



# A Brief Overview of Recent Development for Deepfake Detection and Other Countermeasures

**Jun-Cheng Chen**

[pullpull@citi.sinica.edu.tw](mailto:pullpull@citi.sinica.edu.tw)

Artificial Intelligence and Image Understanding Lab (AIIU)

Research Center of Information Technology Innovation, Academia Sinica

2023/11/15



# Recent Development of Generative AI<sup>(1/2)</sup>

- 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 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:

```
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.

OpenAI Research - API - ChatGPT - Safety - Company - Search - Log in - Try ChatGPT

Research

# DALL-E 3

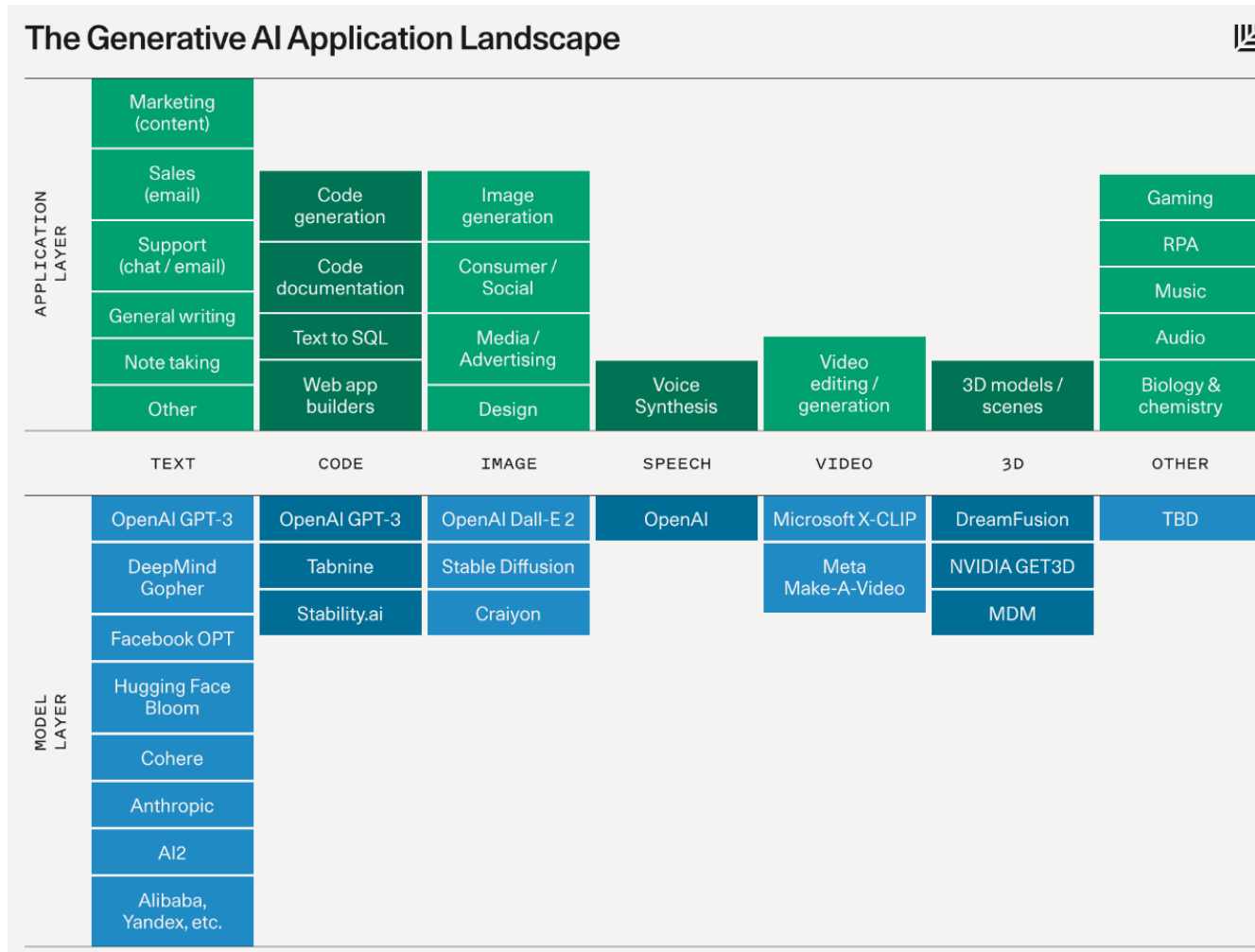
DALL-E 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images.

[Read research paper](#) [Try in ChatGPT](#)

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



# Recent Development of Generative AI<sup>(2/2)</sup>



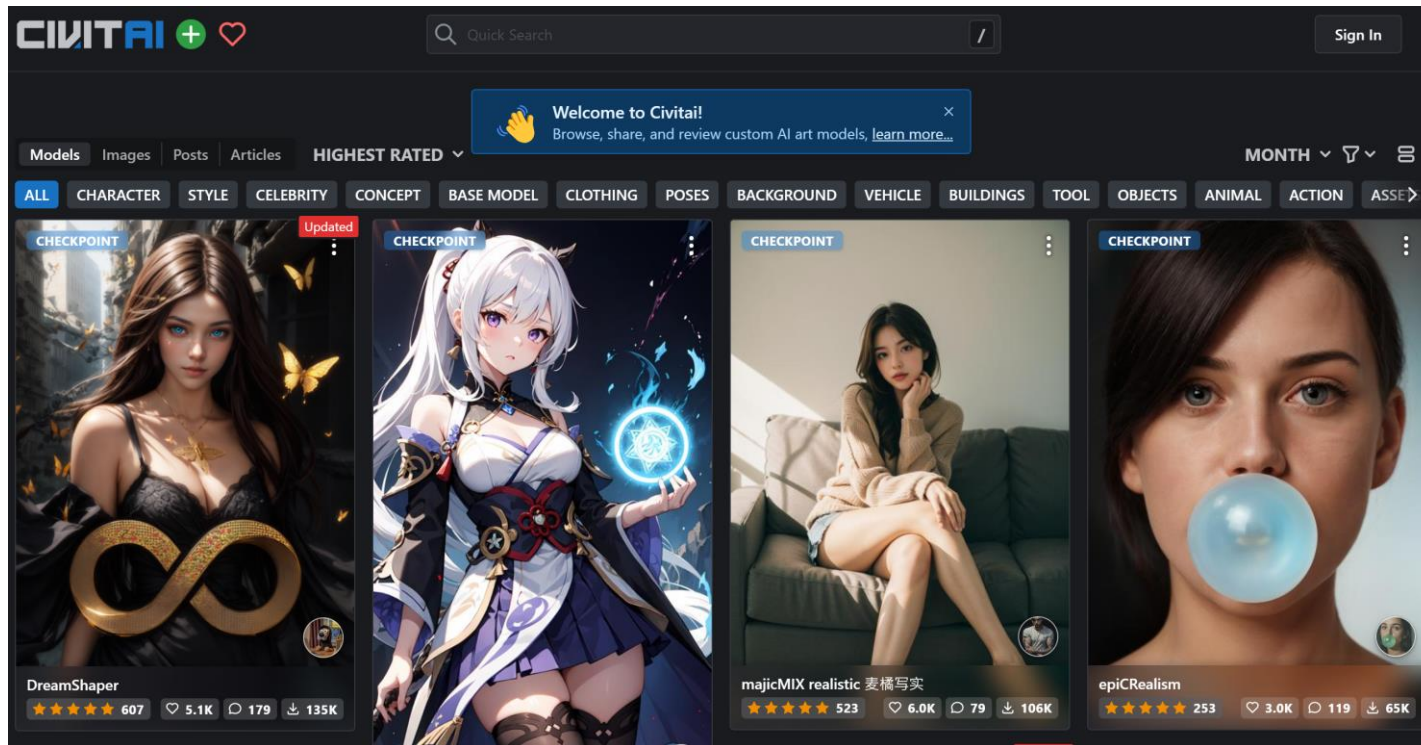
Cited from [https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/?fbclid=IwAR2GNyKNoEc\\_pv1TMjhUw2C7QluyIzNvnnv9HIW2kD080wqwXV3L3zL-14Sk](https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/?fbclid=IwAR2GNyKNoEc_pv1TMjhUw2C7QluyIzNvnnv9HIW2kD080wqwXV3L3zL-14Sk)



# Text-to-Image

- Stable Diffusion
- civitai

➤ <https://civitai.com/>



# Cases Sharing



Images generated by Midjourney v5 · It looks like a Facebook posts by someone.

Ref: FB 社團 IGC視覺藝創聯盟 (原 Midjourney AI 台灣社群)

Slide credit: Teddy Huang

# Cases Sharing



In Taiwan, there are already some commercial applications.

Ref: 和泰集團 ESG · 與美好台灣同行

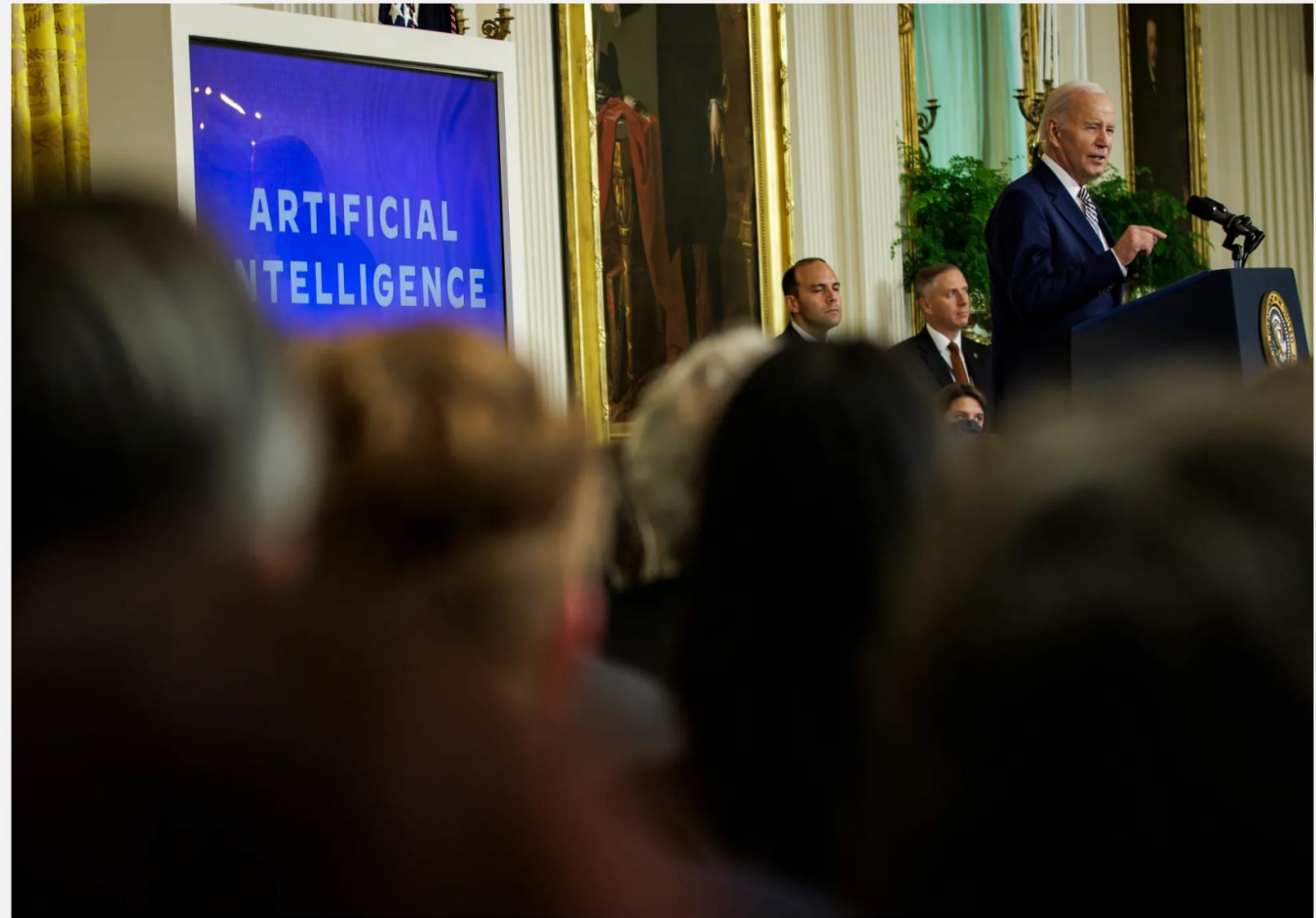
Slide credit: Teddy Huang

# Safe, Secure, Trustworthy Artificial Intelligence

COMMENTARY **PODCAST**

## Unpacking President Biden's executive order on artificial intelligence

Nicol Turner Lee and Fred Dews  
November 3, 2023



<https://www.brookings.edu/articles/unpacking-president-bidens-executive-order-on-artificial-intelligence/>

# Motivation

- Malicious Face Forgery Applications
  - Pornography
  - Politics

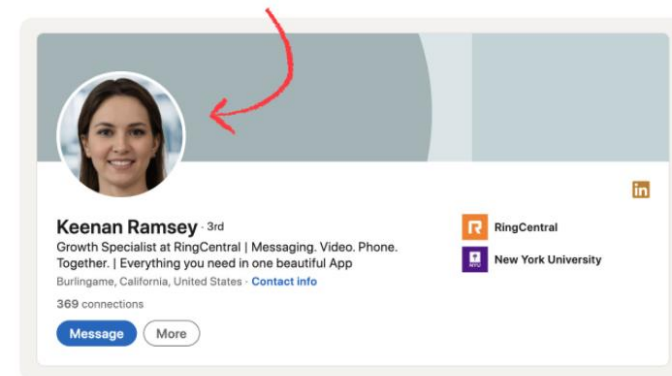


image source: <https://technews.tw/2020/10/25/deepfake-deepnude/>

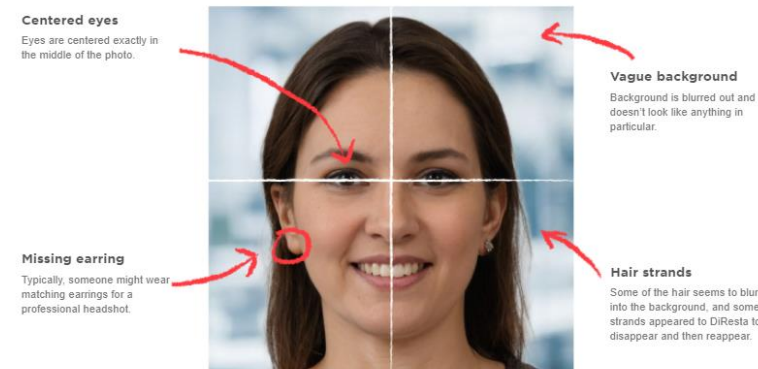
<https://www.youtube.com/watch?v=LFN9r70gk-Q>



At first glance, Keenan Ramsey might seem like a normal person on LinkedIn.



But certain details in her photo stood out to Stanford researcher Renée DiResta:



Source: Stanford Internet Observatory

Source: Stanford Internet Observatory

[https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?fbclid=IwAR3\\_ubq-9niHCYj10LeqIBogoMG9ExSMjz7azLhMlteu2D6-C-shsJhAKUE](https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles?fbclid=IwAR3_ubq-9niHCYj10LeqIBogoMG9ExSMjz7azLhMlteu2D6-C-shsJhAKUE)



# Motivation

## Training Set



*Caption: Living in the light  
with Ann Graham Lotz*

## Generated Image



*Prompt:  
Ann Graham Lotz*

\* Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting training data from diffusion models." *arXiv preprint arXiv:2301.13188* (2023).

\* Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. "Diffusion art or digital forgery? investigating data replication in diffusion models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048-6058. 2023.

# CheapFakes v.s. Deepfakes

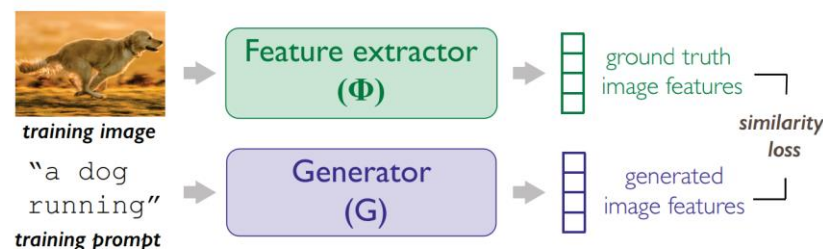
- Cheapfake
  - Photoshop, ... Etc.



Image credit: Daniel Stanley Tan

- Deepfake
  - AI generated Content

## Model training



## Image generation

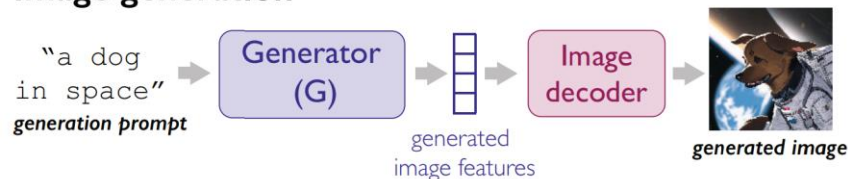
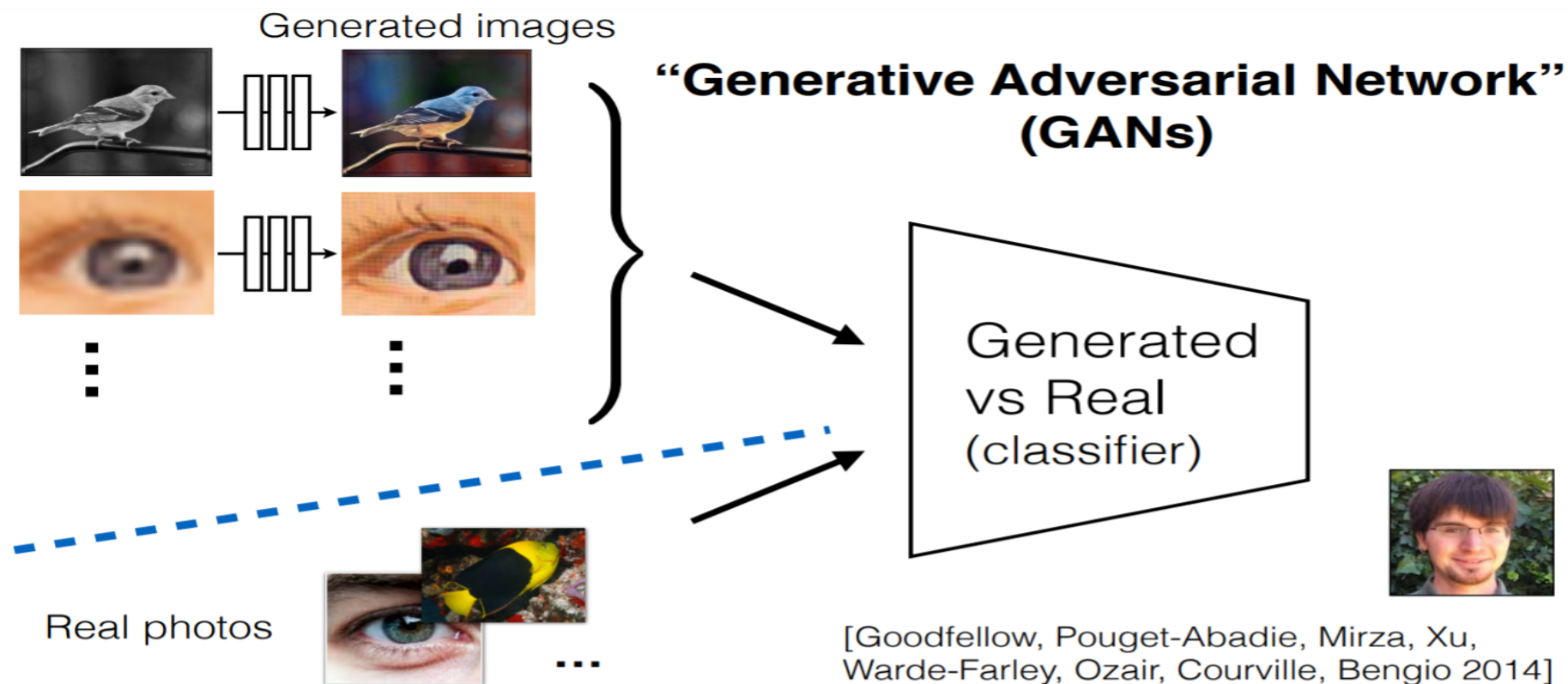


Image credit: [Shawn et al. 2023 GLAZE]

# The Evolution of Content Editing

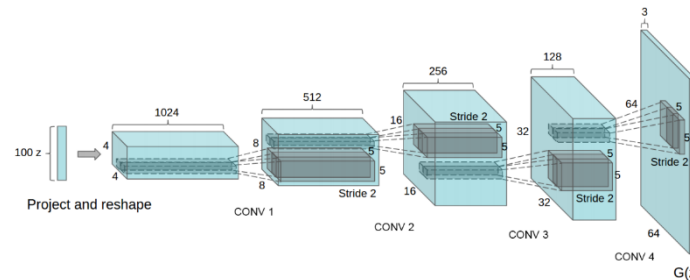
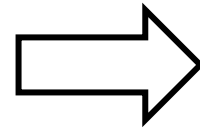


Slide Credit: Prof. Philip Isola, MIT

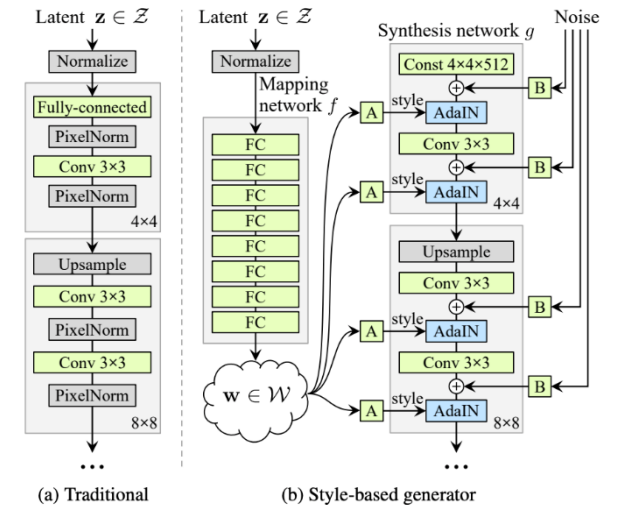
# The Evolution of Content Editing



LightStage  
[USC ICT 2015]



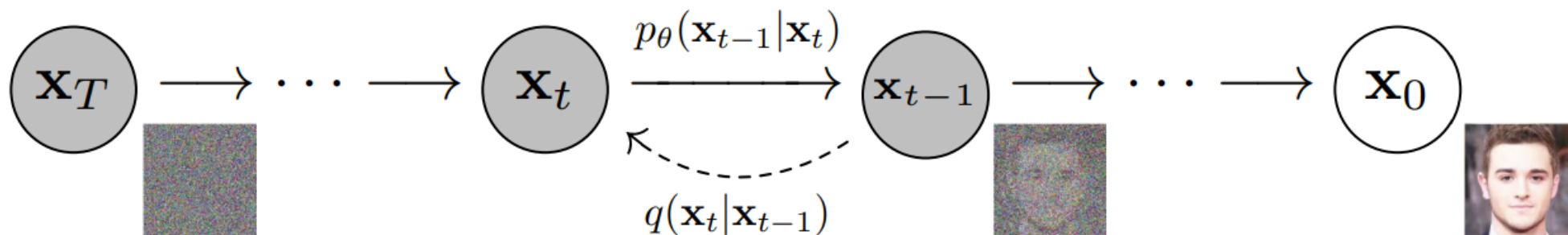
DCGAN  
[Radford et al. 2016]



StyleGAN  
[Karras et al. 2019]



# Denoising Diffusion Probabilistic Model



$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad \Rightarrow \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})$$
$$p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I}) \quad \Rightarrow \quad p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$

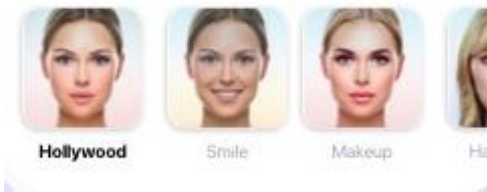
Learned Model (U-Net)

- 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/>

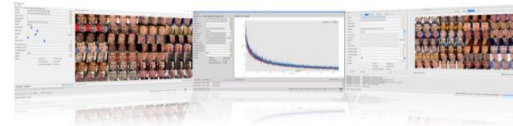
# The Evolution of Content Editing



FaceApp



Faceswap is the leading free and Open Source multi-platform Deepfakes software.

