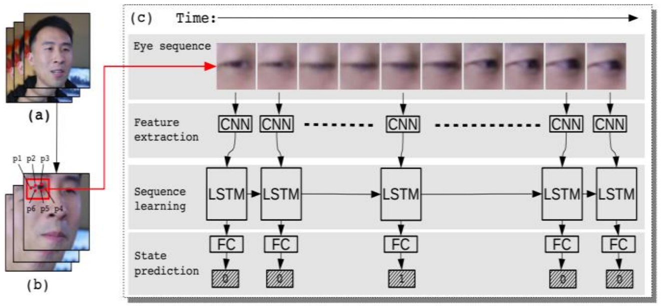


Model	Training set		Test set AUC				
	DF	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	99.38	75.05	49.13	80.39	76.34
HRNet	✓	–	99.26	68.25	39.15	71.39	69.51
Face X-ray	✓	–	99.17	94.14	75.34	93.85	90.62
	✓	✓	99.12	**97.64**	**98.00**	**97.77**	**97.97**
	F2F	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	87.56	99.53	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**

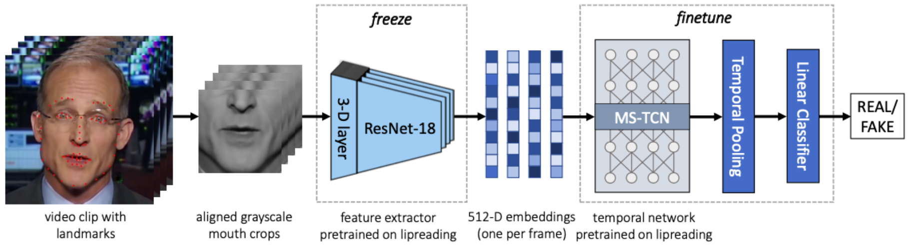
[Li et al. 2020]

Temporal Consistency(1/2)

- Video Inconsistency between frames

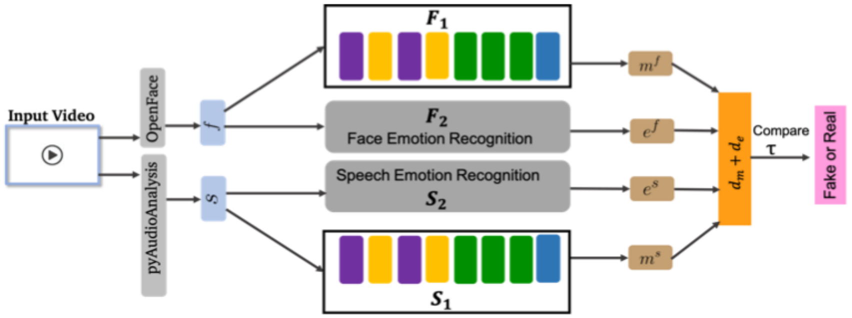


[In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, WIFS 2018]()

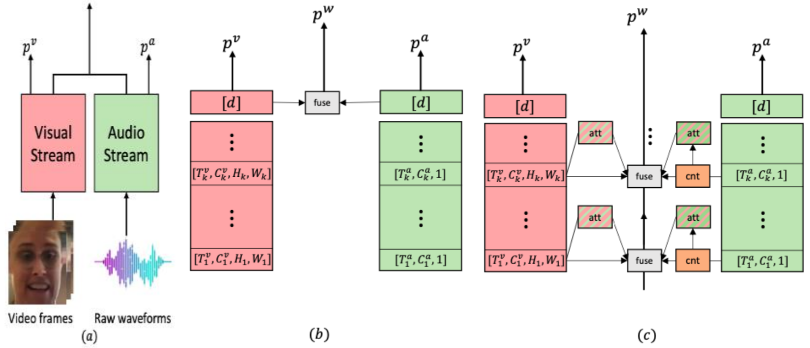


[Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection, CVPR 2021]()

- Audio-visual inconsistency

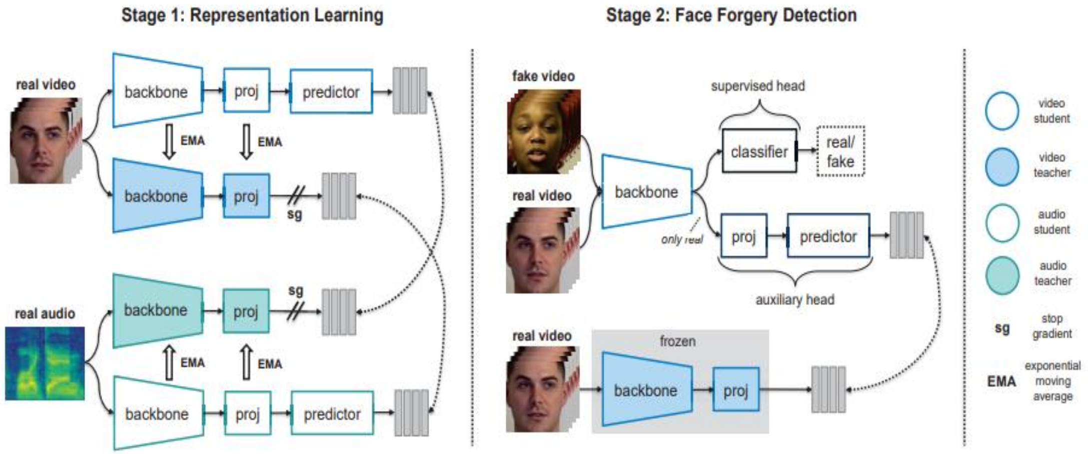


[Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues, ACMMM 2020]()