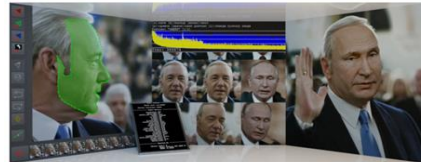


Faceswap

DeepFaceLab

<https://arxiv.org/abs/2005.05335>

the leading software for creating deepfakes



DeepFaceLab



Midjourney

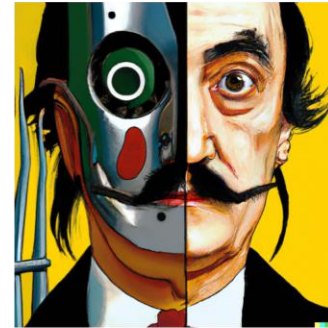
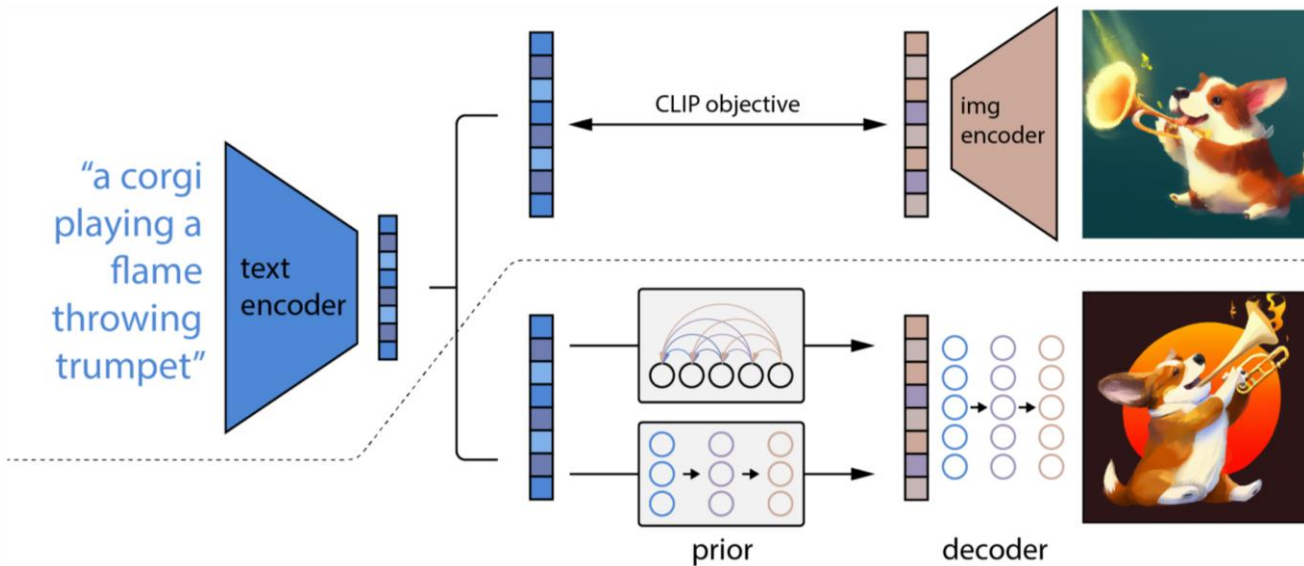


Video credit: *Chris Ume and Miles Fisher*



# DALL·E 2 (DALL·E 3)

From OpenAI



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



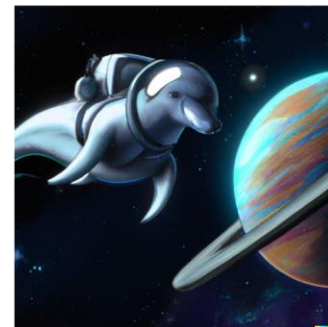
an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula



a dolphin in an astronaut suit on saturn, artstation



a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

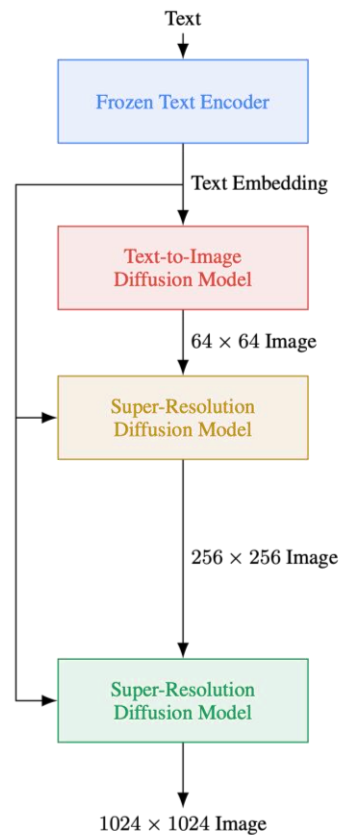


a teddy bear on a skateboard in times square

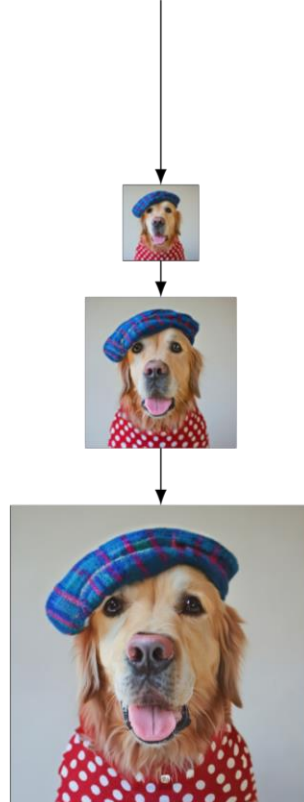
Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents, OpenAI, 2022



# Imagen From Google



“A Golden Retriever dog wearing a blue checked beret and red dotted turtleneck.”



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.



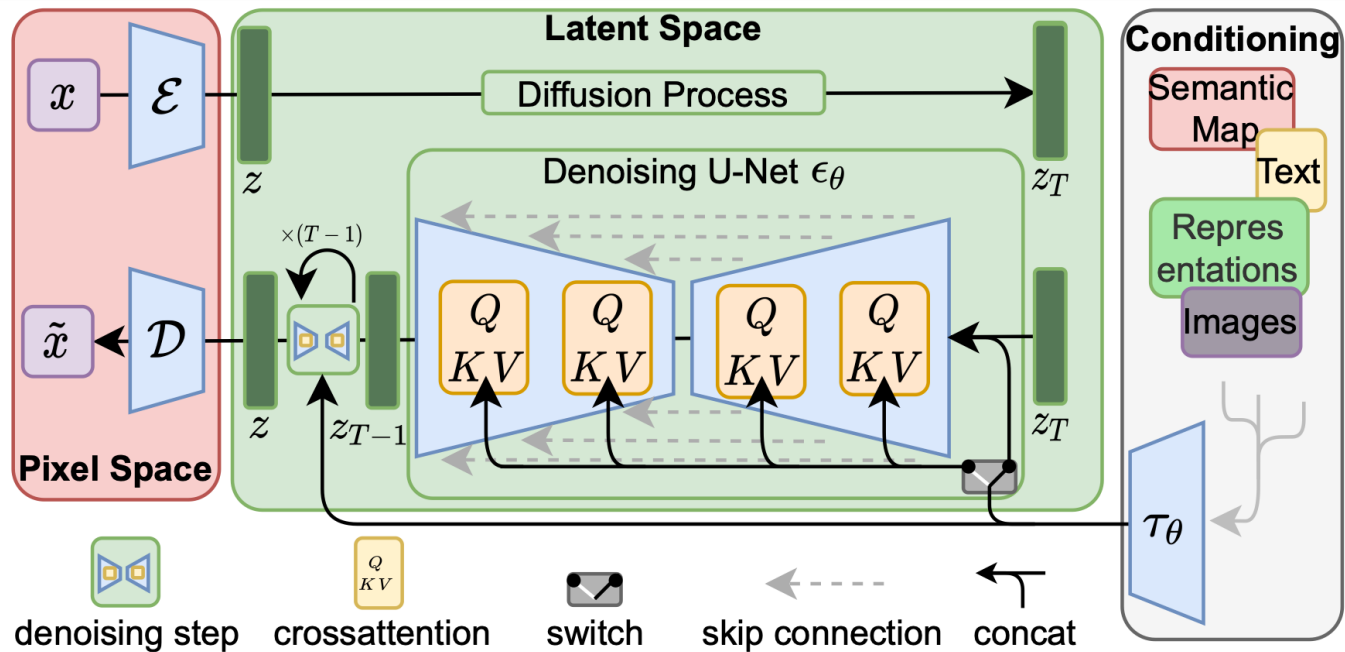
A strawberry splashing in the coffee in a mug under the starry sky.

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





# Stable Diffusion



"A street sign that reads 'Latent Diffusion'"



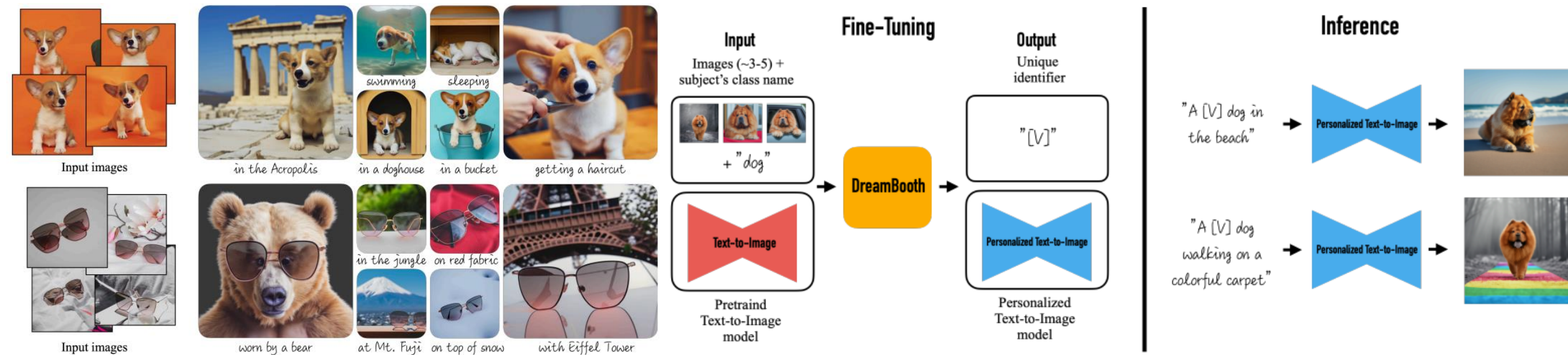
"An oil painting of a space shuttle"

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

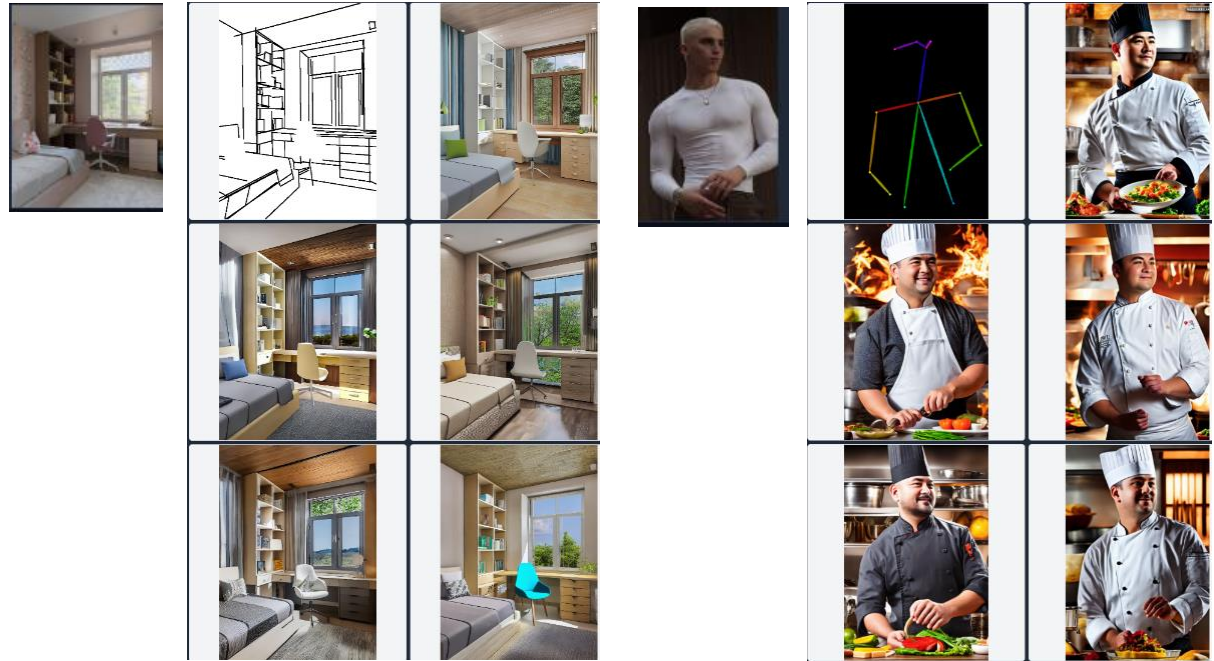
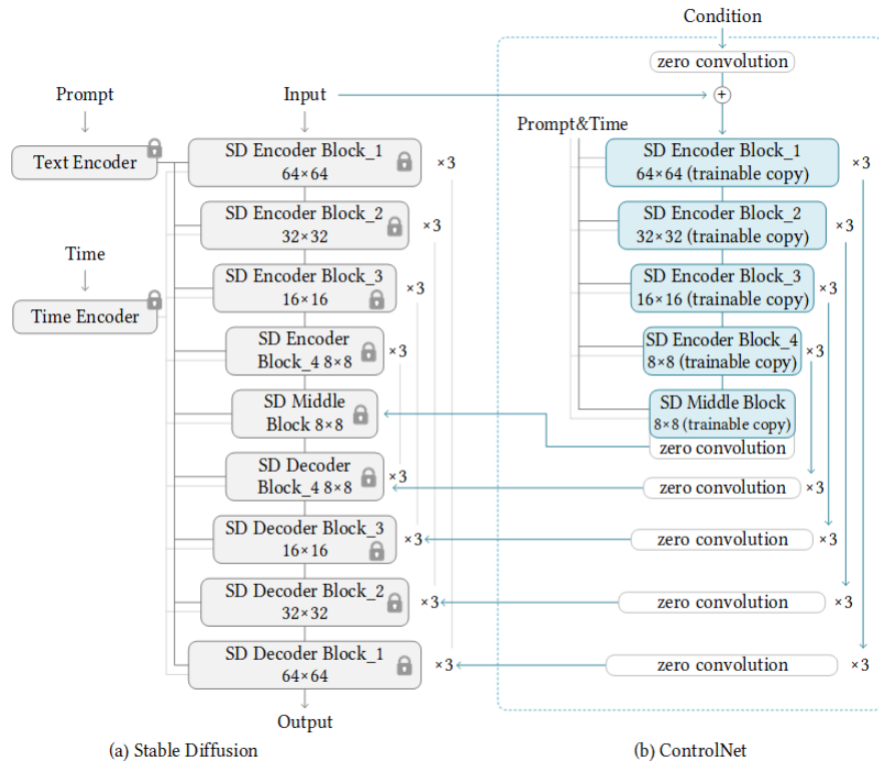


# DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation



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).

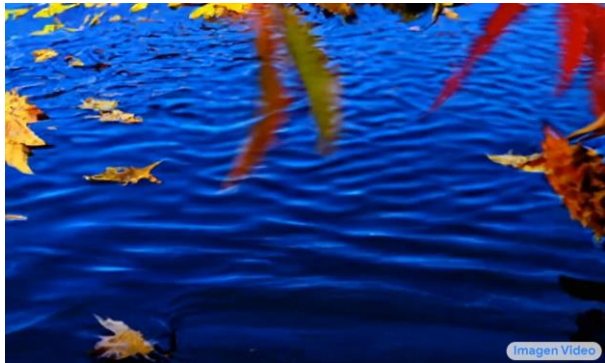
# ControlNet



Zhang, Lvmin, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *arXiv preprint arXiv:2302.05543* (2023).

# 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.

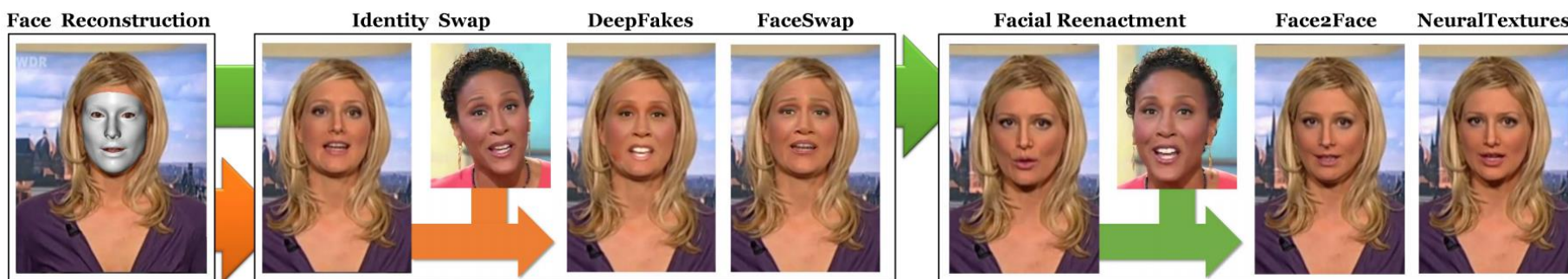
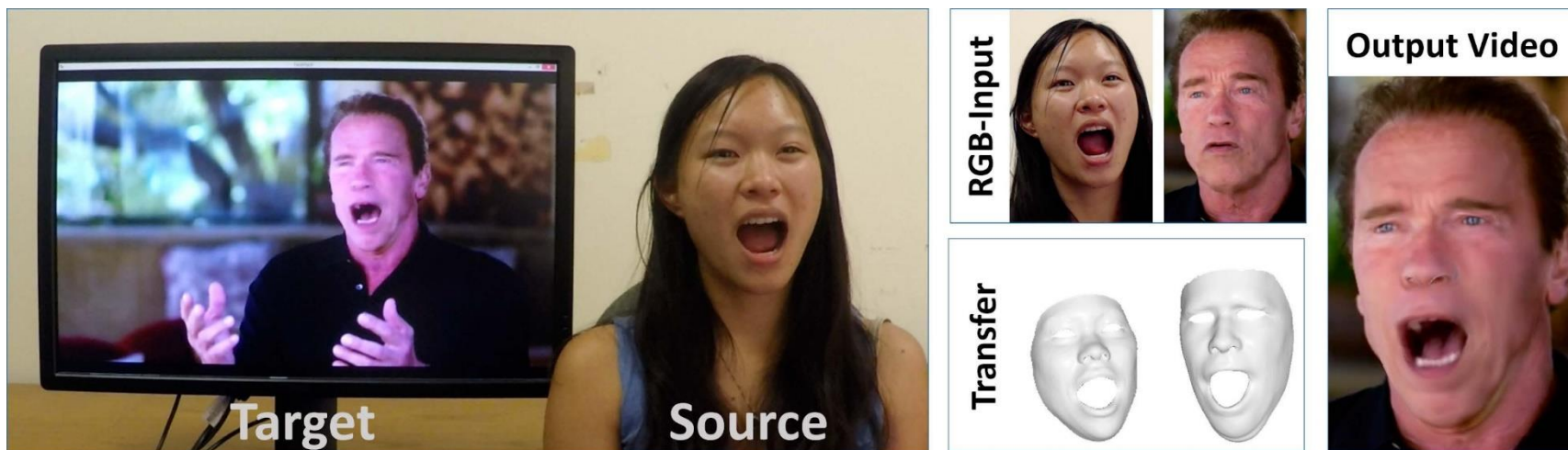


# Possible Countermeasures

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



# Common Deepfake Manipulation



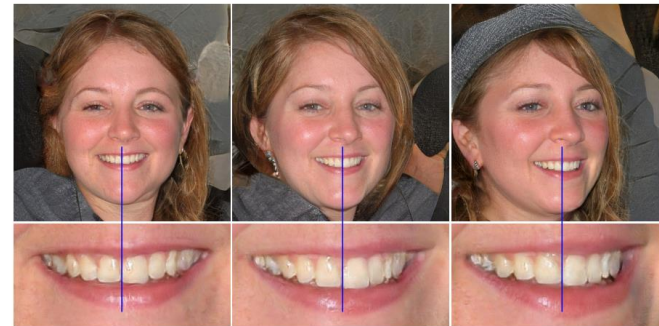
Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1-11).  
 Thies, Justus, Michael Zollhofer, Marc Stamminger, Christian Theobalt, 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]



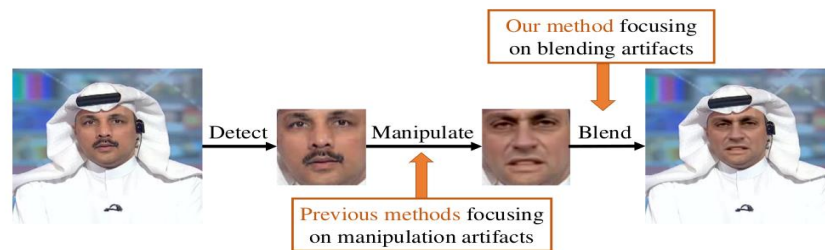
StyleGAN  
[Karras et al. 2019]



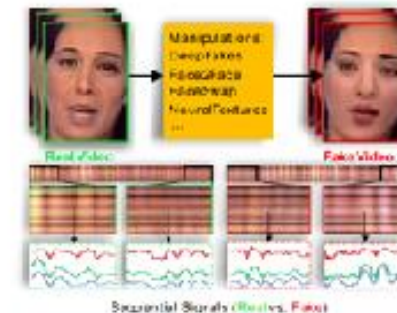
Irregular Pupil Shape  
[Hui et al. 2022]



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



FaceXRay  
[Li et al. 2020]



Deep Rhythm

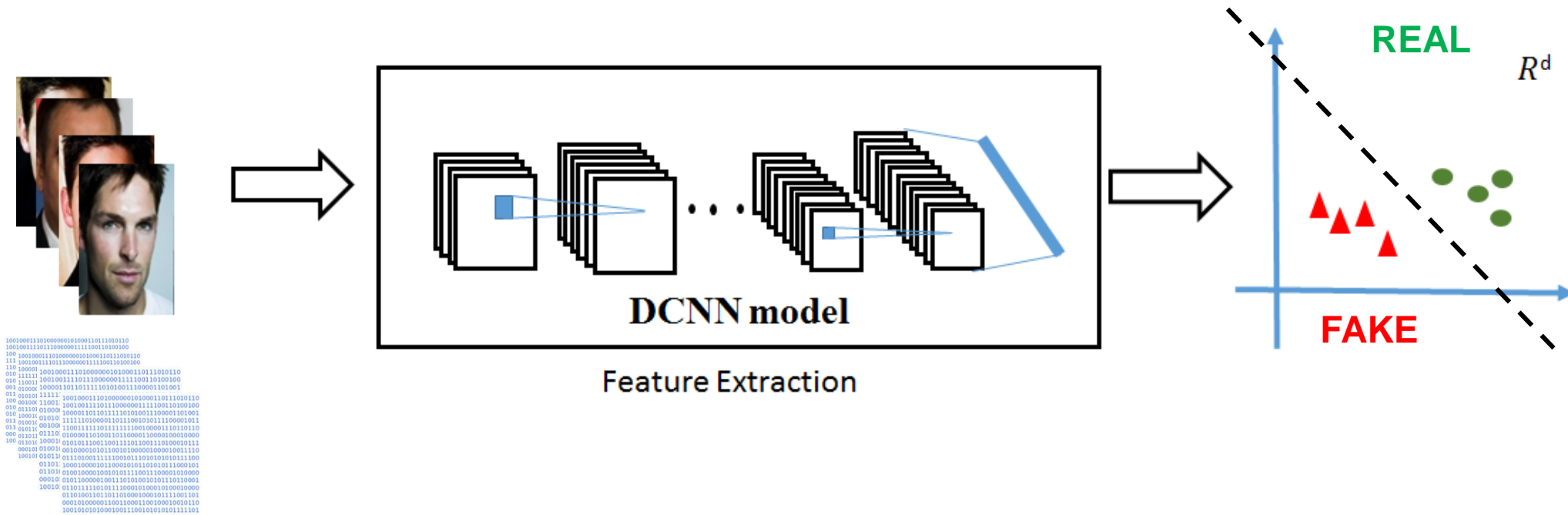


<https://twitter.com/umecha1128/status/1582572128880115712>

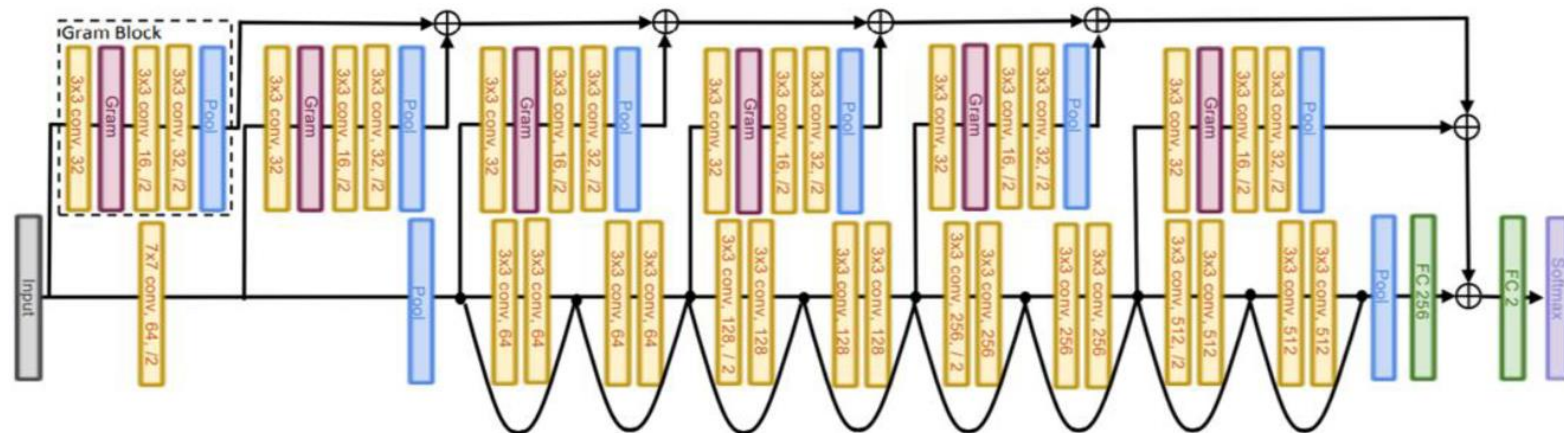


# Deepfake Detection

- 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	<b>99.10 ± 1.36</b>	<b>95.84 ± 1.98</b>	<b>99.05 ± 1.37</b>	<b>92.39 ± 2.66</b>	<b>94.20 ± 5.57</b>	<b>92.47 ± 4.52</b>	<b>95.51</b>
CelebA-HQ vs. CelebA-HQ	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	<b>98.54 ± 1.27</b>	<b>82.40 ± 6.30</b>	<b>94.65 ± 3.28</b>	<b>79.77 ± 6.13</b>	<b>91.96 ± 4.78</b>	<b>88.29 ± 3.44</b>	<b>89.26</b>
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	<b>98.78 ± 0.49</b>	<b>94.66 ± 3.10</b>	<b>97.29 ± 1.05</b>	<b>94.08 ± 3.22</b>	<b>98.55 ± 0.92</b>	<b>70.32 ± 12.04</b>	<b>92.28</b>
CelebA-HQ vs. CelebA-HQ	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	<b>60.89 ± 7.24</b>	85.97
		Gram-Net	<b>98.55 ± 0.89</b>	<b>91.57 ± 2.95</b>	<b>94.28 ± 3.67</b>	<b>83.64 ± 3.43</b>	<b>97.05 ± 1.04</b>	60.07 ± 7.32	<b>87.52</b>
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	<b>75.60 ± 10.75</b>	81.32 ± 5.06	81.50
		Gram-Net	<b>98.96 ± 0.51</b>	<b>89.22 ± 4.44</b>	<b>98.69 ± 0.81</b>	<b>87.86 ± 3.42</b>	70.99 ± 6.07	<b>94.27 ± 2.12</b>	<b>90.00</b>

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

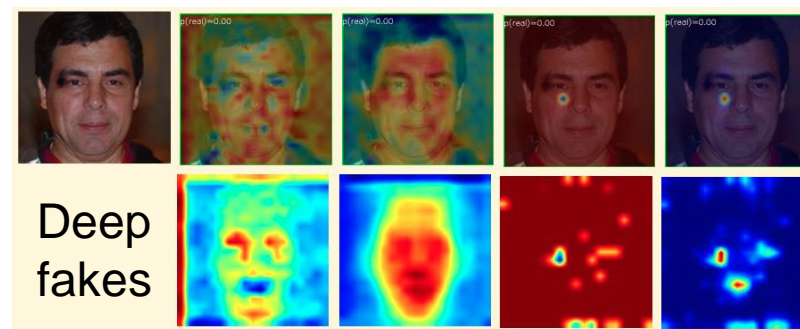
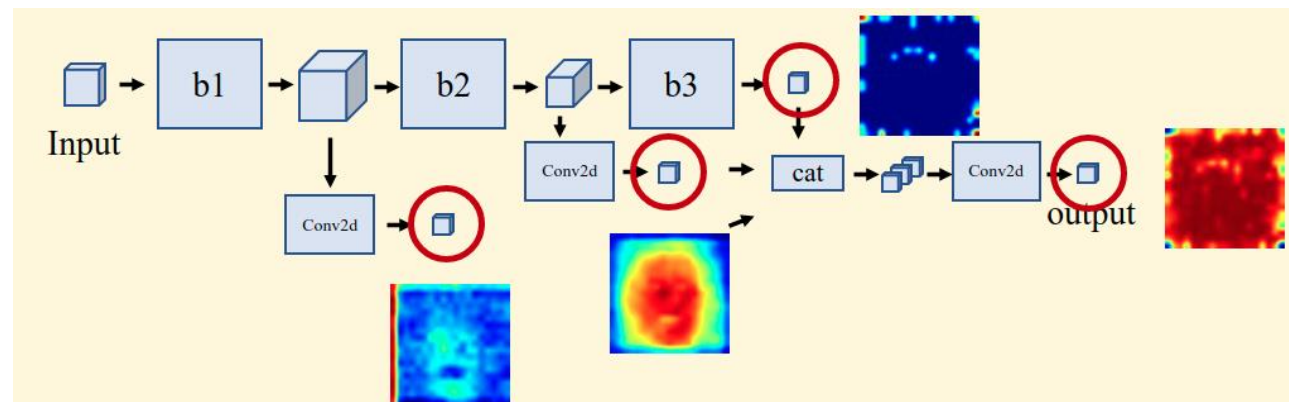
[Liu et al. 2020]



# 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

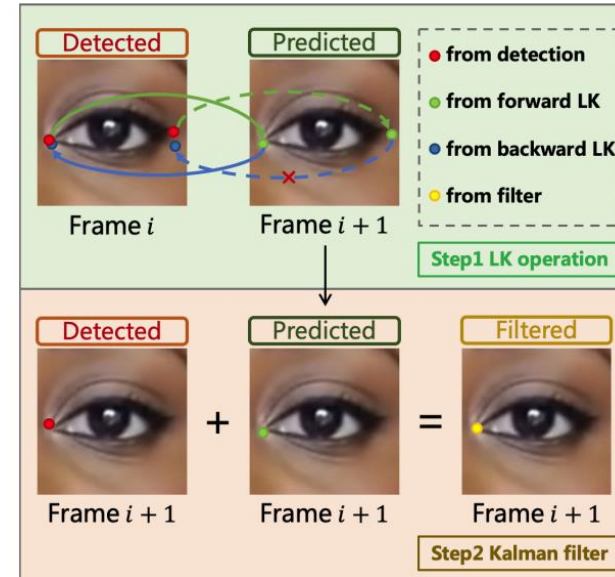
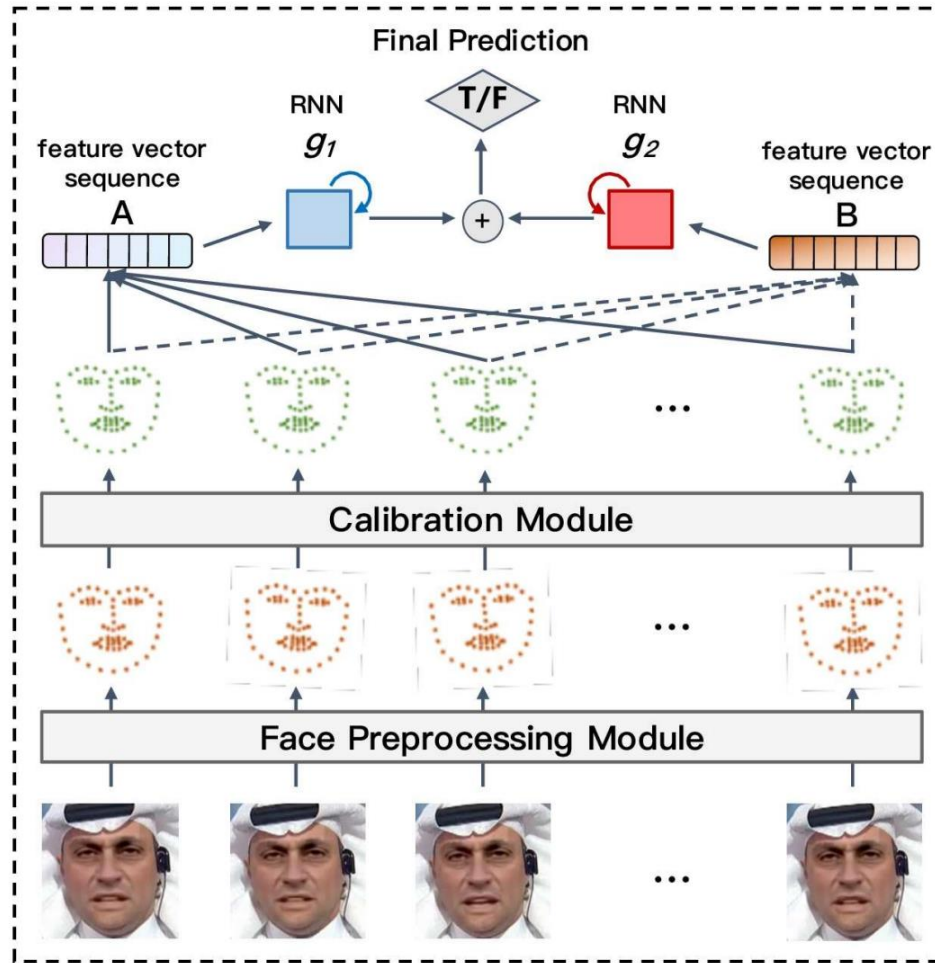


block 1 block 2 block 3 output

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.

# Improving the Efficiency and Robustness of Deepfakes

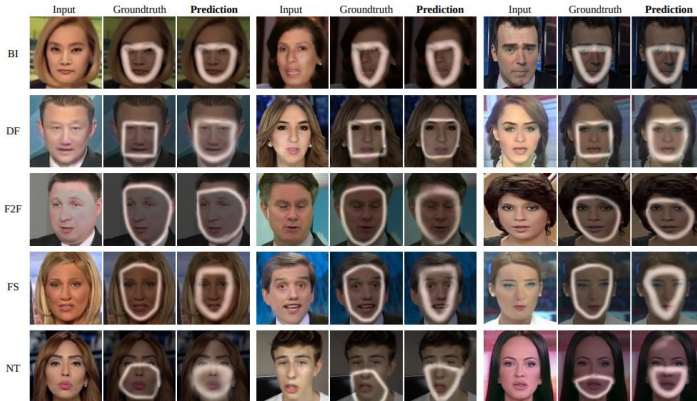
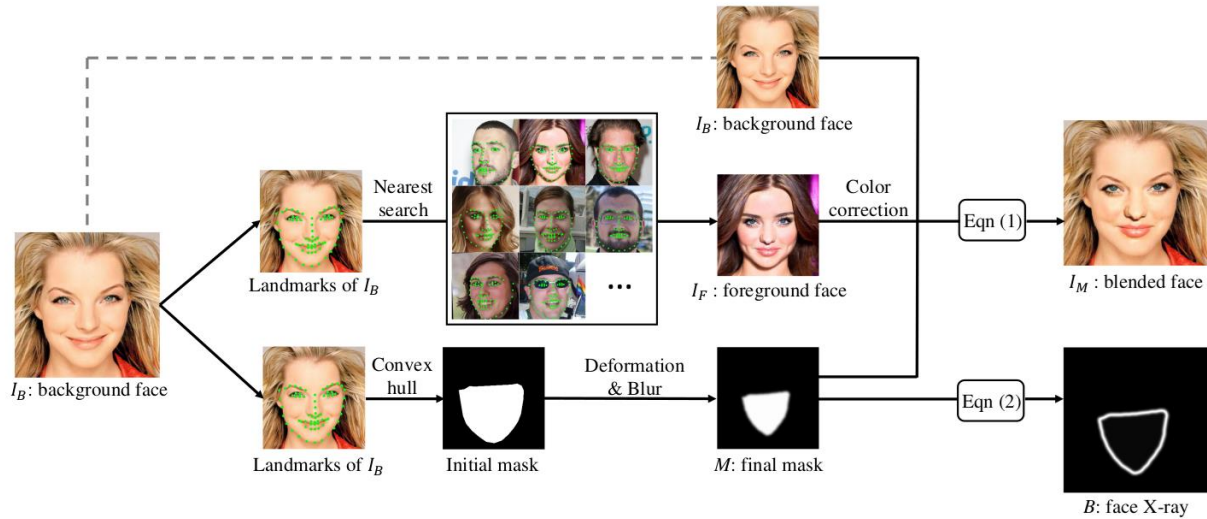
## Detection through Precise Geometric Features



Methods	Configurations			Testing Datasets		
	Size	Aug.	Training	UADFV	FF++	Celeb-DF
Meso4 [1]	0.03 M	×	Unpub.	84.3	84.7	54.8
FWA [18]	26 M	✓	Unpub.	97.4	80.1	56.9
DSP-FWA [18]	28 M	✓	Unpub.	97.7	93.0	<b>64.6</b>
Xception [25]	20.8 M	×	FF++	80.4	99.7	48.2
Capsule [23]	15 M	×	FF++	61.3	96.6	57.5
CNN+RNN [26]	24.3 M	×	FF++	70.9	98.3	61.5
<b>LRNet (ours)</b>	<b>0.18 M</b>	<b>×</b>	<b>FF++</b>	<b>98.5</b>	<b>99.9</b>	56.9

Sun, Zekun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. "Improving the efficiency and robustness of deepfakes detection through precise geometric features." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3609-3618. 2021.