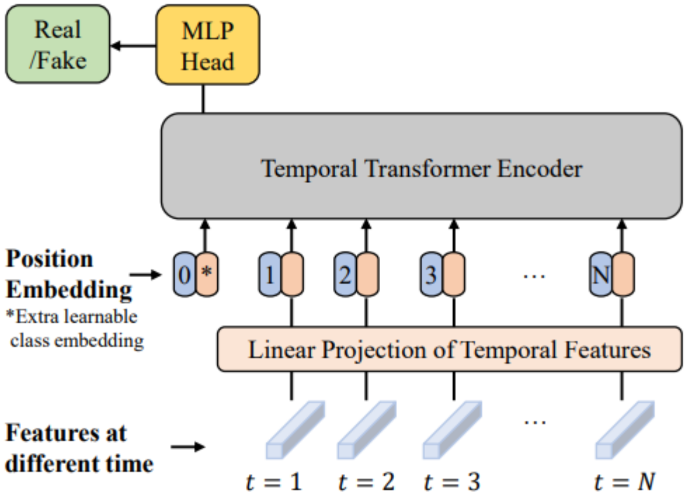


[Joint Audio-Visual Deepfake Detection, ICCV 2021]()

Temporal Consistency$^{(2/2)}$



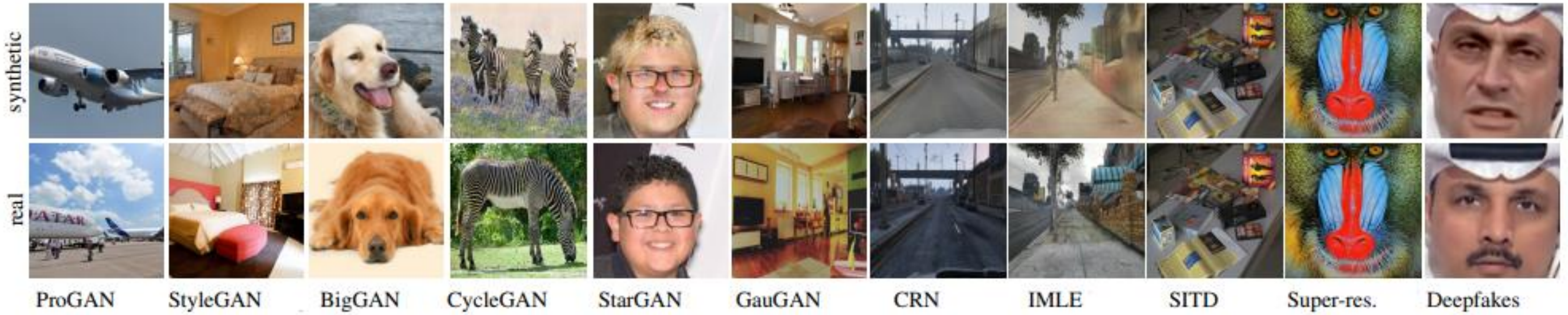
Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection. CVPR 2022



layer		output size
conv_1	$\mathbf{5}\times\mathbf{1}\times\mathbf{1}$, 64, stride 1, 1, 1	$64\times32\times224\times224$
pool_1	$1\times5\times5$ max, stride 1, 4, 4	$256\times32\times56\times56$
res_2	$\begin{bmatrix} 1\times1\times1, 64 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 64 \\ 1\times1\times1, 256 \end{bmatrix}\times3$	$256\times32\times56\times56$
pool_2	$2\times1\times1$ max, stride 2, 1, 1	$256\times16\times56\times56$
res_3	$\begin{bmatrix} 1\times1\times1, 128 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 128 \\ 1\times1\times1, 512 \end{bmatrix}\times4$	$512\times16\times28\times28$
res_4	$\begin{bmatrix} 1\times1\times1, 256 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 256 \\ 1\times1\times1, 1024 \end{bmatrix}\times6$	$1024\times16\times14\times14$
res_5	$\begin{bmatrix} 1\times1\times1, 512 \\ \mathbf{3}\times\mathbf{1}\times\mathbf{1}, 512 \\ 1\times1\times1, 2048 \end{bmatrix}\times3$	$2048\times16\times7\times7$
spatial-related average pool		$2048\times16\times1\times1$

Exploring Temporal Coherence for More General Video Face Forgery Detection, ICCV 2021

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



Family	Name	Training settings					Individual test generators											Total
		Train	Input	No. Class	Augments		Pro-GAN	Style-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	CRN	IMLE	SITD	SAN	Deep-Fake	**mAP**
					Blur	JPEG												
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, 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.

Towards the Detection of Diffusion Model Deepfakes(1/2)

AUROC / Pd@5% / Pd@1%	Wang et al. (2020)		Gragnaniello et al. (2021)		Mandelli et al. (2022a)
	Blur+JPEG (0.5)	Blur+JPEG (0.1)	ProGAN	StyleGAN2	
ProGAN	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	91.2 / 54.6 / 27.5
StyleGAN	98.7 / 93.7 / 81.4	99.0 / 95.5 / 84.4	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	89.6 / 43.6 / 14.7
ProjectedGAN	94.8 / 73.8 / 49.1	90.9 / 61.8 / 34.5	**100.0 / 99.9 / 99.3**	99.9 / 99.6 / 97.8	59.4 / 8.4 / 2.4
Diff-StyleGAN2	99.9 / 99.6 / 97.9	100.0 / 99.9 / 99.3	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	100.0 / 100.0 / 99.9
Diff-ProjectedGAN	93.8 / 69.5 / 43.3	88.8 / 54.6 / 27.2	**99.9 / 99.9 / 99.2**	99.8 / 99.6 / 96.6	62.1 / 10.5 / 2.8
Average	97.4 / 87.3 / 74.3	95.7 / 82.4 / 69.1	**100.0 / 100.0 / 99.7**	99.9 / 99.8 / 98.9	80.4 / 43.4 / 29.5
DDPM	85.2 / 37.8 / 14.2	80.8 / 29.6 / 9.3	**96.5 / 79.4 / 39.1**	95.1 / 69.5 / 30.7	57.4 / 3.8 / 0.6
IDDPM	81.6 / 30.6 / 10.6	79.9 / 27.6 / 7.8	**94.3 / 64.8 / 25.7**	92.8 / 58.0 / 21.2	62.9 / 7.0 / 1.3
ADM	68.3 / 13.2 / 3.4	68.8 / 14.1 / 4.0	**77.8 / 20.7 / 5.2**	70.6 / 13.0 / 2.5	60.5 / 8.2 / 1.8
PNDM	79.0 / 27.5 / 9.2	75.5 / 22.6 / 6.3	91.6 / 52.0 / 16.6	**91.5 / 53.9 / 22.2**	71.6 / 15.4 / 4.0
LDM	78.7 / 24.7 / 7.4	77.7 / 24.3 / 6.9	96.7 / 79.9 / 42.1	**97.0 / 81.8 / 48.9**	54.8 / 7.7 / 2.1
Average	78.6 / 26.8 / 9.0	76.6 / 23.7 / 6.8	**91.4 / 59.3 / 25.7**	89.4 / 55.2 / 25.1	61.4 / 8.4 / 2.0