# Face X-ray for More General Face Forgery Detection



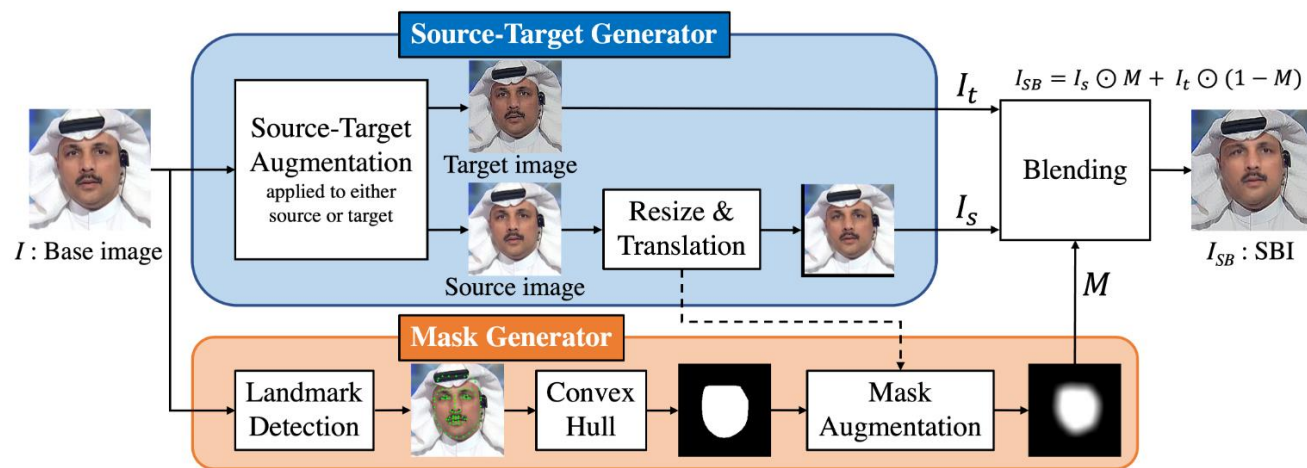
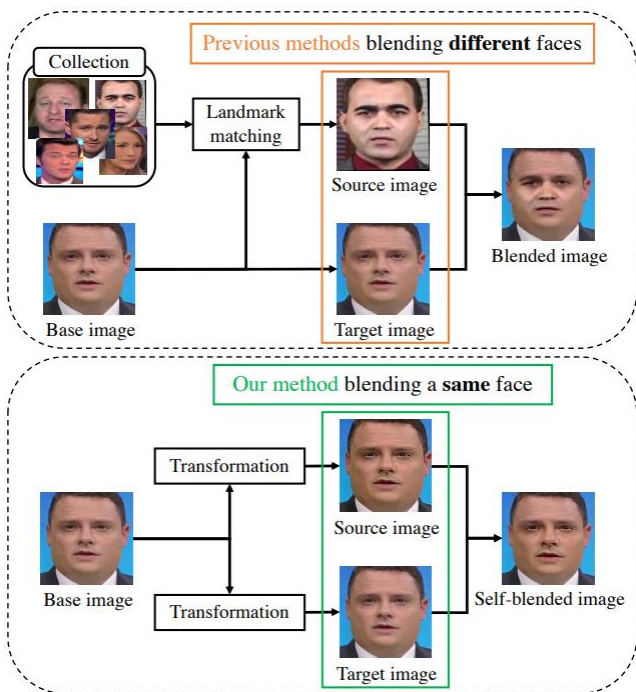
Model	Training set		Test set AUC				
	DF	BI	DF	F2F	FS	NT	FF++
Xception [36]	✓	–	<b>99.38</b>	75.05	49.13	80.39	76.34
HRNet	✓	–	<b>99.26</b>	68.25	39.15	71.39	69.51
Face X-ray	✓	–	<b>99.17</b>	94.14	75.34	93.85	90.62
	✓	✓	<b>99.12</b>	<b>97.64</b>	<b>98.00</b>	<b>97.77</b>	<b>97.97</b>
	F2F	BI	DF	F2F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

[Li et al. 2020]



# More Self-Supervised Deepfake Detection Approaches

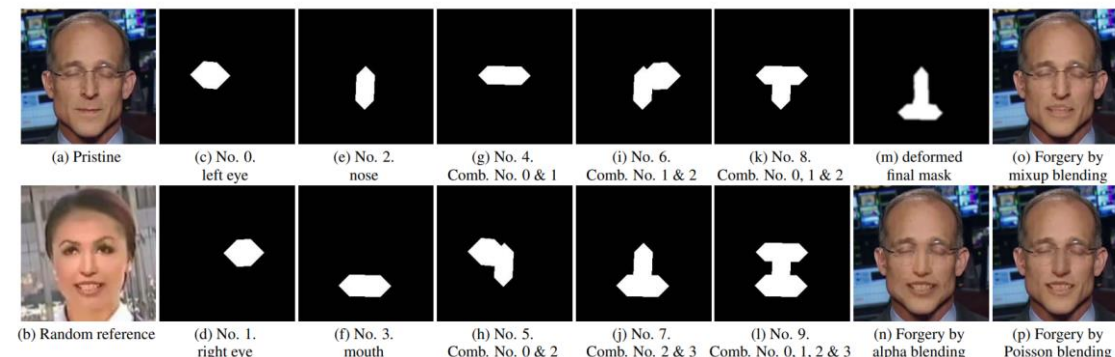
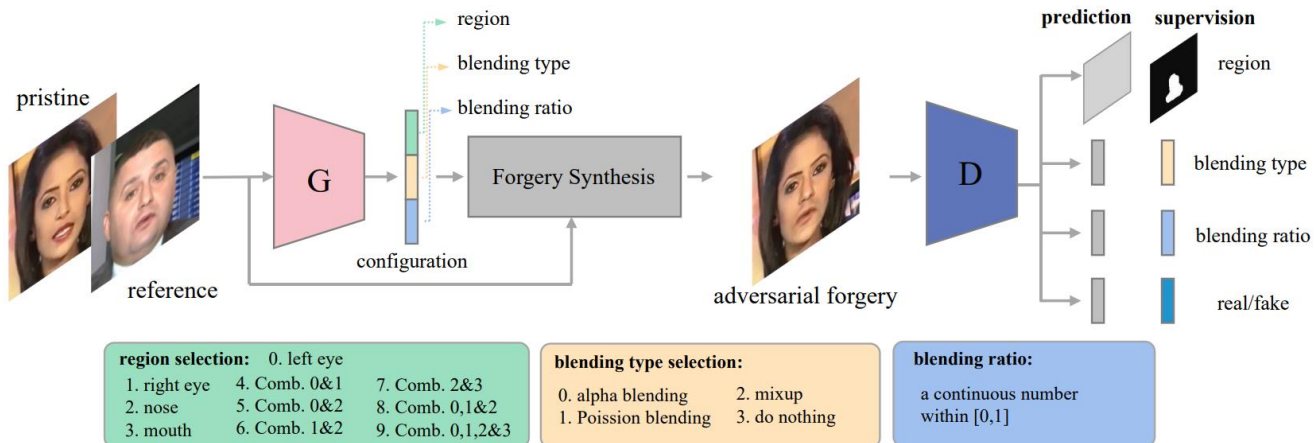
- Self-Blended Image



Method	Input Type	Training Set		Test Set AUC (%)				
		Real	Fake	CDF	DFD	DFDC	DFDCP	FFIW
DSP-FWA [41]	Frame	✓	✓	69.30	-	-	-	-
Face X-ray + BI [39]	Frame	✓		-	93.47	-	71.15	-
Face X-ray + BI [39]	Frame	✓	✓	-	95.40	-	80.92	-
LRL [13]	Frame	✓	✓	78.26	89.24	-	76.53	-
FRDM [44]	Frame	✓	✓	79.4	91.9	-	79.7	-
PCL + I2G [65]	Frame	✓		<u>90.03</u>	<b>99.07</b>	67.52	74.37	-
Two-branch [47]	Video	✓	✓	76.65	-	-	-	-
DAM [67]	Video	✓	✓	75.3	-	-	72.8	-
LipForensics [27]	Video	✓	✓	82.4	-	-	-	-
FTCN [66]	Video	✓	✓	86.9	94.40*	71.00*	74.0	74.47*
EFNB4 + SBIs (Ours)	Frame	✓		<b>93.18</b>	<u>97.56</u>	<b>72.42</b>	<b>86.15</b>	<b>84.83</b>

Shiohara, Kaede, and Toshihiko Yamasaki. "Detecting deepfakes with self-blended images." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720-18729. 2022.

# More Self-Supervised Deepfake Detection Approaches

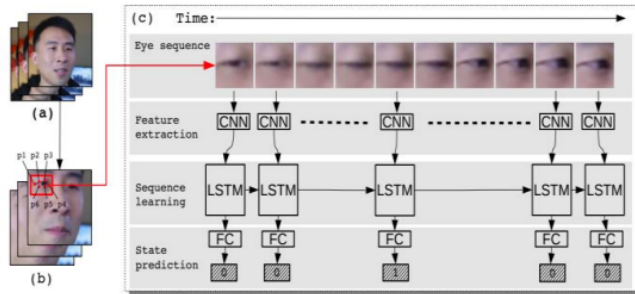


Method	DF			F2F			FS			NT			Avg.
	DFDC	CelebDF	DF1.0	DFDC	CelebDF	DF1.0	DFDC	CelebDF	DF1.0	DFDC	CelebDF	DF1.0	
Xception [41]	0.654	0.681	0.617	0.708	0.598	0.745	0.708	0.601	0.605	0.646	0.625	0.838	0.669
Face X-ray [24]	0.609	0.554	0.668	0.633	0.684	0.766	0.646	0.697	<b>0.795</b>	0.613	0.703	0.866	0.686
F3Net [39]	0.682	0.664	0.658	0.679	0.654	0.761	0.679	0.636	0.651	0.672	0.689	0.932	0.696
RFM [47]	0.758	0.723	0.717	0.736	0.663	0.732	0.714	0.591	0.714	0.726	0.600	0.846	0.710
SRM [30]	0.679	0.650	0.720	0.687	0.693	0.775	0.671	0.643	0.771	0.656	0.651	<b>0.936</b>	0.711
Ours	<b>0.772</b>	<b>0.730</b>	<b>0.742</b>	<b>0.787</b>	<b>0.781</b>	<b>0.786</b>	<b>0.742</b>	<b>0.800</b>	0.695	<b>0.741</b>	<b>0.759</b>	0.889	<b>0.769</b>

Chen, Liang, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18710-18719. 2022.

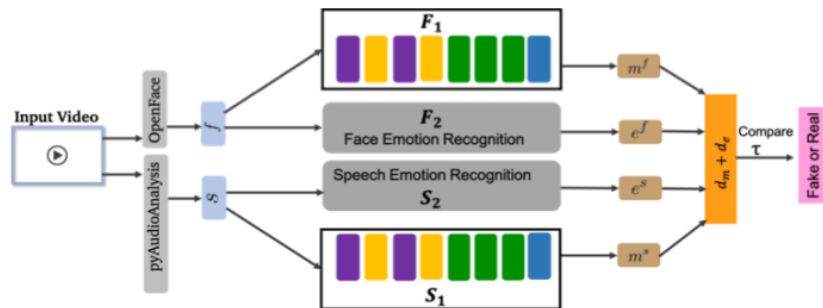
# Video Consistency<sup>(1/2)</sup>

- Temporal Inconsistency between frames

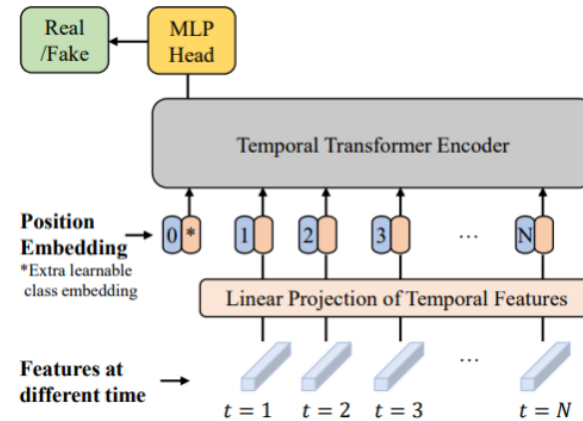


In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking, WIFS 2018

- Audio-visual inconsistency

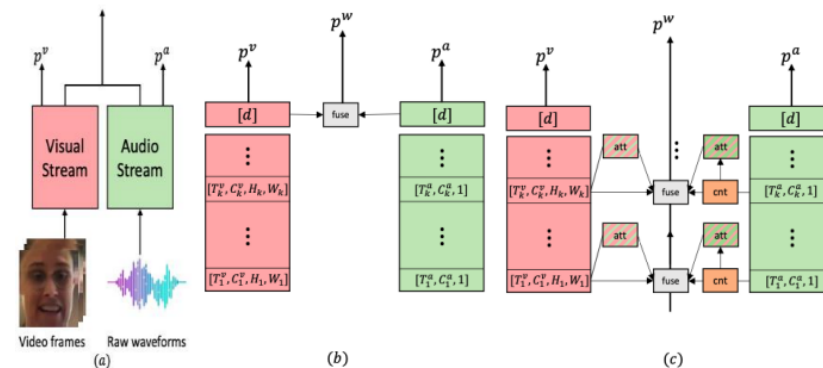


Emotions Don't Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues, ACM MM 2020



	layer	output size
conv <sub>1</sub>	5 × 1 × 1, 64, stride 1, 1, 1	64 × 32 × 224 × 224
pool <sub>1</sub>	1 × 5 × 5 max, stride 1, 4, 4	256 × 32 × 56 × 56
res <sub>2</sub>	$\begin{bmatrix} 1 \times 1 \times 1, 64 \\ 3 \times 1 \times 1, 64 \\ 1 \times 1 \times 1, 256 \end{bmatrix} \times 3$	256 × 32 × 56 × 56
pool <sub>2</sub>	2 × 1 × 1 max, stride 2, 1, 1	256 × 16 × 56 × 56
res <sub>3</sub>	$\begin{bmatrix} 1 \times 1 \times 1, 128 \\ 3 \times 1 \times 1, 128 \\ 1 \times 1 \times 1, 512 \end{bmatrix} \times 4$	512 × 16 × 28 × 28
res <sub>4</sub>	$\begin{bmatrix} 1 \times 1 \times 1, 256 \\ 3 \times 1 \times 1, 256 \\ 1 \times 1 \times 1, 1024 \end{bmatrix} \times 6$	1024 × 16 × 14 × 14
res <sub>5</sub>	$\begin{bmatrix} 1 \times 1 \times 1, 512 \\ 3 \times 1 \times 1, 512 \\ 1 \times 1 \times 1, 2048 \end{bmatrix} \times 3$	2048 × 16 × 7 × 7
	spatial-related average pool	2048 × 16 × 1 × 1

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

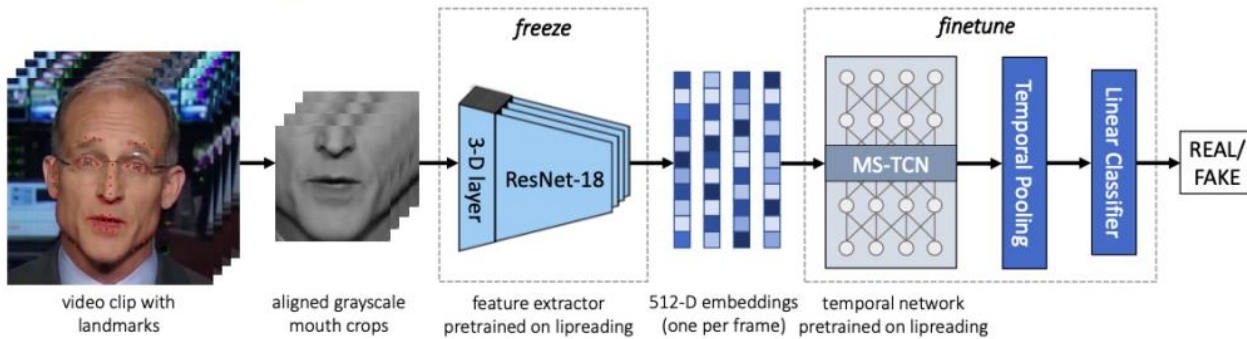


Joint Audio-Visual Deepfake Detection, ICCV 2021



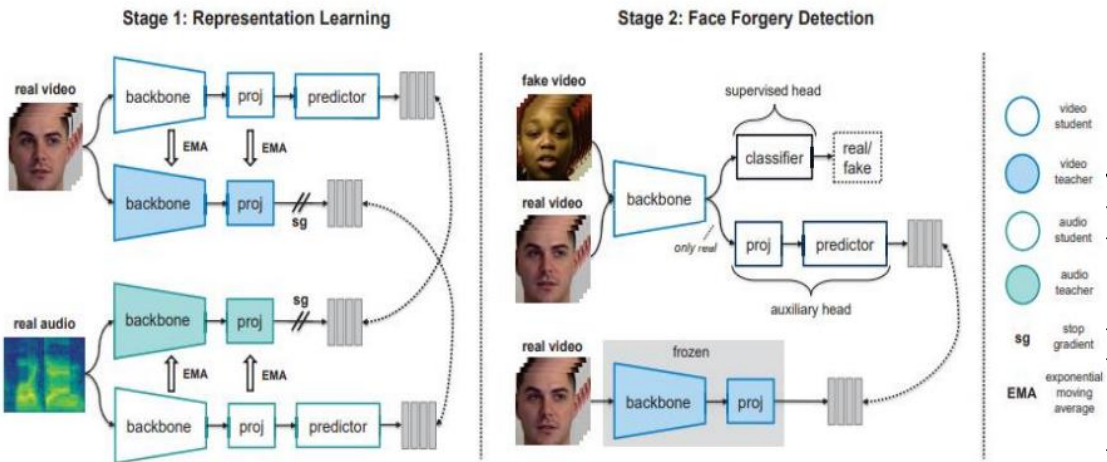
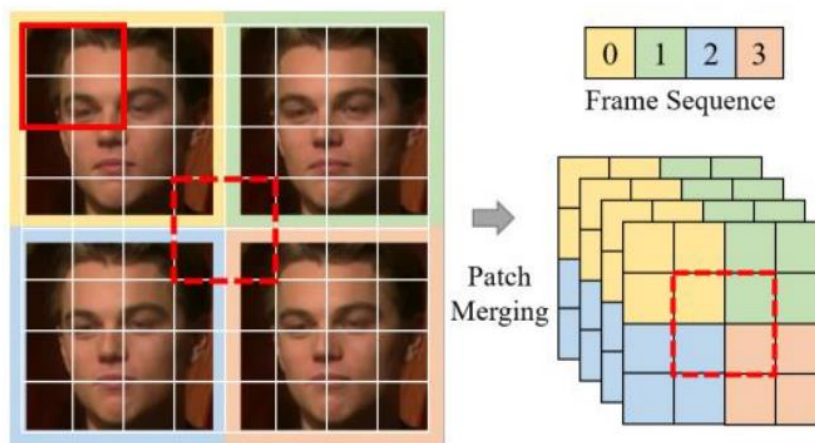
# Video Consistency<sup>(2/2)</sup>

- More temporal inconsistency



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

- Spatio-temporal inconsistency



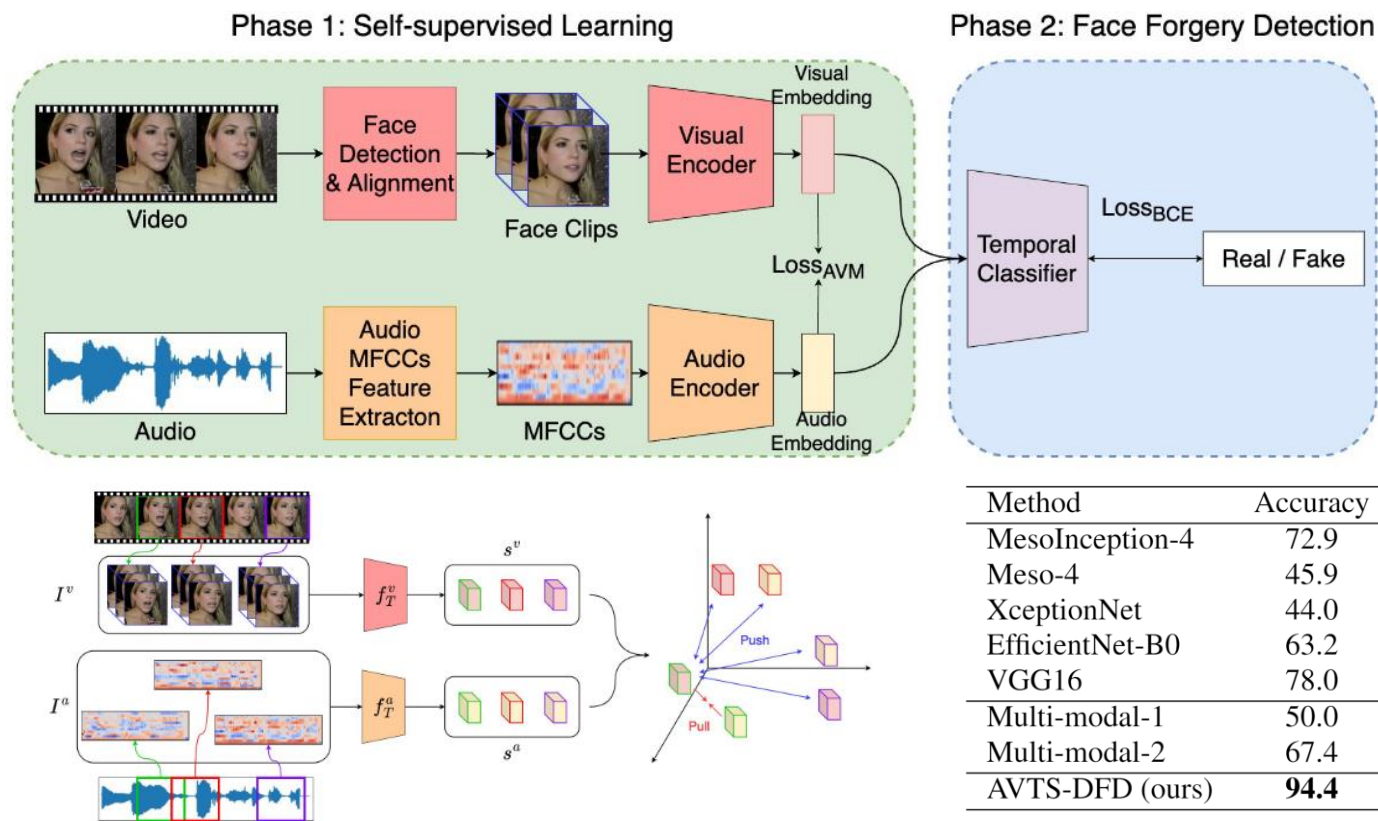
Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection. CVPR 2022

TALL: Thumbnail Layout for Deepfake Video Detection, ICCV 2023



# SELF-SUPERVISED AUDIO-VISUAL MUTUAL LEARNING FOR DEEPFAKE DETECTION (ICASSP 2023)

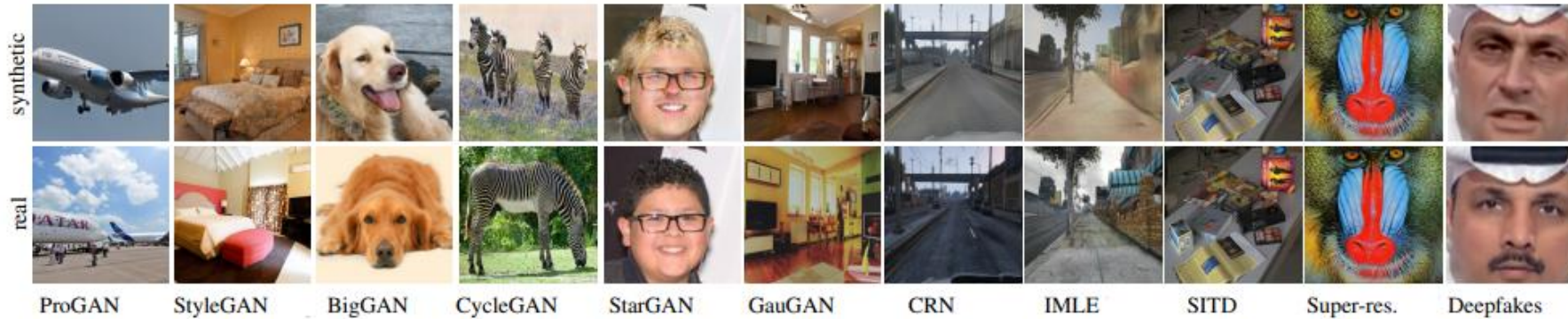
- Develop an effective audio-visual self-supervised pretraining-based feature extractor which can significantly improve the generalization of finetuned Deepfake detector for unseen Deepfakes.



ChangSung Sung, Jun-Cheng Chen, Chu-Song Chen, "Hearing and Seeing Abnormality: Self-supervised Audio-Visual Mutual Learning for Deepfake Detection," *EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.



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



Family	Name	Training settings					Individual test generators										Total mAP	
		Train	Input	No. Class	Augments		Pro-GAN	Style-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN	CRN	IMLE	SITD	SAN		Deep-Fake
					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	<b>98.1</b>	<b>99.0</b>	<b>99.7</b>	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	<b>93.6</b>	<b>98.2</b>	90.1
	Blur only	ProGAN	RGB	20	✓		100.	99.0	82.5	90.1	100.	74.7	66.6	66.7	<b>99.6</b>	53.7	95.1	84.4
	JPEG only	ProGAN	RGB	20		✓	100.	99.0	87.8	93.2	91.8	97.5	<b>99.0</b>	99.5	88.7	78.1	88.1	<b>93.0</b>
	Blur+JPEG (0.5)	ProGAN	RGB	20	✓	✓	100.	98.5	<b>88.2</b>	96.8	95.4	<b>98.1</b>	98.9	99.5	92.7	63.9	66.3	90.8
Blur+JPEG (0.1)	ProGAN	RGB	20	†	†	100.	<b>99.6</b>	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<sup>(1/2)</sup>

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	<b>100.0 / 100.0 / 100.0</b>	<b>100.0 / 100.0 / 100.0</b>	<b>100.0 / 100.0 / 100.0</b>	<b>100.0 / 100.0 / 100.0</b>	91.2 / 54.6 / 27.5
StyleGAN	98.7 / 93.7 / 81.4	99.0 / 95.5 / 84.4	<b>100.0 / 100.0 / 100.0</b>	<b>100.0 / 100.0 / 100.0</b>	89.6 / 43.6 / 14.7
ProjectedGAN	94.8 / 73.8 / 49.1	90.9 / 61.8 / 34.5	<b>100.0 / 99.9 / 99.3</b>	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	<b>100.0 / 100.0 / 100.0</b>	<b>100.0 / 100.0 / 100.0</b>	100.0 / 100.0 / 99.9
Diff-ProjectedGAN	93.8 / 69.5 / 43.3	88.8 / 54.6 / 27.2	<b>99.9 / 99.9 / 99.2</b>	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	<b>100.0 / 100.0 / 99.7</b>	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	<b>96.5 / 79.4 / 39.1</b>	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	<b>94.3 / 64.8 / 25.7</b>	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	<b>77.8 / 20.7 / 5.2</b>	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	<b>91.5 / 53.9 / 22.2</b>	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	<b>97.0 / 81.8 / 48.9</b>	54.8 / 7.7 / 2.1
Average	78.6 / 26.8 / 9.0	76.6 / 23.7 / 6.8	<b>91.4 / 59.3 / 25.7</b>	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<sup>(2/2)</sup>

		Fine-tuned on												
		ProGAN	StyleGAN	ProjectedGAN	Diff-StyleGAN2	Diff-ProjectedGAN	DDPM	IDDPM	ADM	PNDM	LDM	GAN	DM	All
Tested on	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												
		ProGAN	StyleGAN	ProjectedGAN	Diff-StyleGAN2	Diff-ProjectedGAN	DDPM	IDDPM	ADM	PNDM	LDM	GAN	DM	All
Tested on	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

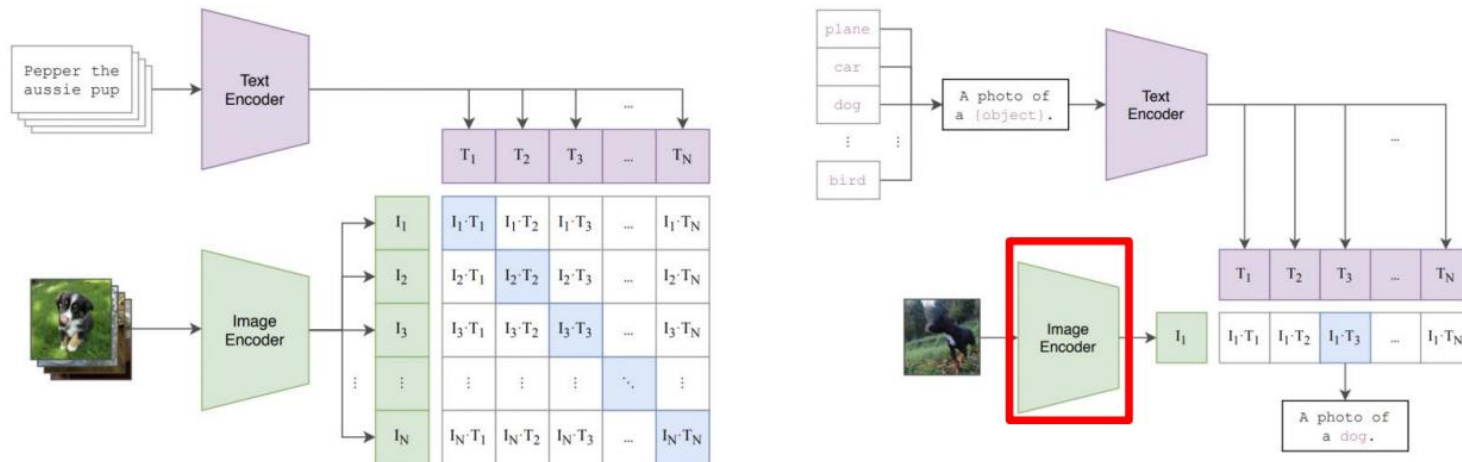
(b) PD@1%.

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



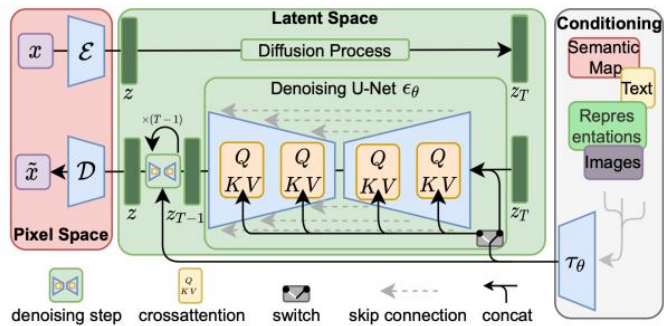
# Foundation Model

- “Any model that is **trained on broad data** (generally using self-supervision at scale) that can be adapted (e.g. fine-tuned) to a **wide range of downstream tasks**”.
- Foundation Models: BERT, GPT-n, **CLIP**, ...
- CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on 400 millions of (image, text) pairs.

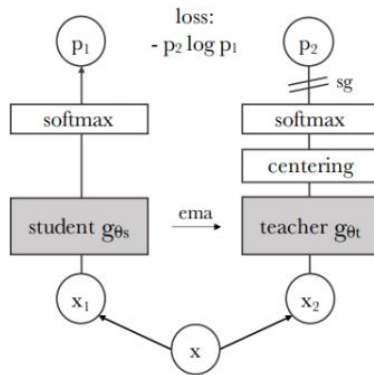


- The Center for Research on Foundation Models(CRFM), Stanford Institute for Human-Centered Artificial Intelligence's(HAI) coined the term "Foundation Model" in August 2021.
- Alec Radford, et al., "Learning transferable visual models from natural language supervision," International conference on machine learning (ICML), 2021.

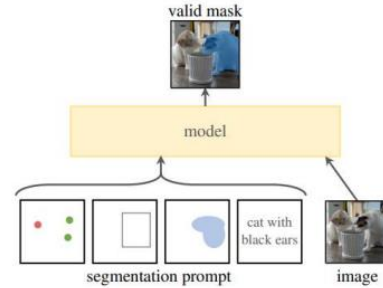
# Other Popular Vision Foundation Models



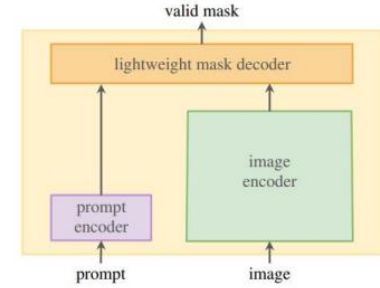
Stable Diffusion



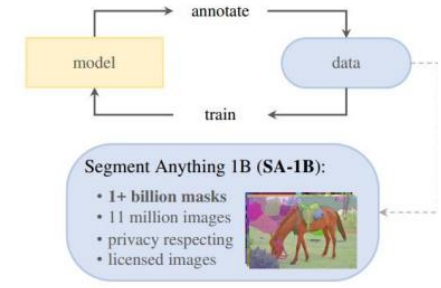
Dino/Dinov2



(a) Task: promptable segmentation



(b) Model: Segment Anything Model (SAM)



(c) Data: data engine (top) & dataset (bottom)

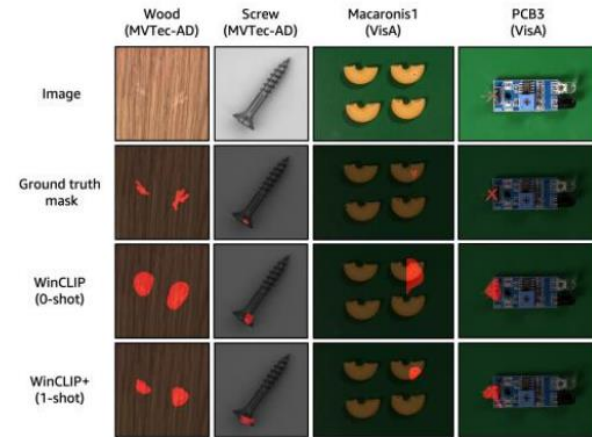
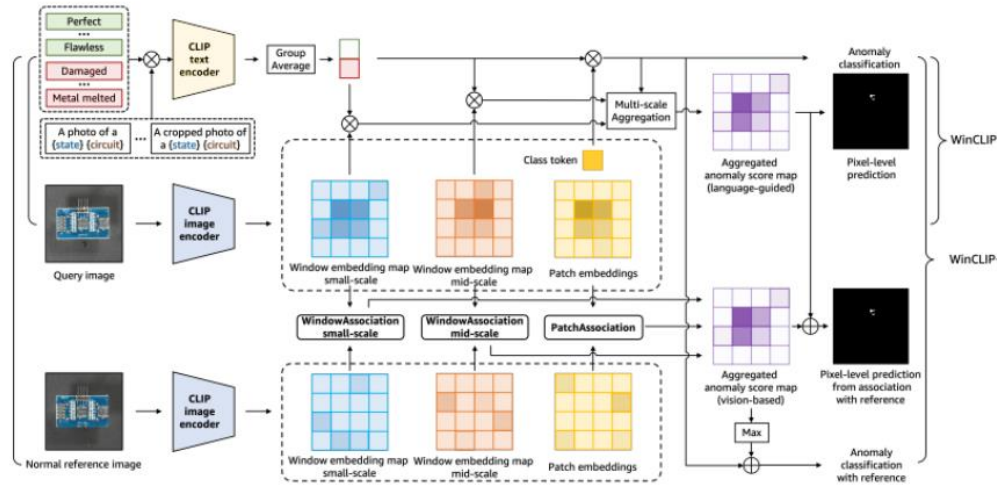
Segment Anything



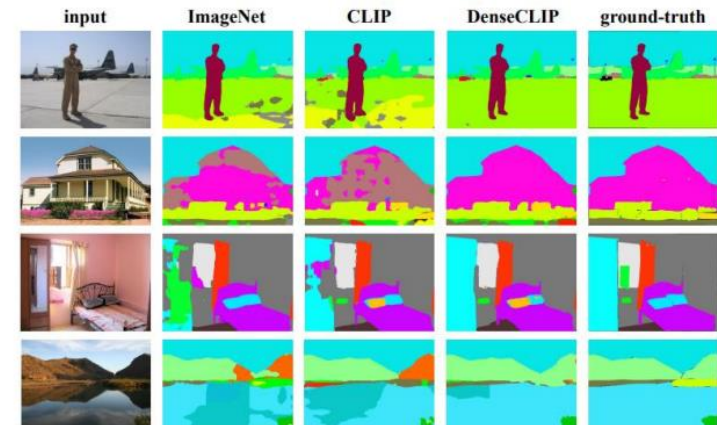
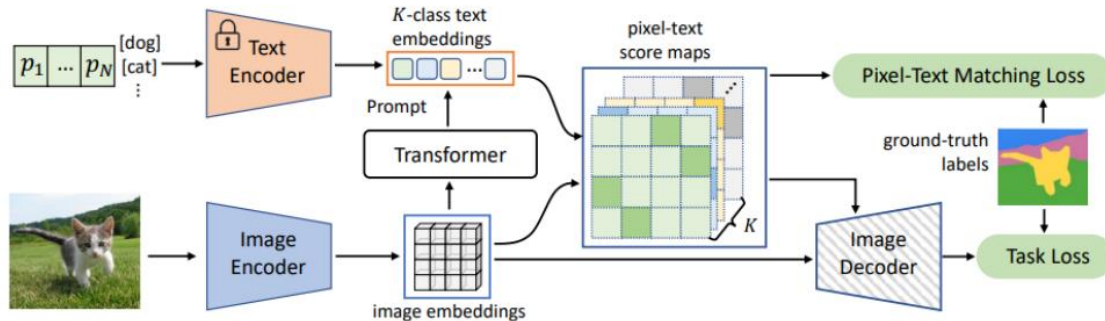
- Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models", CVPR. 2022.
- Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao et al. "Segment anything." ICCV 2023.
- Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez et al. "Dino v2: Learning robust visual features without supervision." arXiv preprint arXiv:2304.07193 (2023).

# Strong Generalization Capability to Many Downstream Tasks<sup>(1/2)</sup>

- WinCLIP



- DenseCLIP



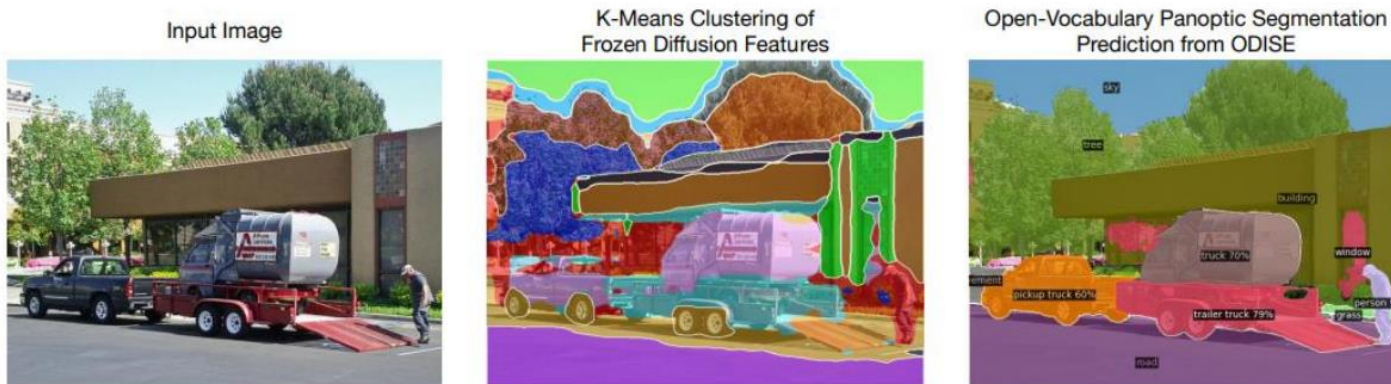
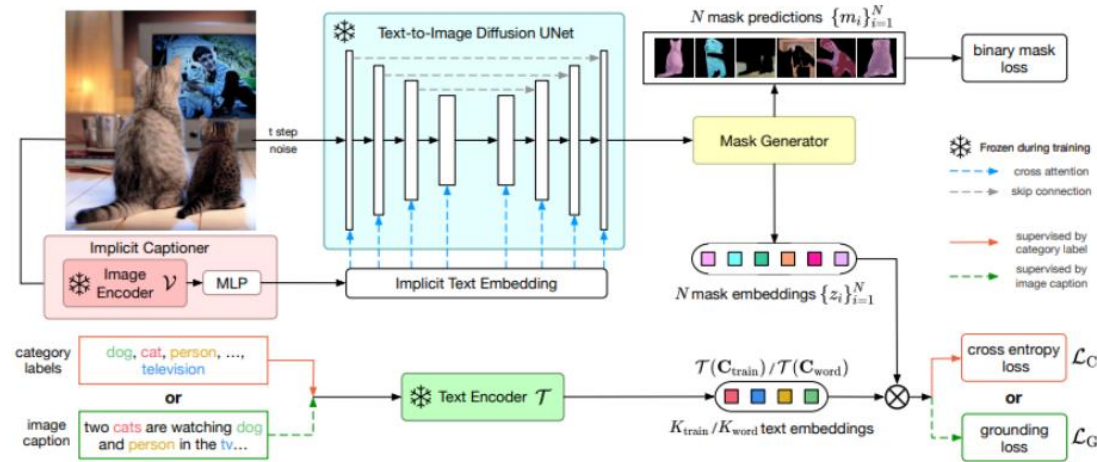
Jeong, Jongheon, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. "Winclip: Zero-/few-shot anomaly classification and segmentation." CVPR. 2023.  