Ricker, Jonas, Simon Damm, Thorsten Holz, and Asja Fischer. "Towards the Detection of Diffusion Model Deepfakes." *arXiv preprint arXiv:2210.14571* (2022)

Towards the Detection of Diffusion Model Deepfakes$^{(2/2)}$

Fine-tuned on (columns) / Tested on (rows)

Tested on \ Fine-tuned on	ProGAN	StyleGAN	ProjectedGAN	Diff-StyleGAN2	Diff-ProjectedGAN	DDPM	IDDPM	ADM	PNDM	LDM	GAN	DM	All
ProGAN	100.0	99.9	100.0	100.0	100.0	99.2	99.3	99.6	99.3	99.8	100.0	99.9	100.0
StyleGAN	98.5	100.0	98.0	99.9	98.9	95.5	98.2	98.1	96.8	97.8	100.0	99.9	100.0
ProjectedGAN	93.7	87.8	100.0	94.6	100.0	90.9	91.8	88.2	92.3	86.8	100.0	98.1	100.0
Diff-StyleGAN2	99.4	99.8	99.3	100.0	99.5	97.6	96.9	84.6	99.1	99.3	100.0	98.9	100.0
Diff-ProjectedGAN	93.2	86.4	100.0	93.9	100.0	91.0	91.8	89.5	90.3	91.0	100.0	98.6	100.0
DDPM	82.4	77.5	81.8	88.0	82.6	100.0	99.9	99.7	99.1	98.2	92.4	100.0	100.0
IDDPM	80.5	77.9	79.7	82.7	81.7	99.8	100.0	99.9	97.9	97.9	89.6	100.0	100.0
ADM	65.7	61.3	65.2	60.3	67.9	93.3	97.4	99.9	85.9	93.0	69.5	100.0	100.0
PNDM	77.5	74.3	79.9	83.1	81.8	99.8	99.9	99.8	100.0	99.2	93.3	100.0	100.0
LDM	76.7	68.1	74.1	77.6	78.0	97.6	99.2	99.5	94.6	100.0	87.9	100.0	100.0

(a) AUROC

Fine-tuned on (columns) / Tested on (rows)

Tested on \ Fine-tuned on	ProGAN	StyleGAN	ProjectedGAN	Diff-StyleGAN2	Diff-ProjectedGAN	DDPM	IDDPM	ADM	PNDM	LDM	GAN	DM	All
ProGAN	100.0	98.9	99.4	100.0	99.8	95.9	96.5	97.6	94.8	98.0	100.0	99.4	100.0
StyleGAN	78.5	100.0	70.1	98.8	83.0	60.7	81.8	85.5	67.8	73.6	100.0	98.4	100.0
ProjectedGAN	40.2	20.0	100.0	55.2	100.0	43.8	51.7	48.8	40.2	37.6	100.0	86.2	100.0
Diff-StyleGAN2	88.9	96.1	88.9	100.0	92.9	75.2	72.2	39.1	89.5	92.0	100.0	89.7	100.0
Diff-ProjectedGAN	39.8	17.9	99.5	49.2	100.0	42.0	50.0	47.4	31.0	45.1	100.0	88.9	100.0
DDPM	10.9	7.1	9.3	19.9	10.0	99.7	99.0	93.5	81.8	72.9	30.3	100.0	100.0
IDDPM	8.4	7.0	7.4	12.2	9.2	96.1	99.3	98.9	64.5	69.2	22.7	100.0	100.0
ADM	2.4	1.8	2.4	1.7	3.2	42.0	65.9	97.7	16.5	38.6	4.0	99.7	99.9
PNDM	6.6	6.3	7.4	14.4	10.9	96.7	98.7	96.6	100.0	89.1	39.4	100.0	100.0
LDM	5.7	2.9	4.4	7.8	6.6	68.4	86.2	92.7	40.8	100.0	20.6	100.0	100.0

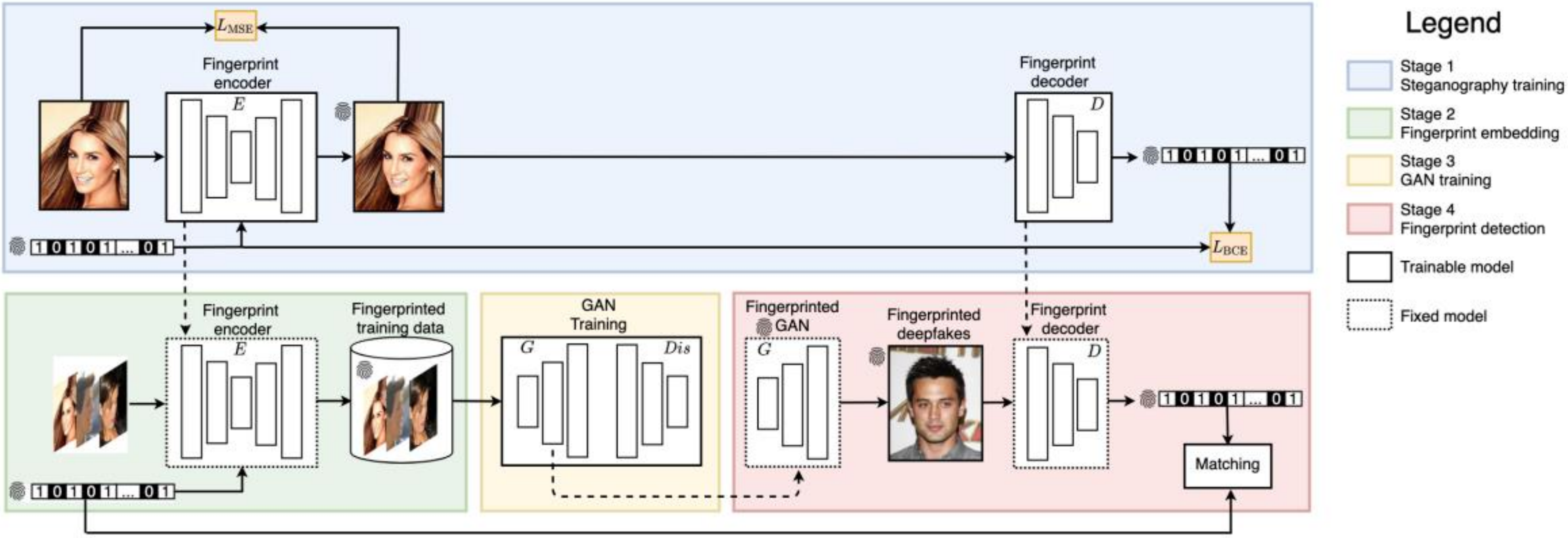
(b) PD@1%.

Ricker, Jonas, Simon Damm, Thorsten Holz, and Asja Fischer. "Towards the Detection of Diffusion Model Deepfakes." *arXiv preprint arXiv:2210.14571* (2022)

Possible Countermeasures

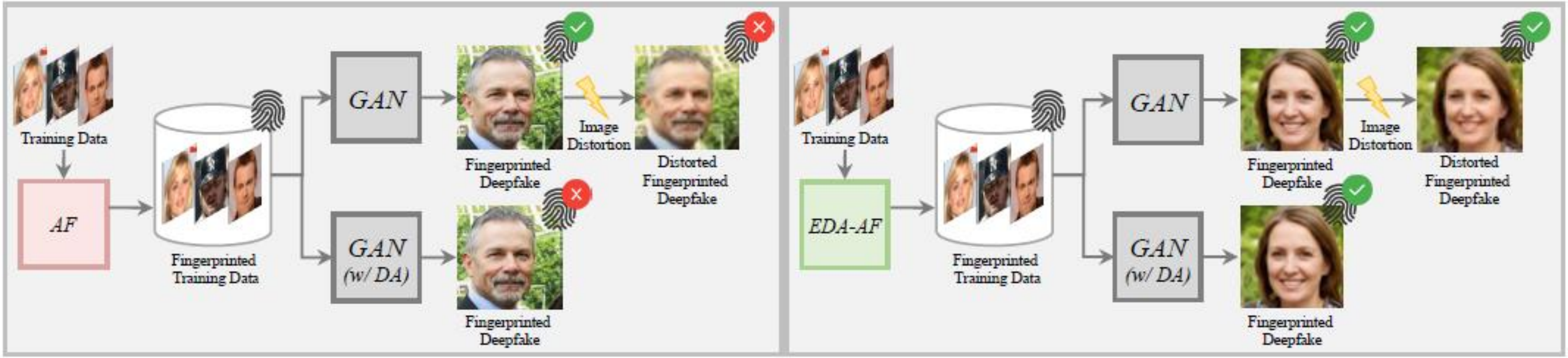
- Passive Defense
  - Deepfake Detection
  - Digital Watermark
- Proactive Defense
  - Adversarial Attack