 Rao, Yongming, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. "Denseclip: Language-guided dense prediction with context-aware prompting." CVPR 2022.





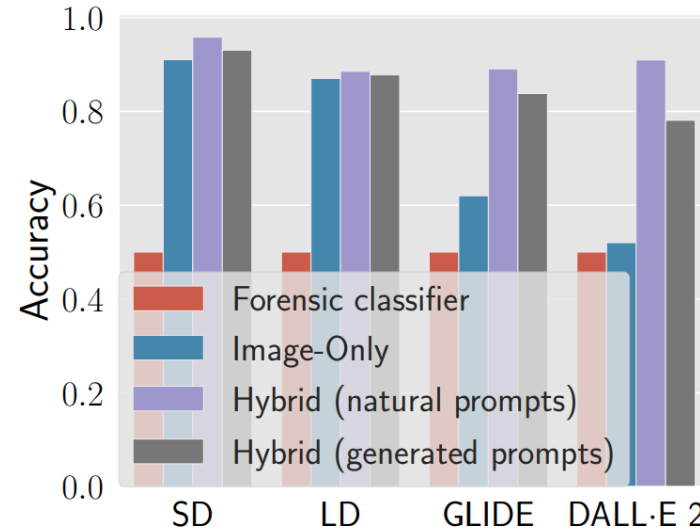
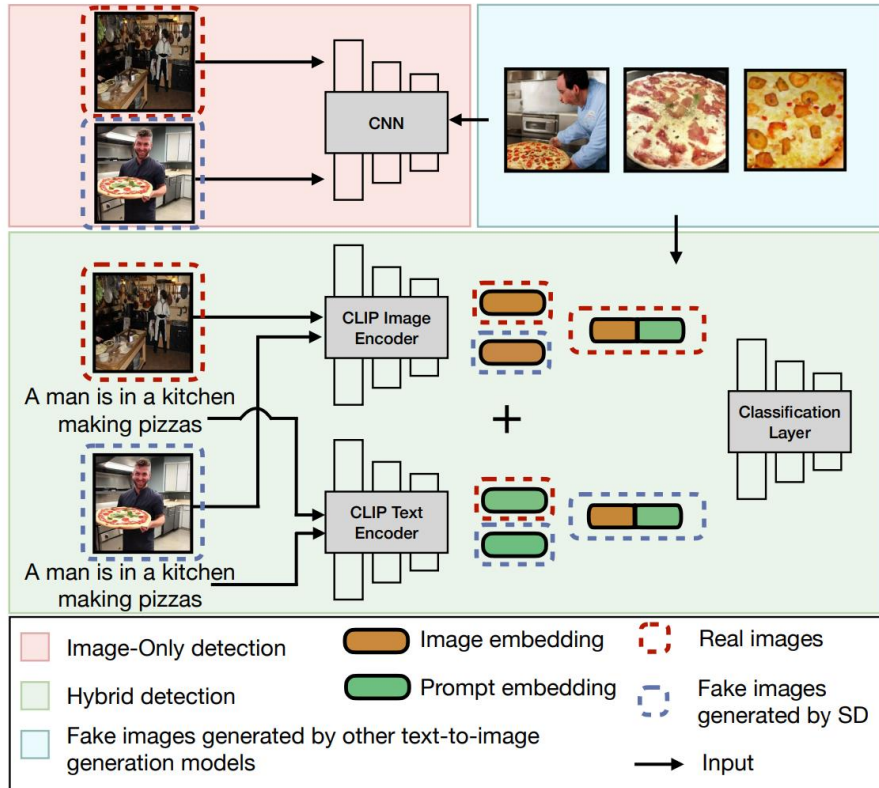
# Strong Generalization Capability to Many Downstream Tasks<sup>(2/2)</sup>

- Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models (ODISE)

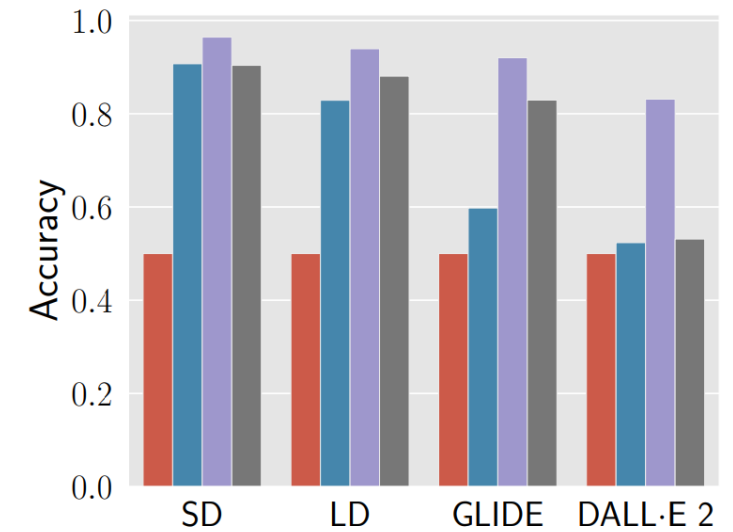


Xu, Jiarui, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. "Open-vocabulary panoptic segmentation with text-to-image diffusion models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2955-2966. 2023.

# DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models



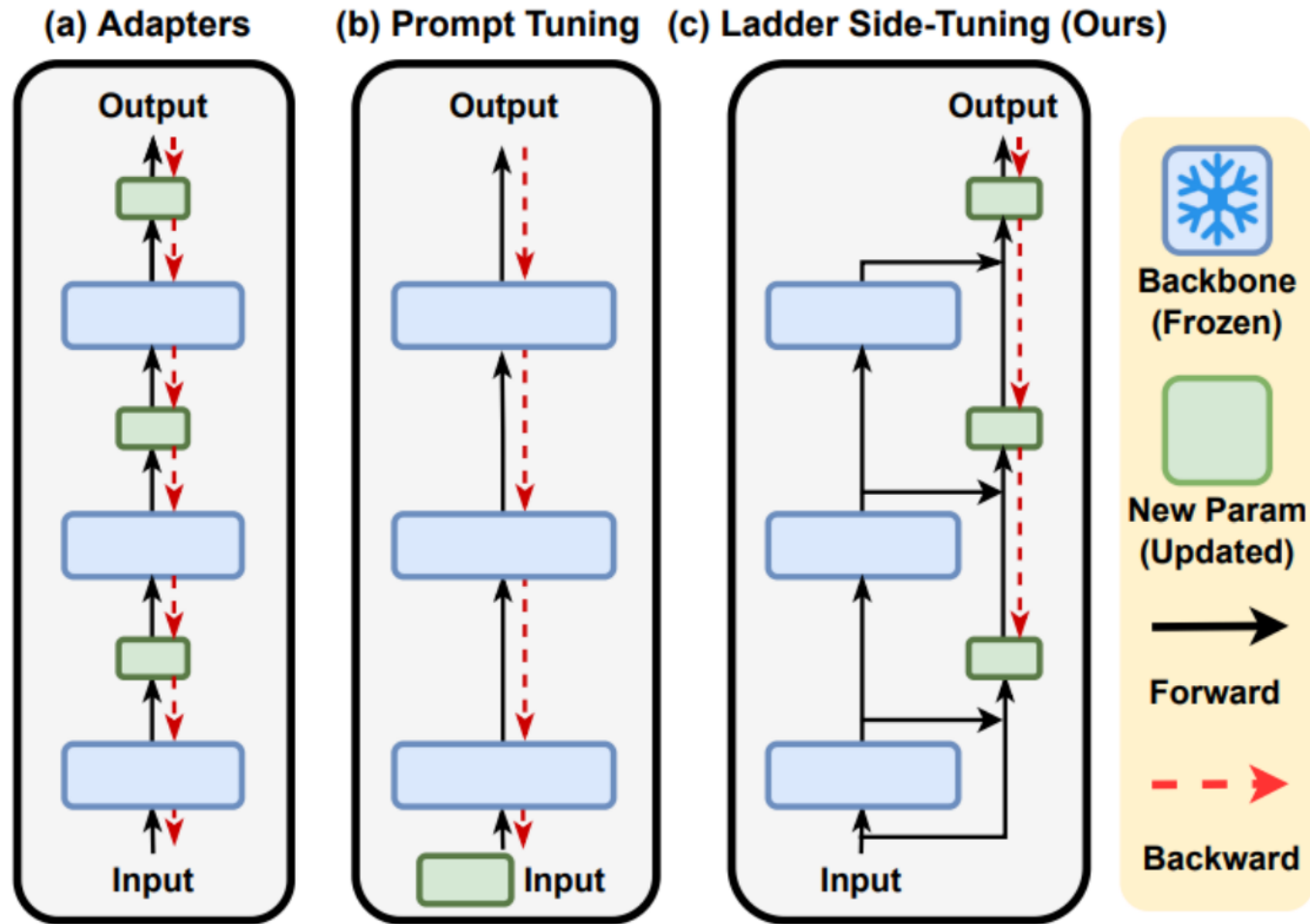
(a) MSCOCO



(b) Flickr30k

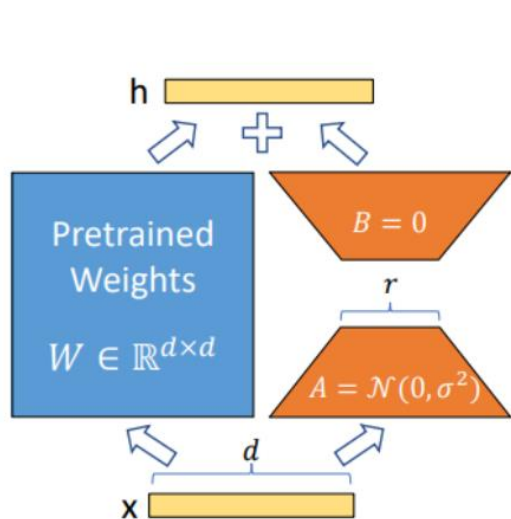
Sha, Zeyang, Zheng Li, Ning Yu, and Yang Zhang. "De-fake: Detection and attribution of fake images generated by text-to-image diffusion models." *ACM CCS 2023x`*

# Parameter Efficient Transfer Learning (PETL)

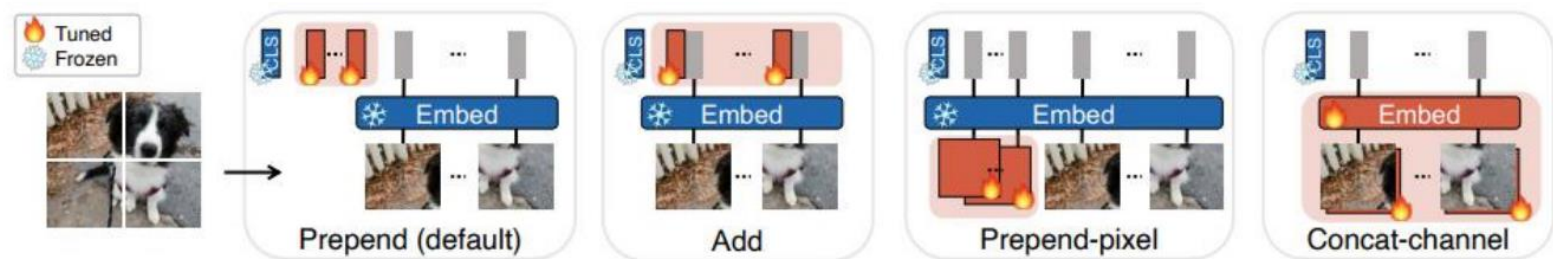
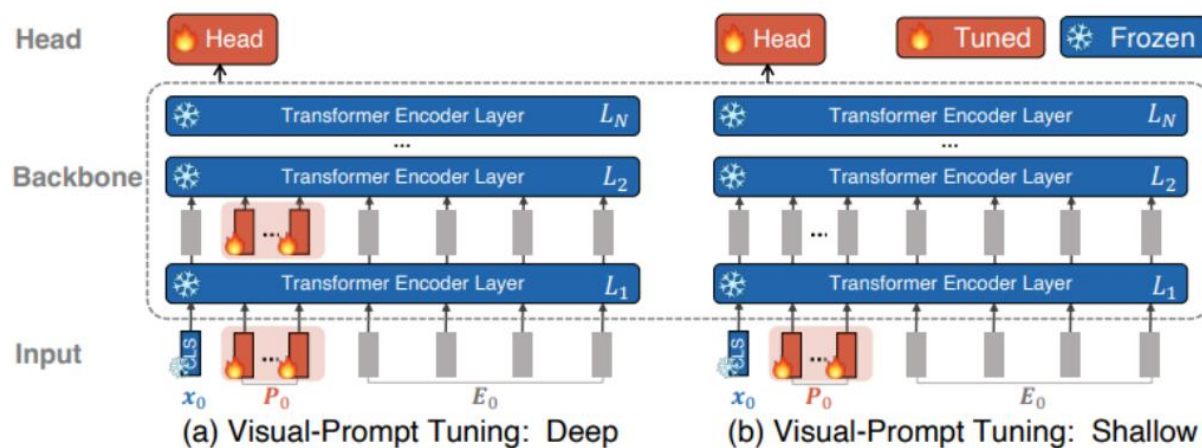
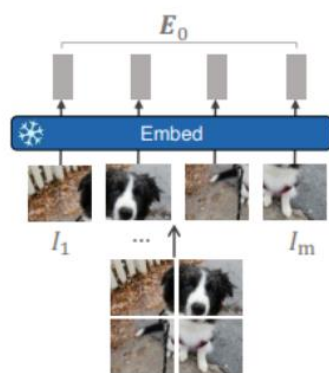


Sung, Yi-Lin, Jaemin Cho, and Mohit Bansal. "Lst: Ladder side-tuning for parameter and memory efficient transfer learning." Advances in Neural Information Processing Systems 35 (2022): 12991-13005.

# Parameter Efficient Transfer Learning (PETL)



Adapter (LORA)

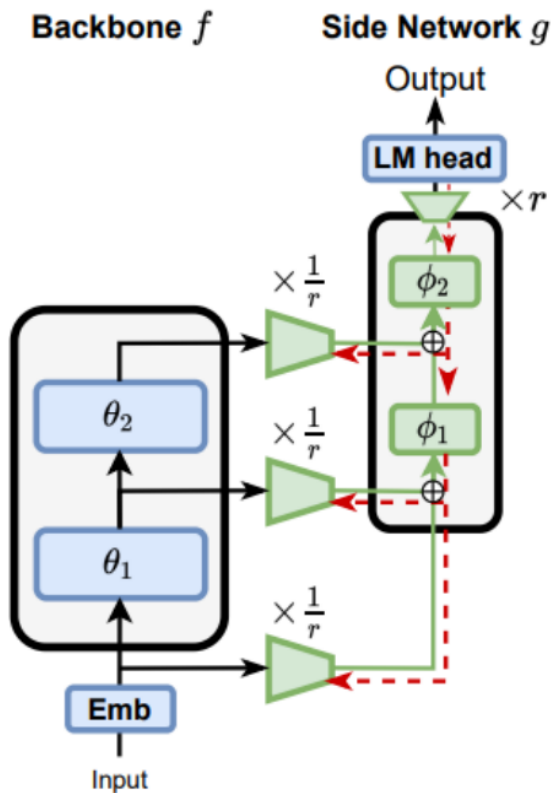


## Visual Prompt Tuning

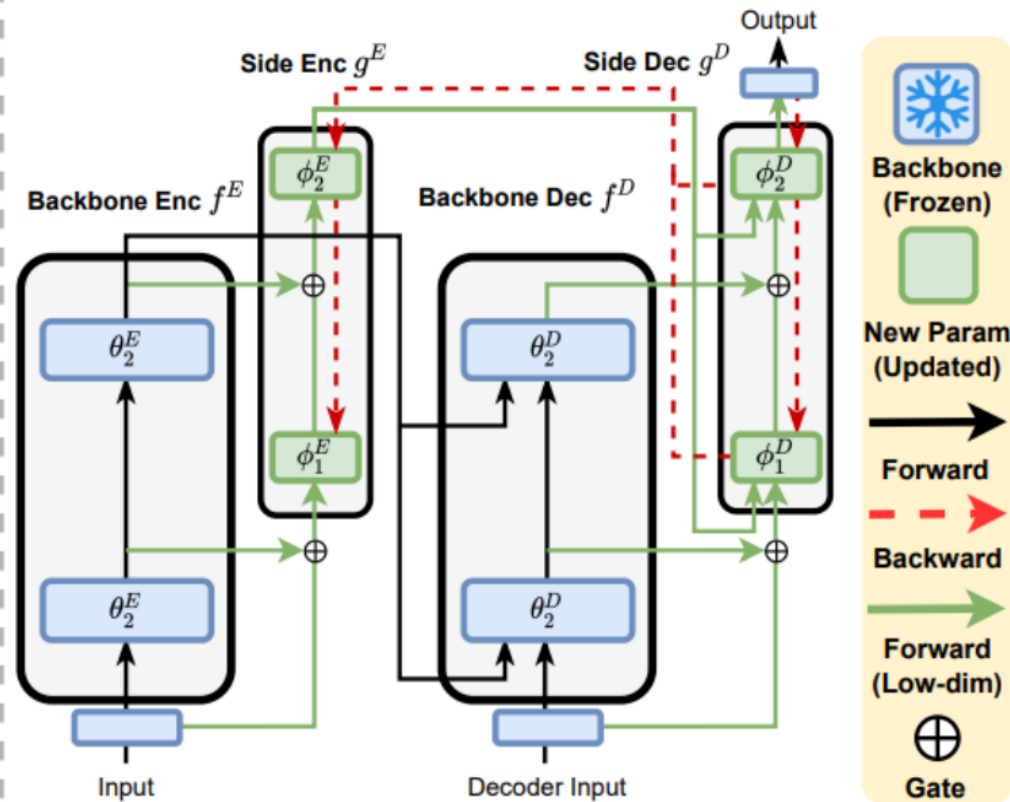
- Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." *arXiv preprint arXiv:2106.09685* (2021).
- Jia, Menglin, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. "Visual prompt tuning." ECCV 2022.

# Ladder Side Tuning

(a) Ladder Side Network

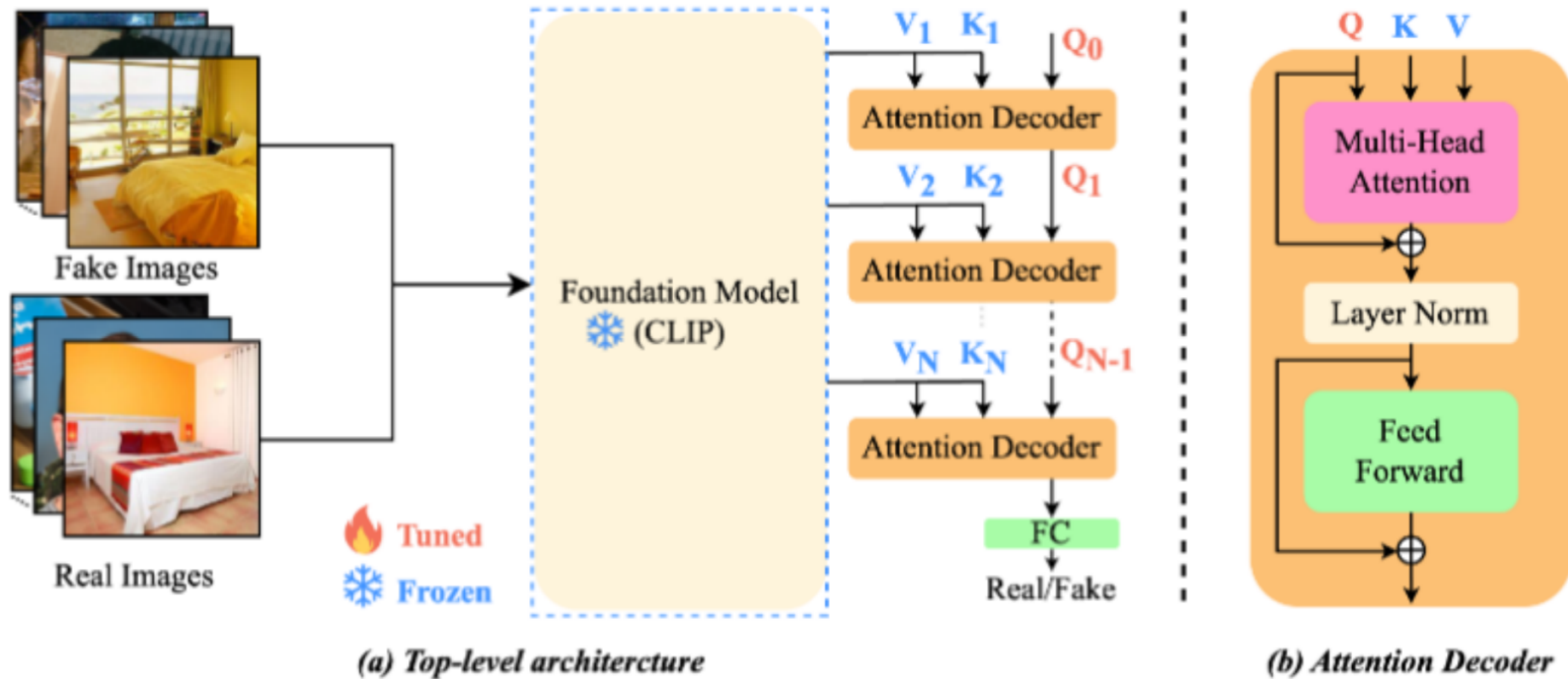


(b) Ladder Side Network for Encoder-Decoder



Sung, Yi-Lin, Jaemin Cho, and Mohit Bansal. "Lst: Ladder side-tuning for parameter and memory efficient transfer learning." Advances in Neural Information Processing Systems 35 (2022): 12991-13005.