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

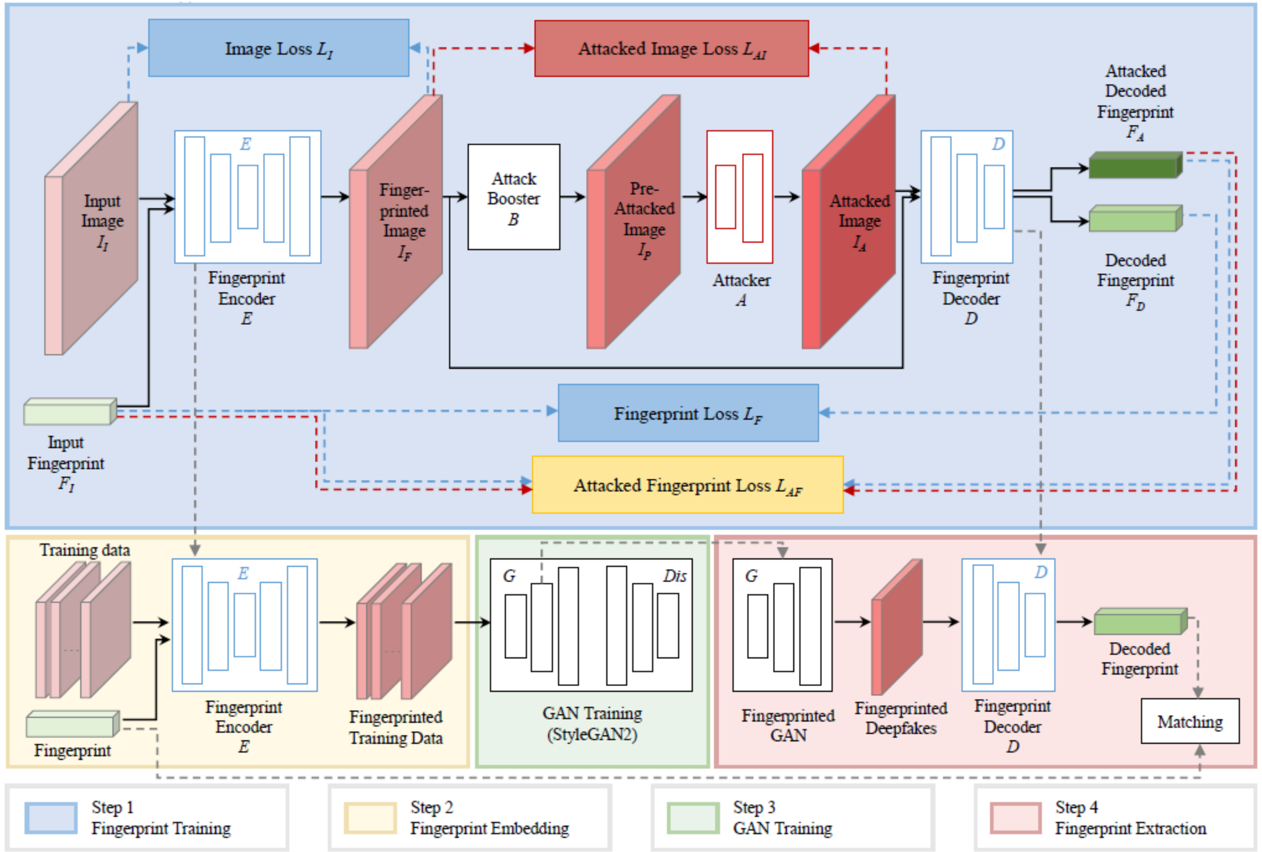


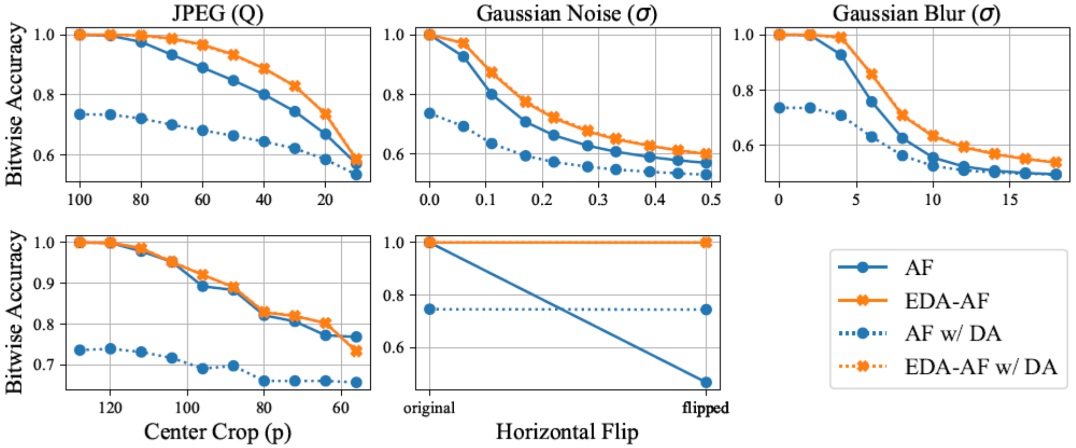
Yu, Ning, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14448-14457. 2021.

Enhancing the Robustness of Deep Learning Based Fingerprinting to Improve Deepfake Attribution (1/2)



Enhancing the Robustness of Deep Learning Based Fingerprinting to Improve Deepfake Attribution (2/2)





Method	Bit Acc ↑	SSIM ↑	PSNR ↑
AF [1]	0.74	**0.967**	**36.938**
EDA-AF	**0.99**	0.961	35.577

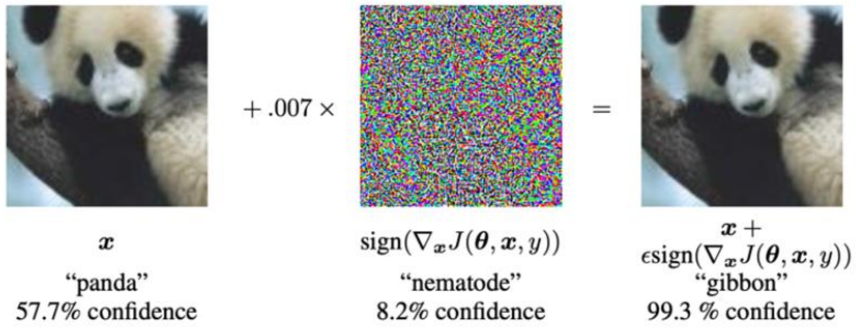
Dataset	Model	Method	Bit Acc ↑	FID ↓
CelebA	StyleGAN2	AF [1]	0.74	**7.60**
		EDA-AF	**0.99**	10.73
	StyleGAN3	AF [1]	0.73	**7.15**
		EDA-AF	**0.99**	9.32

Chieh-Yin Liao, Chen-Hsiu Huang, Jun-Cheng Chen, Ja-Ling Wu, “Enhancing the Robustness of Deep Learning Based Fingerprinting to Improve Deepfake Attribution,” in ACM Multimedia Asia conference (MMAsia), 2022.

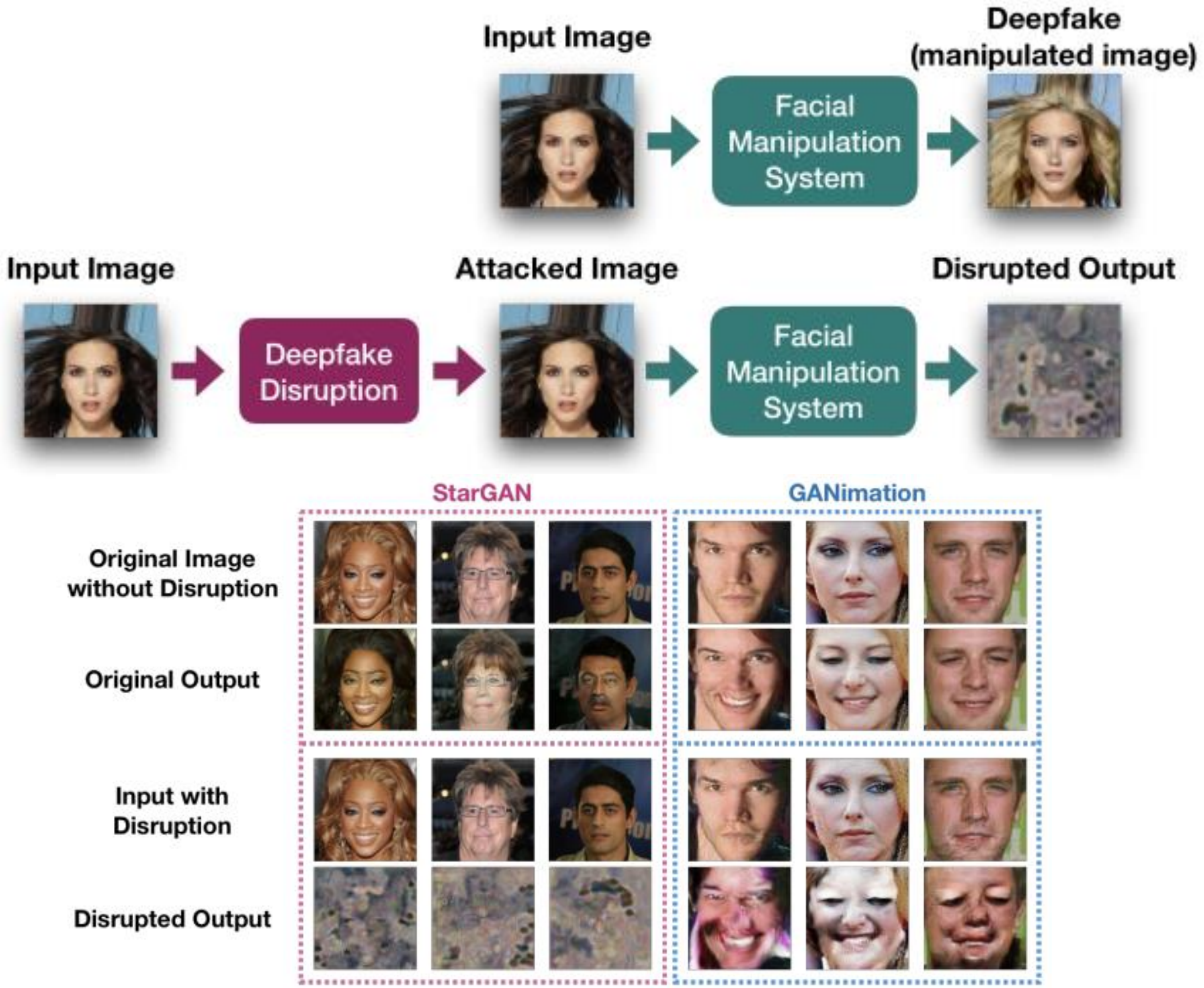
Possible Countermeasures

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

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

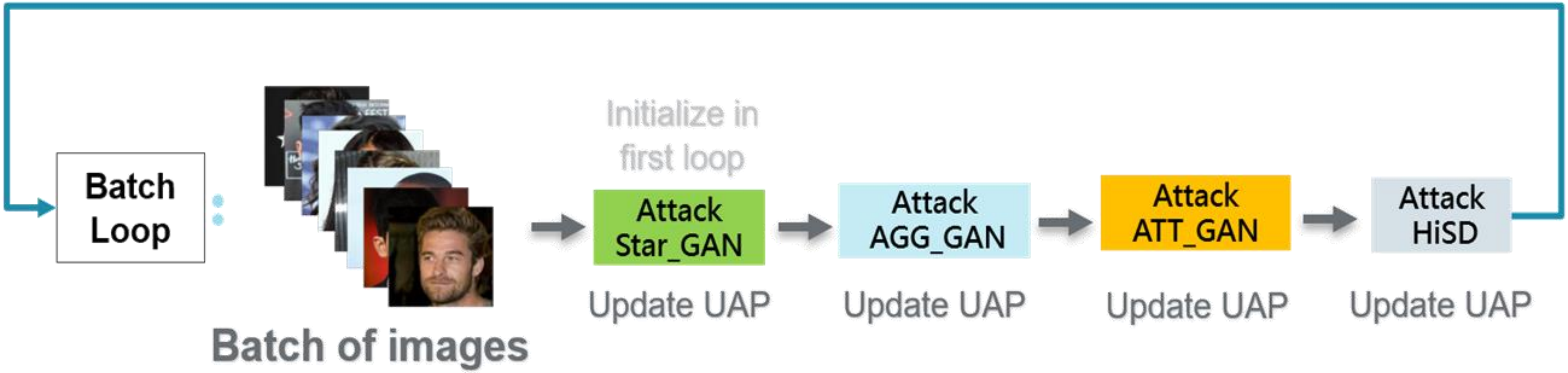


[Goodfellow et al. 2015]



[Ruiz et al. 2020]

Cross-Model Universal Adversarial Watermark



Hao Huang, Yongtao Wang, Zhaoyu Chen, Yuze Zhang, Yuheng Li, Zhi Tang, Wei Chu, Jingdong Chen, Weisi Lin, and Kai-Kuang Ma. "Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 989-997. 2022.

The Proposed Cross-Model Universal Adversarial Watermark

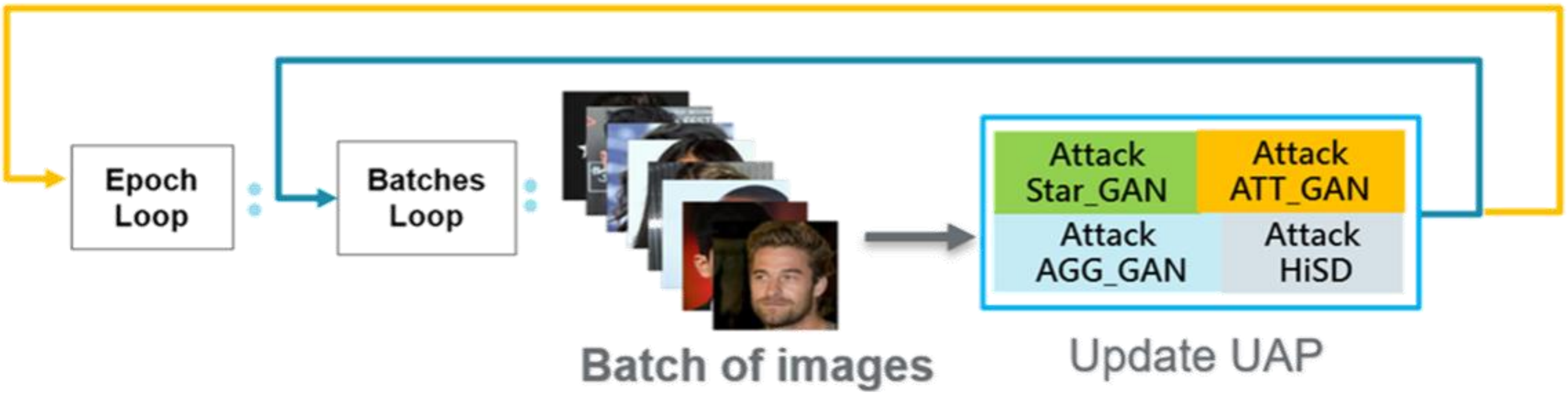


TABLE I: The evaluation results of cross-model UAP.