# Deepfake Detection through Foundation Model



# Deepfake Detection through Foundation Model

Detection method	Training dataset	Generation model	Testing DMs					Testing GANs					Total Avg.
			ADM	DDPM	IDDPM	PNDM	LDM	Pro-GAN	Style-GAN	Projected-GAN	Diff-StyleGAN2	Diff-ProjectedGAN	
CNNDet [4]	LSUN	ProGAN	50.1/66.4	50.3/82.5	50.1/78.9	50.1/77.5	50.2/75.9	99.7/100	59.2/97.1	52.6/92.6	80.9/99.7	51.6/91.2	59.5/86.2
GANDet [30]	LSUN	ProGAN	50.0/61.2	50.0/59.0	50.1/64.6	50.4/73.4	50.1/56.9	54.4/89.9	51.4/90.0	50.1/58.5	95.1/99.6	50.3/62.8	55.2/71.6
SBI [32]	FF++	Multiple	49.5/49.3	50.2/50.0	50.7/50.6	49.6/49.9	50.2/50.2	50.0/50.2	49.8/50.2	49.4/49.5	50.1/50.0	50.8/50.9	50.0/50.1
TwoStream [33]	FF++	Multiple	50.0/49.9	50.0/52.3	50.0/50.9	50.0/56.2	50.0/50.3	50.0/53.4	50.0/53.0	50.0/47.9	50.0/52.9	50.0/49.2	50.0/51.6
CNNDet* (Tuning)	LSUN-B	ADM	94.3/99.4	95.0/99.6	98.0/99.9	97.5/99.6	90.0/98.9	98.6/99.8	86.2/97.0	70.9/88.6	70.2/86.8	68.6/88.6	86.9/95.8
CNNDet* (Re-train)	LSUN-B	ADM	92.8/99.4	93.3/99.5	98.4/99.9	95.4/99.8	85.4/98.7	94.3/98.8	82.5/97.5	63.5/82.6	61.5/87.4	61.8/85.6	82.9/94.9
AdaptCLIP (ours)	LSUN-B.	ADM	<b>99.2/100</b>	<b>100/100</b>	<b>99.7/100</b>	<b>99.9/100</b>	<b>99.8/100</b>	<b>100/100</b>	<b>99.8/100</b>	<b>99.7/100</b>	<b>99.9/100</b>	<b>99.8/100</b>	<b>99.8/100</b>
	LSUN-B.	IDDPM	98.5/100	99.6/100	99.6/100	99.8/100	99.8/100	99.7/100	99.5/100	99.4/100	99.8/100	99.2/99.9	99.5/99.9
	LSUN-B.	LDM	72.5/91.3	94.0/99.1	87.4/97.5	99.2/100	99.3/100	99.5/100	98.9/99.9	98.2/99.8	99.4/100	98.6/99.9	94.7/98.8

Tested on	Trained on						
	ADM*	LDM	ProjectedGAN	Diff-StyleGAN2	StyleGAN	StyleGAN2	StyleGAN3
ADM*	99.9	50.9	51.9	69.0	55.7	58.6	84.2
LDM	71.3	100.0	53.8	55.2	55.4	78.3	68.3
ProjectedGAN	55.8	50.6	100.0	97.9	90.7	82.4	83.6
Diff-StyleGAN2	60.4	50.1	72.0	99.9	87.1	77.7	95.8
StyleGAN	60.8	50.4	82.3	99.1	100.0	96.8	99.2
StyleGAN2	81.2	61.1	83.7	99.3	99.5	100.0	98.8
StyleGAN3	58.5	50.1	52.0	80.3	71.0	58.6	99.6

(a) ACC

Tested on	Trained on						
	ADM*	LDM	ProjectedGAN	Diff-StyleGAN2	StyleGAN	StyleGAN2	StyleGAN3
ADM*	100.0	96.4	87.6	97.7	92.1	96.6	98.6
LDM	97.1	100.0	91.4	87.7	91.6	99.5	92.3
ProjectedGAN	87.9	88.8	100.0	100.0	99.7	99.0	97.7
Diff-StyleGAN2	95.7	83.3	98.8	100.0	99.6	99.2	99.8
StyleGAN	95.0	91.8	99.5	100.0	100.0	100.0	100.0
StyleGAN2	99.4	99.4	99.7	100.0	100.0	100.0	99.9
StyleGAN3	93.1	85.0	91.2	99.1	97.9	95.4	100.0

(b) AUROC

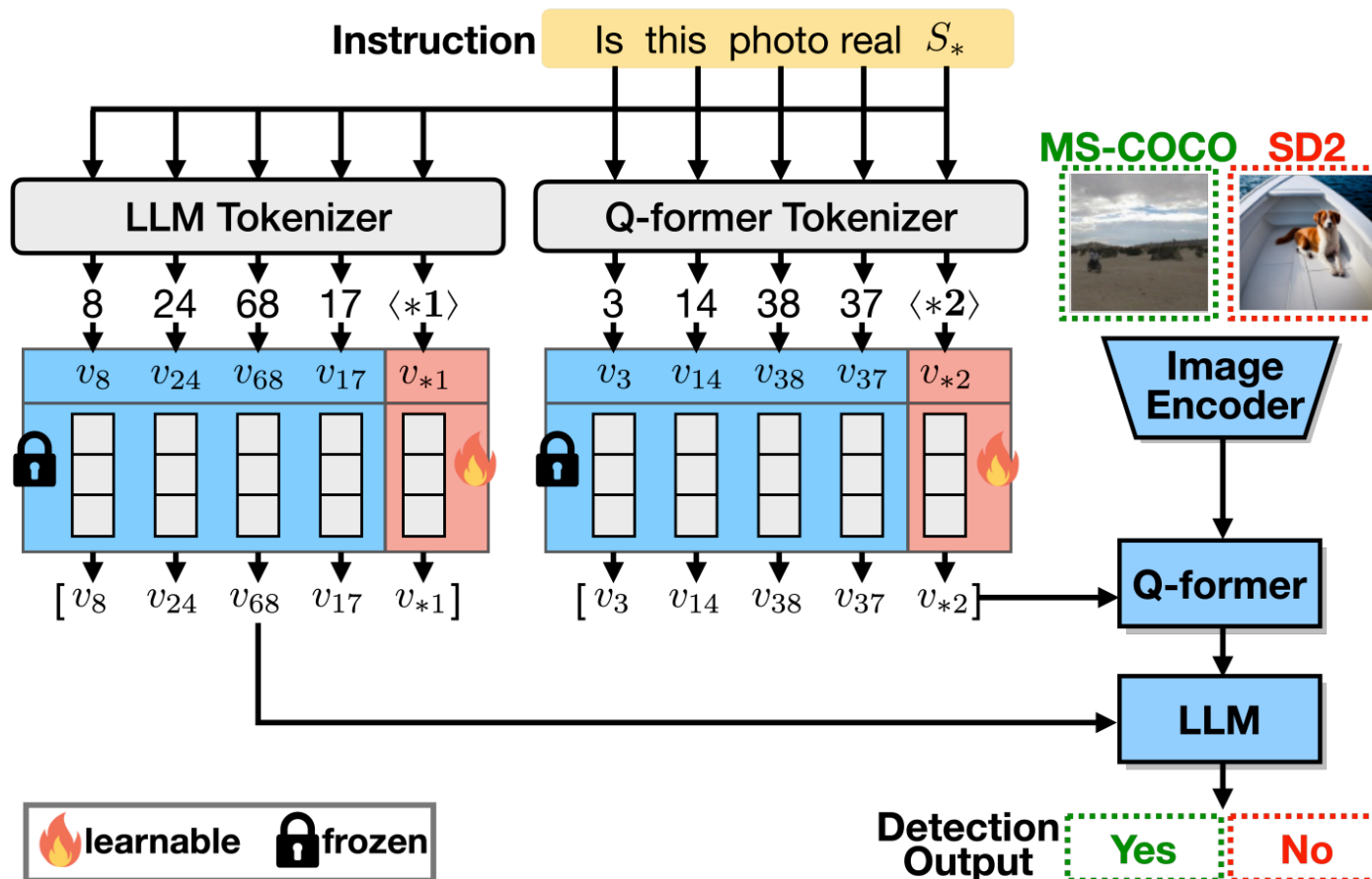
FFHQ

Adaptation Method	LSUN-B (Text2Image)		FFHQ	MSCOCO
	ACC	AUC		
(a) CNNDet*(Tuning)	56.2	60.3	50.4/63.4	47.6/45.5
(b) LinearCLIP	76.6	85.3	51.8/62.2	50.2/51.8
(c) AdaptCLIP (ours)	<b>99.7</b>	<b>100</b>	<b>60.8/77.2</b>	<b>57.4/60.9</b>

ACC(%)/AUC(%) scores



# More Deepfake Detection through Foundation Model



Chang, You-Ming, Chen Yeh, Wei-Chen Chiu, and Ning Yu. "AntifakePrompt: Prompt-Tuned Vision-Language Models are Fake Image Detectors." *arXiv preprint arXiv:2310.17419* (2023).



# More Deepfake Detection through Foundation Model

Methods	Training set	No. of param.	MS COCO	Flickr	SD2	SDXL	IF	DALLE-2
Wange-2020	ImageNet vs. ProGAN	23.51M	96.87	96.67	0.17	0.17	19.17	3.40
DE-FAKE	MS COCO vs. SD2	308.02M	85.97	90.67	97.10	90.50	<b>99.20</b>	68.97
DIRE	LSUN Bedroom vs. StyleGAN	23.51M	81.77	77.53	3.83	18.17	6.93	2.13
InstructBLIP	-	188.84M	<b>98.93</b>	<b>99.63</b>	40.27	23.07	20.63	41.77
InstructBLIP + LoRA	MS COCO vs. SD2	4.19M	95.73	91.83	<b>98.03</b>	96.33	86.60	<b>99.57</b>
AntifakePrompt	MS COCO vs. SD2	<b>4.86K</b>	95.37	91.00	97.83	<b>97.27</b>	89.73	<b>99.57</b>
	MS COCO vs. SD2+LaMa	<b>4.86K</b>	90.83	81.04	97.10	97.10	88.37	99.07

Methods	Training set	No. of param.	SGXL	ControlNet	Inpainting		Super Res.	
					LaMa	SD2	LTE	SD2
Wange-2020	ImageNet vs. ProGAN	23.51M	79.30	11.43	7.53	0.17	15.27	1.40
DE-FAKE	MS COCO vs. SD2	308.02M	56.90	63.97	13.03	16.00	9.97	29.70
DIRE	LSUN Bedroom vs. StyleGAN	23.51M	45.27	9.90	13.23	11.37	12.53	2.77
InstructBLIP	-	188.84M	69.53	33.97	10.90	44.23	97.23	69.10
InstructBLIP + LoRA	MS COCO vs. SD2	4.19M	97.67	92.87	<b>59.50</b>	<b>93.03</b>	99.53	<b>99.97</b>
AntifakePrompt	MS COCO vs. SD2	<b>4.86K</b>	<b>99.97</b>	91.47	39.03	85.20	99.90	99.93
	MS COCO vs. SD2+LaMa	<b>4.86K</b>	99.93	<b>93.27</b>	58.53	90.70	<b>100.00</b>	<b>99.97</b>

Methods	Training set	No. of param.	Deeper-Forensics	Adver.	Attack Backdoor	Data Poisoning	Average
Wange-2020	ImageNet vs. ProGAN	23.51M	0.30	4.93	15.50	0.97	22.08
DE-FAKE	MS COCO vs. SD2	308.02M	86.97	60.40	22.23	55.87	59.22
DIRE	LSUN Bedroom vs. StyleGAN	23.51M	0.27	1.60	1.93	1.00	18.14
InstructBLIP	-	188.84M	13.83	5.50	3.17	1.60	42.09
InstructBLIP + LoRA	MS COCO vs. SD2	4.19M	<b>98.80</b>	64.30	53.40	50.87	86.13
AntifakePrompt	MS COCO vs. SD2	<b>4.86K</b>	97.90	96.70	93.00	91.57	91.59
	MS COCO vs. SD2+LaMa	<b>4.86K</b>	97.77	<b>97.20</b>	<b>97.10</b>	<b>93.63</b>	<b>92.60</b>

Chang, You-Ming, Chen Yeh, Wei-Chen Chiu, and Ning Yu. "AntifakePrompt: Prompt-Tuned Vision-Language Models are Fake Image Detectors." *arXiv preprint arXiv:2310.17419* (2023).

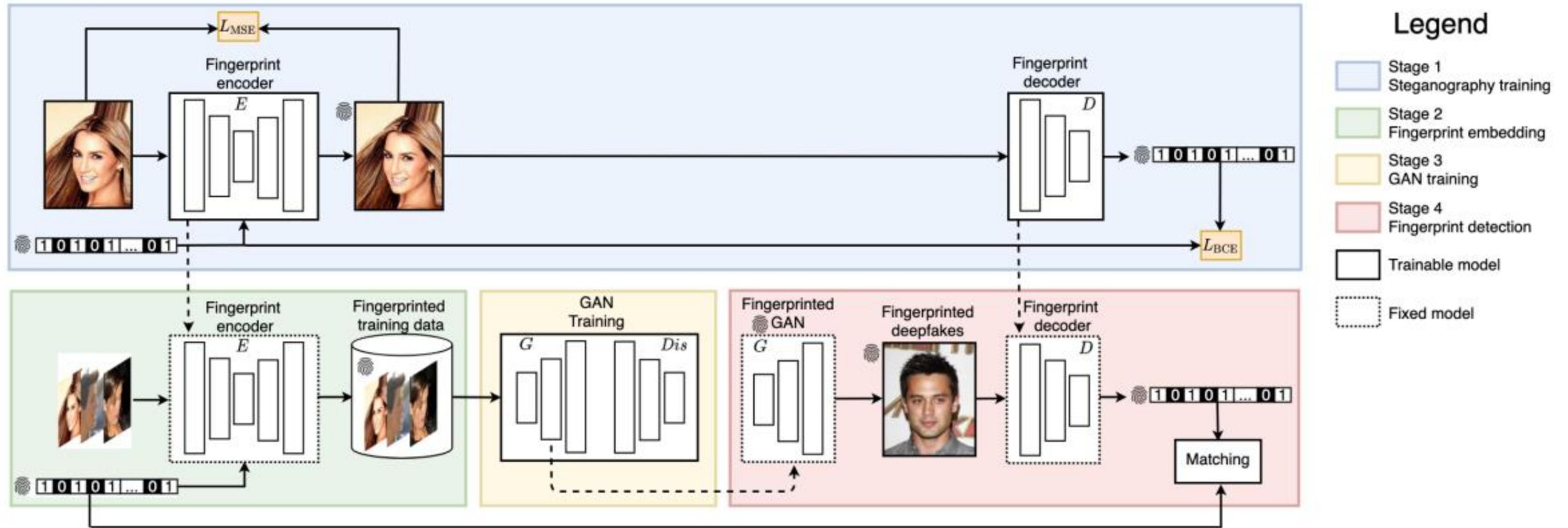


# Possible Countermeasures

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

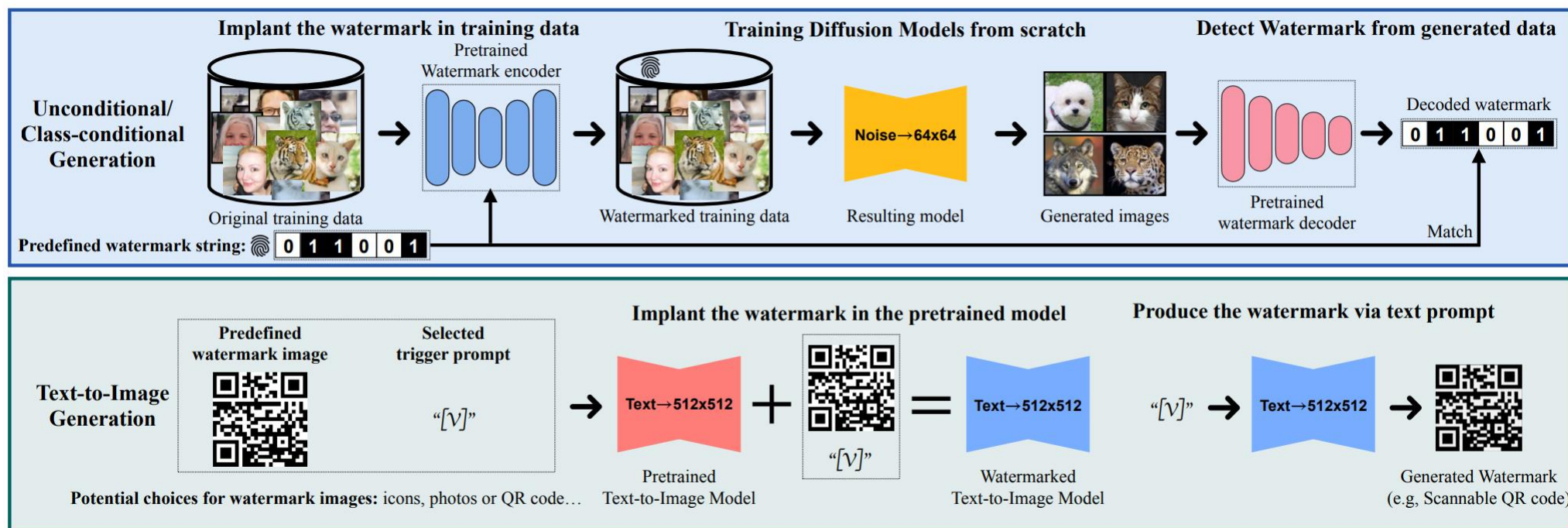


# Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data



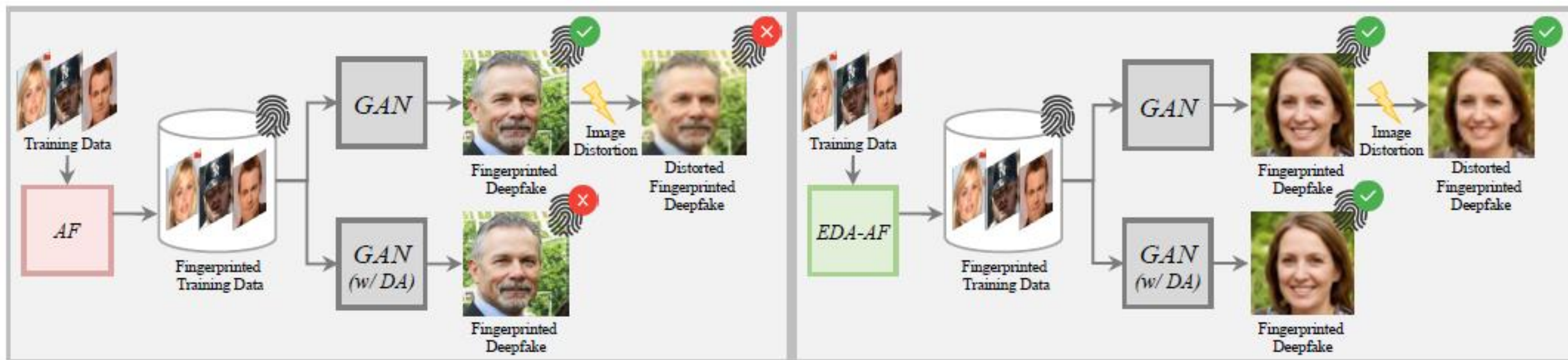
Yu, Ning, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. "Artificial fingerprinting for generative models: Rooting deepfake attribution in training data." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14448-14457. 2021.