Attacks \ DGMs	StarGAN ASR	StarGAN ℓ_2	AggGAN ASR	AggGAN ℓ_2	ATTGAN ASR	ATTGAN ℓ_2	HiSD ASR	HiSD ℓ_2
CMUA_v1	**100.0%**	0.457	99.2%	0.107	20.2%	0.037	N/A[1]	N/A[1]
CMUA_v2	**100.0%**	0.199	**100.0%**	0.128	95.3%	0.066	**100.0%**	0.108
CMUA_12	97.0%	0.081	**100.0%**	0.046	86.1%	0.047	**100.0%**	0.107
Ours	**100.0%**	**0.766**	**100.0%**	**0.133**	**98.2%**	**0.124**	**100.0%**	**0.113**

TABLE II: The results of visual quality for perturbed images.

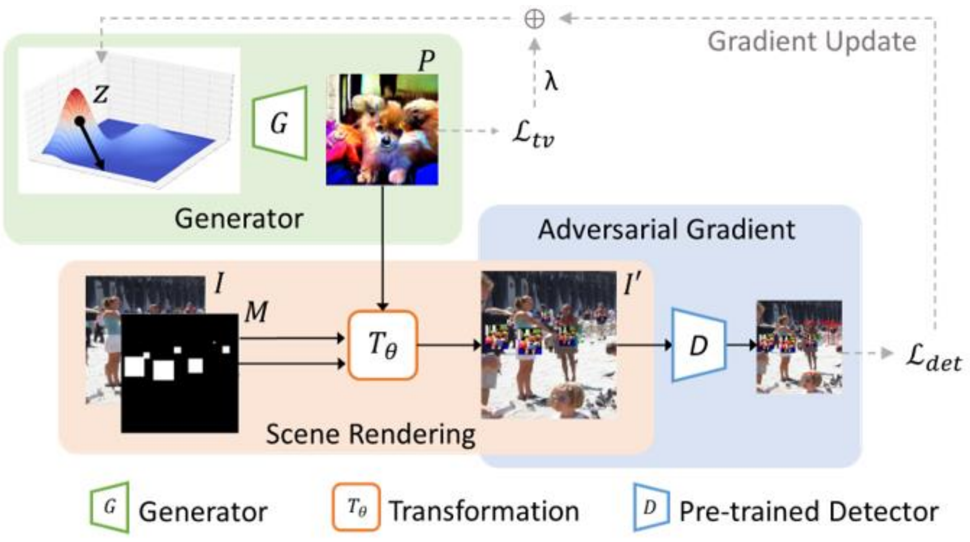
	PSNR↑	SSIM↑	Time	Batch size
CMUA_v2	32.4939	0.7821	N/A[2]	64
CMUA_12	32.2755	0.7736	800s	12
Ours	**33.4651**	**0.8186**	170s	8

Shuo-Yen Lin, Jun-Cheng Chen, Jia-Ching Wang, "A Comparative Study of Cross-Model Universal Adversarial Perturbation for Face Forgery," in IEEE International Conference on Visual Communications and Image Processing (VCIP), 2022.

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.

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

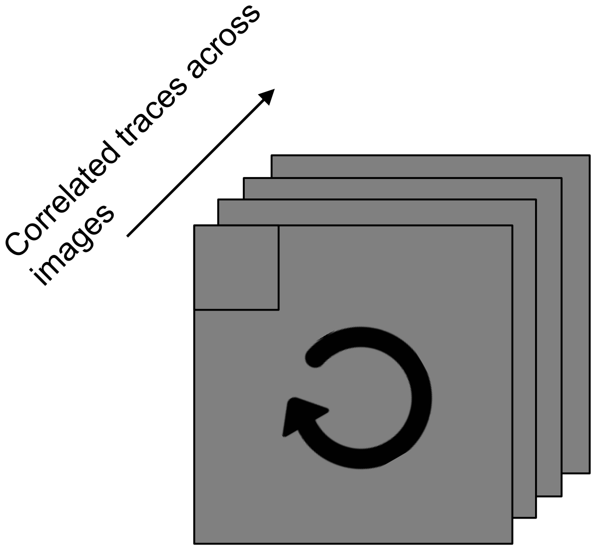
†trained on YOLOv2+YOLOv3+YOLOv4tiny * trained on YOLO ** trained on FasterRCNN

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.

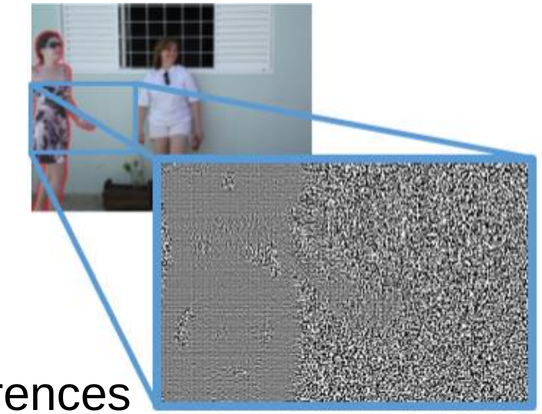
Traces in Images Allow us to Detect Forgery

Correlated traces across images
- Photo-response non uniformity noise (PRNU)

Correlated traces within images (usually periodic)
- Compression (e.g. blocking)
- Resampling
- Demosaicing

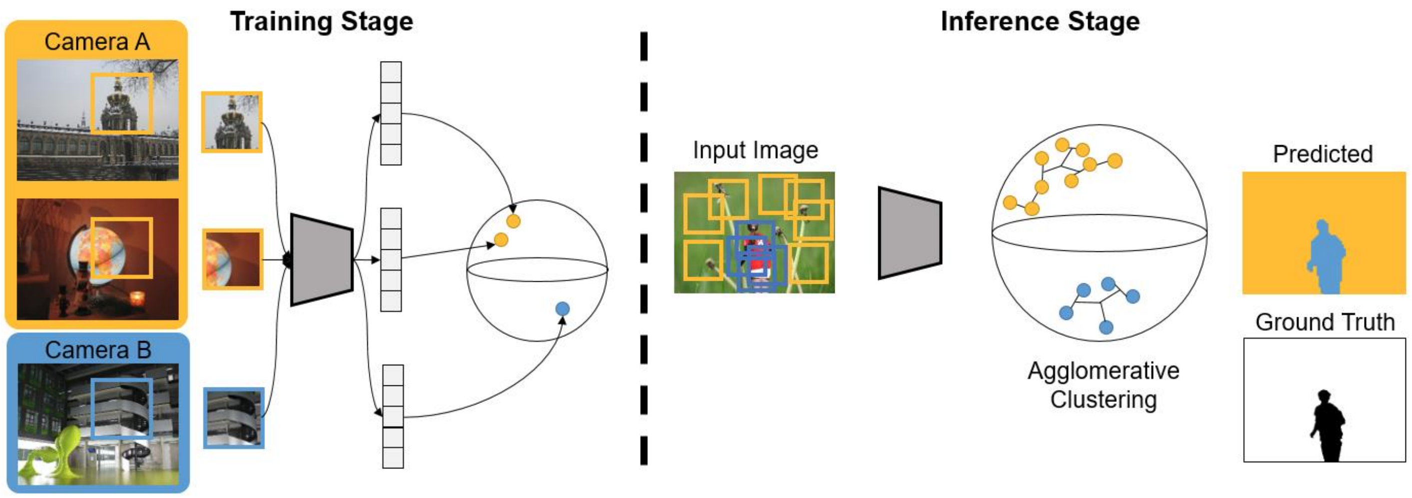


Correlated traces within images

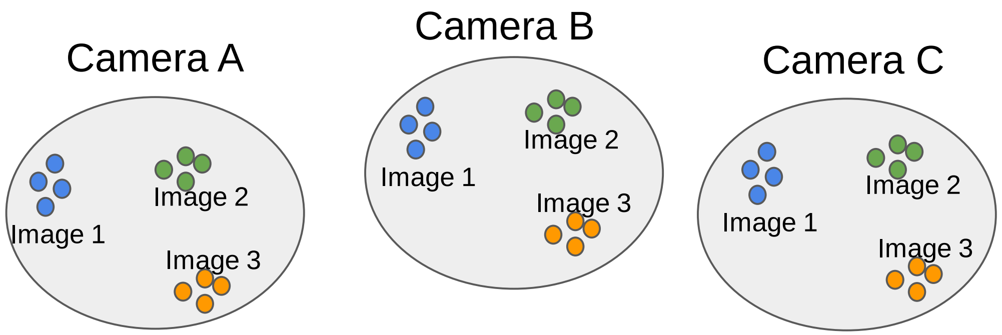


PRNU Differences

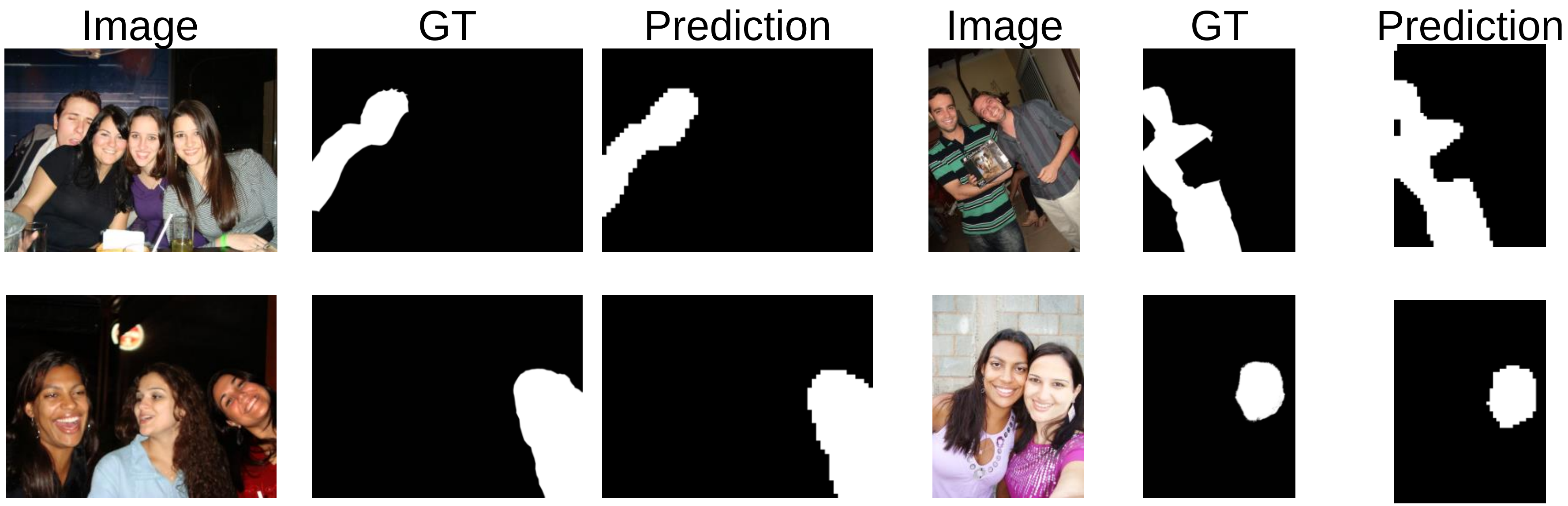
Patch Contrastive Learning



- 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 Preliminary Results



Thank You!

Any Questions?

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.
- [Zheng et al. 2021] Zheng, Yinglin, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. "Exploring temporal coherence for more general video face forgery detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15044-15054. 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.
- [Hui et al. 2022] Guo, Hui, Shu Hu, Xin Wang, Ming-Ching Chang, and Siwei Lyu. "Eyes Tell All: Irregular Pupil Shapes Reveal GAN-generated Faces." International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2022.
- [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.
- [Ho et al. 2020] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in Neural Information Processing Systems 33 (2020): 6840-6851.
- [Haliassos et al. 2022] Haliassos, Alexandros, Rodrigo Mira, Stavros Petridis, and Maja Pantic. "Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14950-14962. 2022.

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.
- [Chris Ume and Miles Fisher] https://www.youtube.com/watch?v=nwOywe7xLhs