# A Recipe for Watermarking Diffusion Models

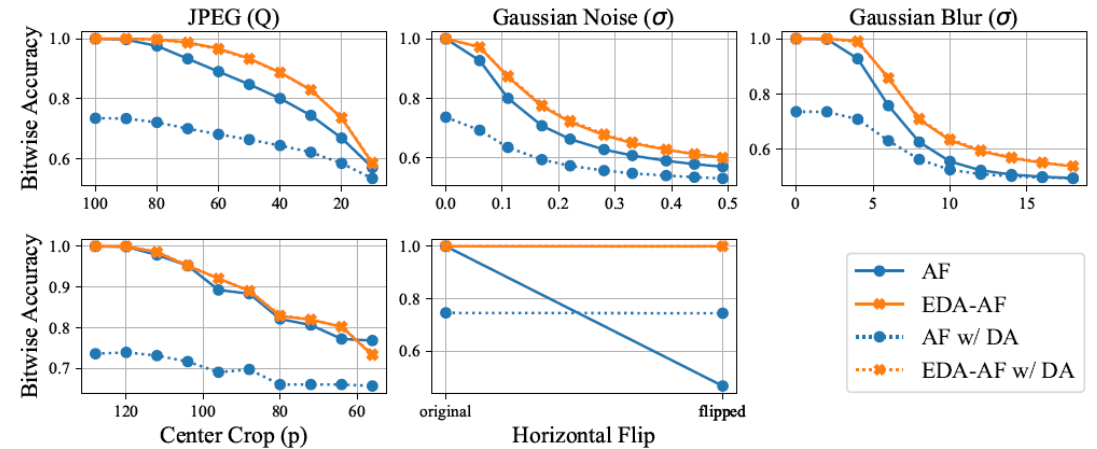
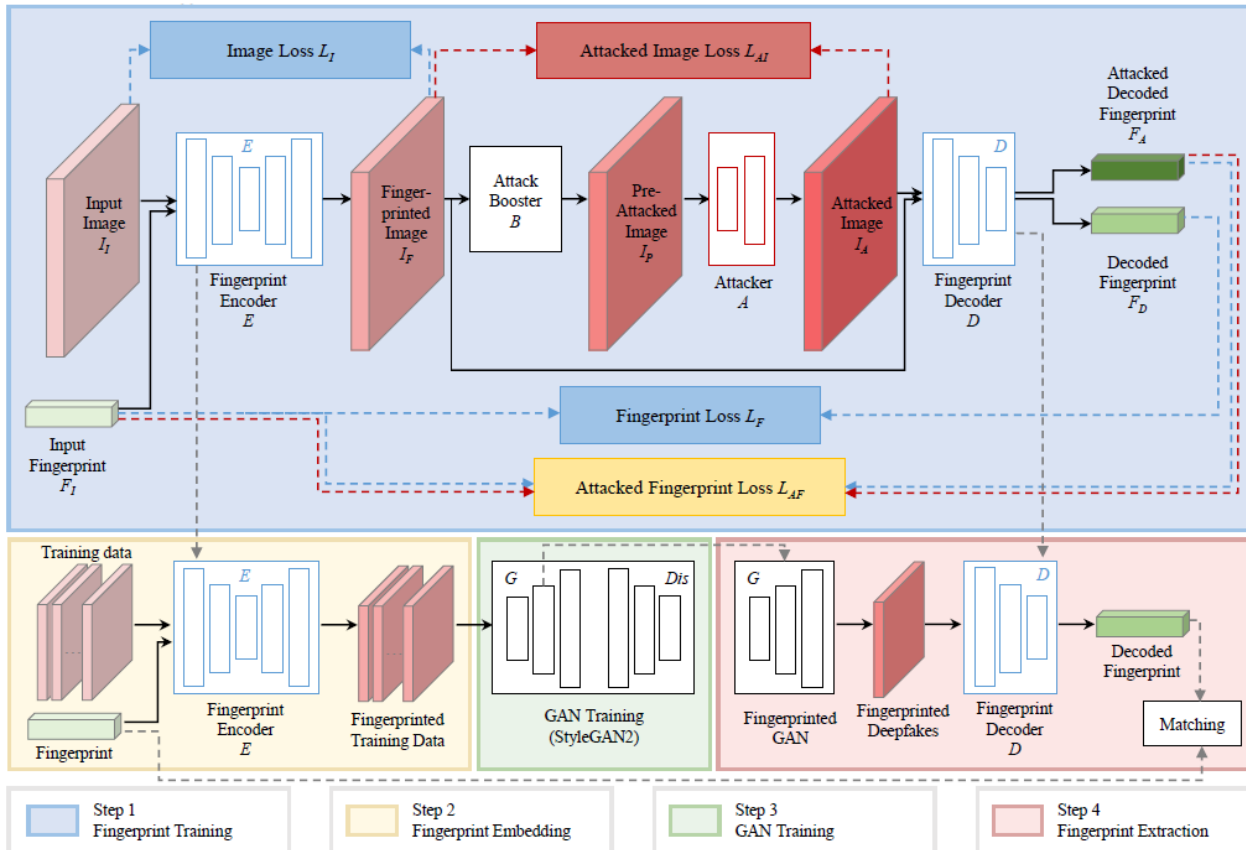


Zhao, Yunqing, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. "A recipe for watermarking diffusion models." arXiv preprint arXiv:2303.10137 (2023).

# 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)



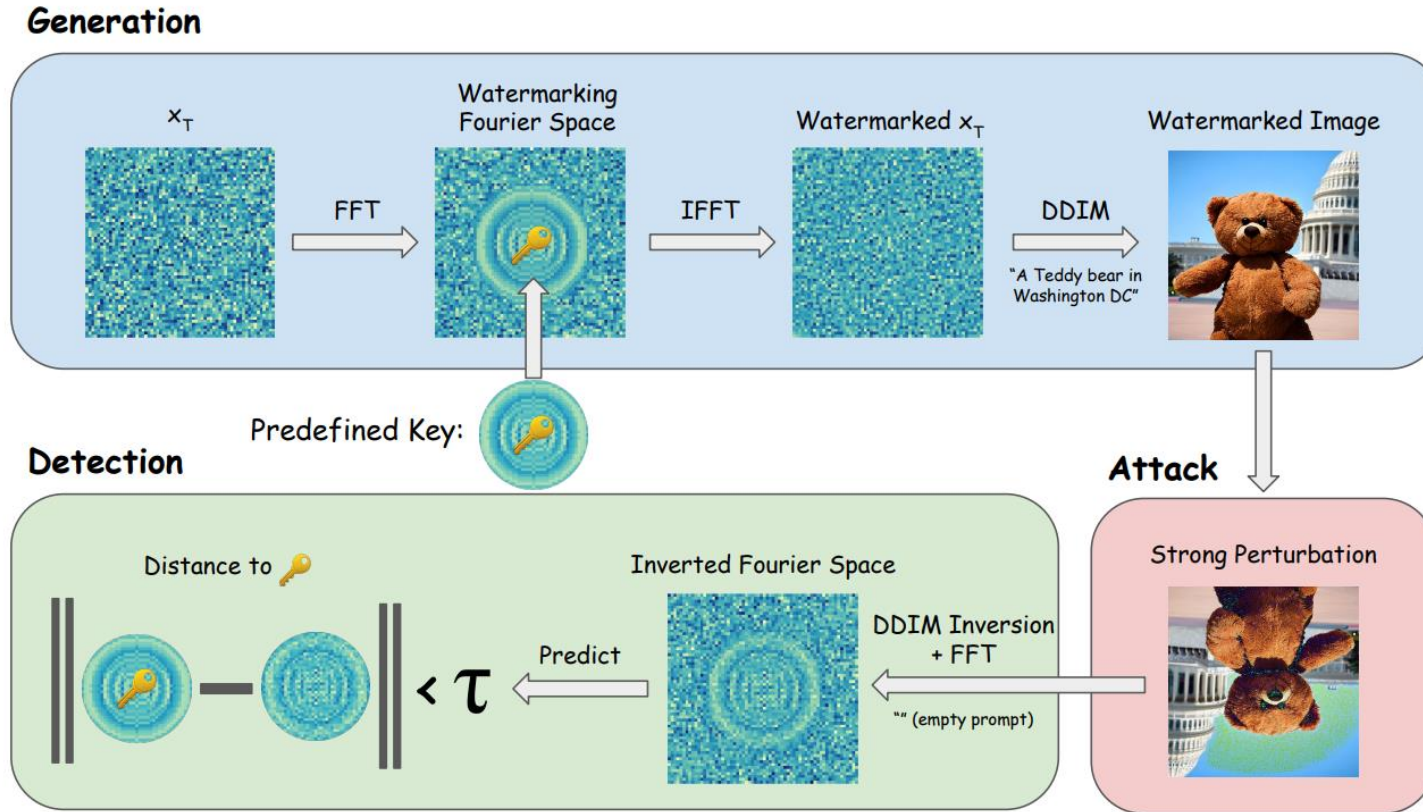
Method	Bit Acc $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$
AF [1]	0.74	0.967	36.938
EDA-AF	0.99	0.961	35.577

Dataset	Model	Method	Bit Acc $\uparrow$	FID $\downarrow$
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.



# Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust



Wen, Yuxin, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. "Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust." *arXiv preprint arXiv:2305.20030* (2023).

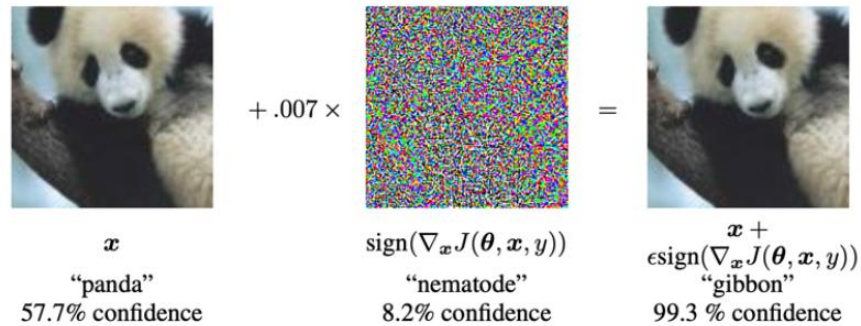
# Possible Countermeasures

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

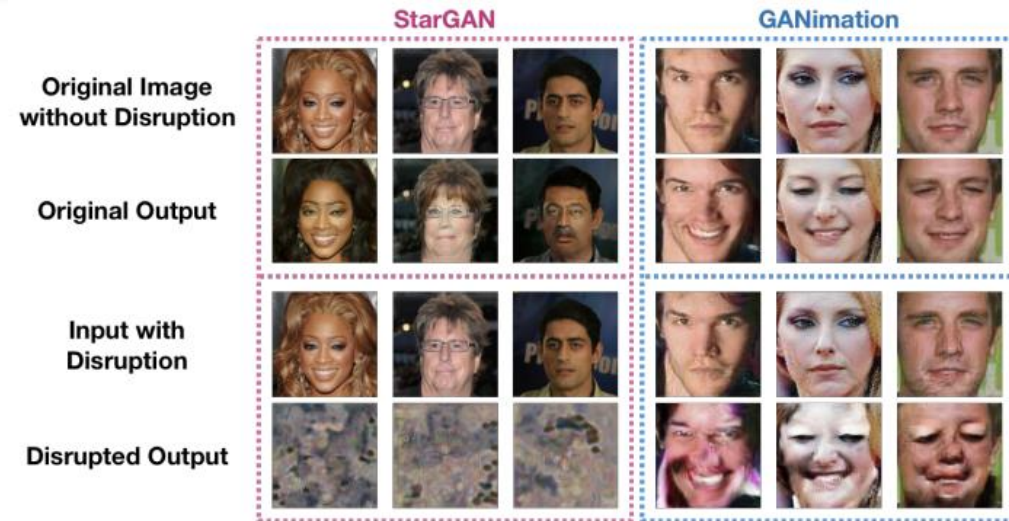
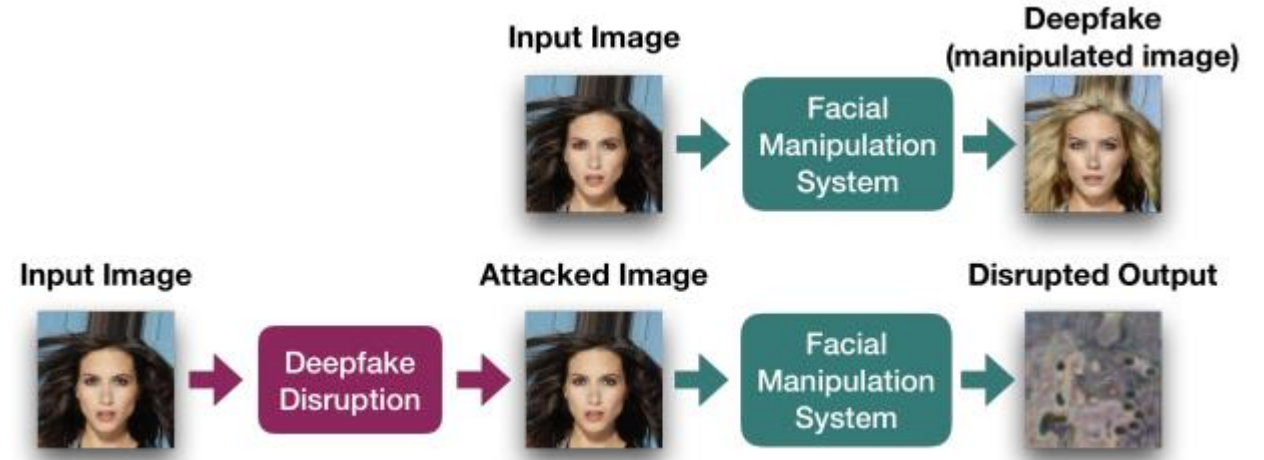




# Disrupting Deepfakes: Adversarial Attacks Against Conditional Image Translation Networks and Facial Manipulation Systems

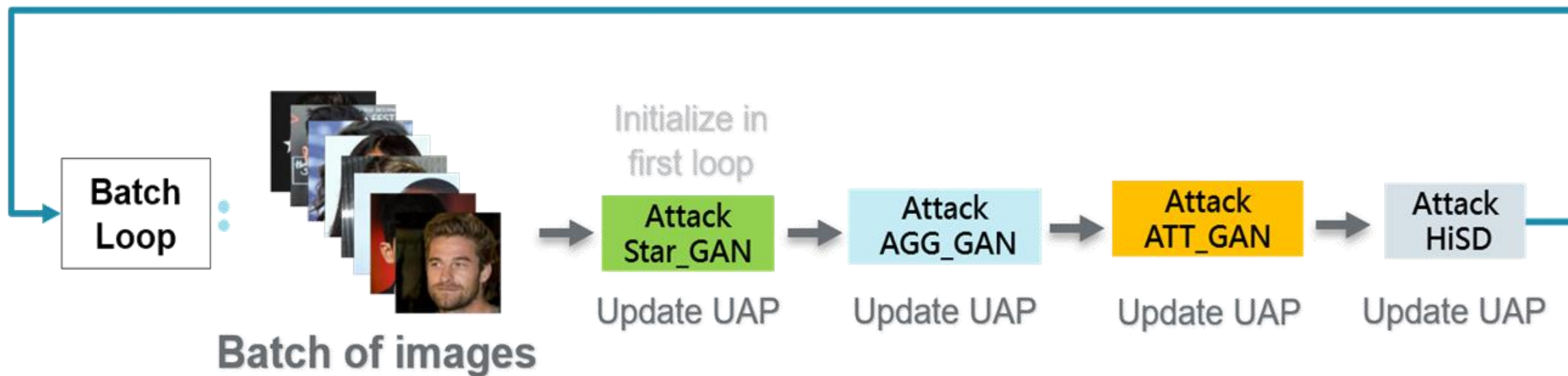


[Goodfellow et al. 2015]



[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

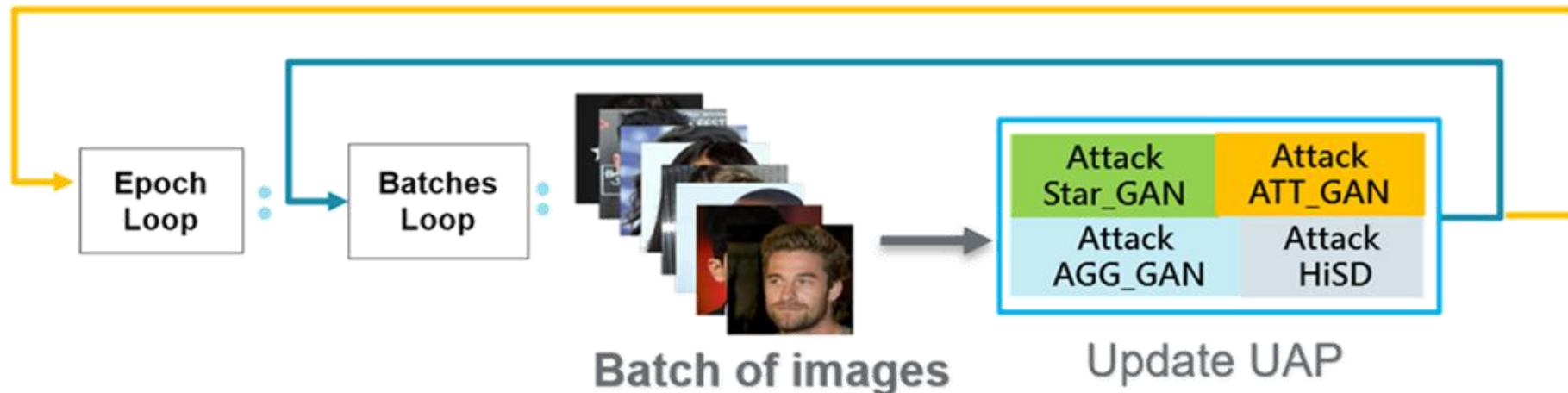


TABLE I: The evaluation results of cross-model UAP.

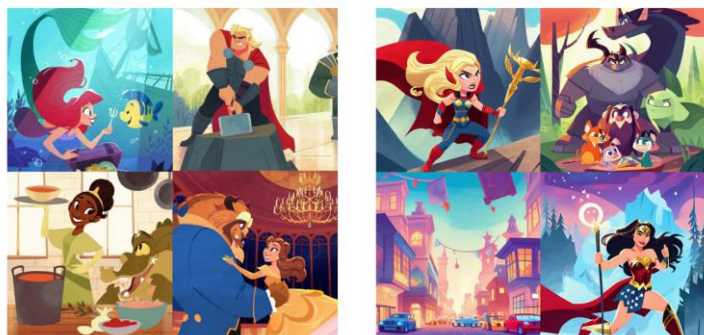
Attacks \ DGMs	StarGAN		AggGAN		ATTGAN		HiSD	
	ASR	$\ell_2$	ASR	$\ell_2$	ASR	$\ell_2$	ASR	$\ell_2$
CMUA_v1	<b>100.0%</b>	0.457	99.2%	0.107	20.2%	0.037	N/A <sup>1</sup>	N/A <sup>1</sup>
CMUA_v2	<b>100.0%</b>	0.199	<b>100.0%</b>	0.128	95.3%	0.066	<b>100.0%</b>	0.108
CMUA_12	97.0%	0.081	<b>100.0%</b>	0.046	86.1%	0.047	<b>100.0%</b>	0.107
Ours	<b>100.0%</b>	<b>0.766</b>	<b>100.0%</b>	<b>0.133</b>	<b>98.2%</b>	<b>0.124</b>	<b>100.0%</b>	<b>0.113</b>

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

	PSNR $\uparrow$	SSIM $\uparrow$	Time	Batch size
CMUA_v2	32.4939	0.7821	N/A <sup>2</sup>	64
CMUA_12	32.2755	0.7736	800s	12
Ours	<b>33.4651</b>	<b>0.8186</b>	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.

# GLAZE



Original artwork  
by Hollie Mengert

Mimicked artwork  
in Hollie's style

## Artist (V)

Original artwork



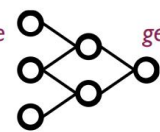
Cloaked artwork



## Mimic



*fine-tune*



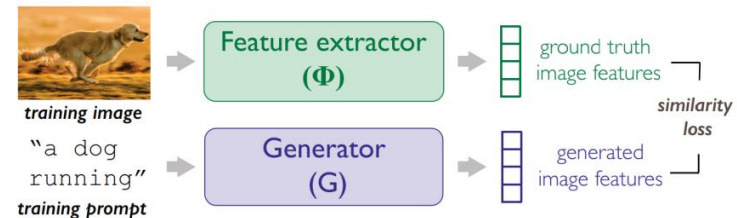
Style-specific  
model

*generate*

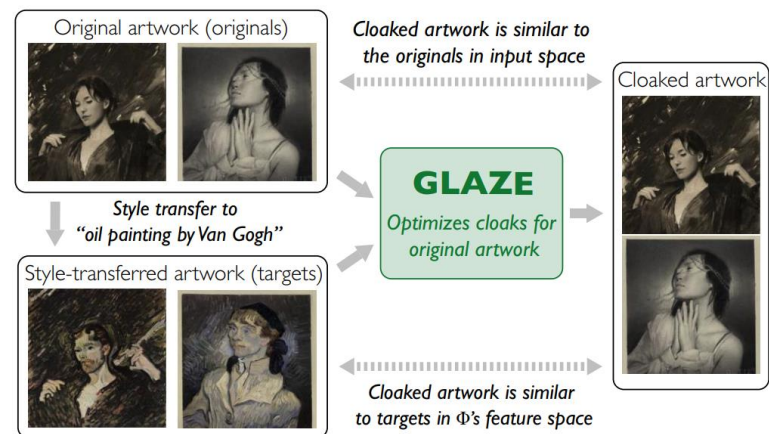
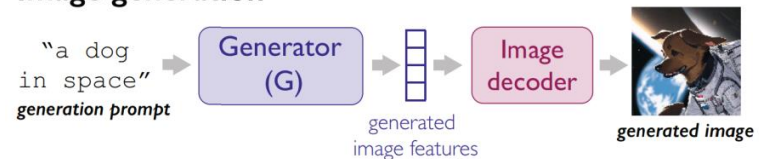


Fails to mimic  
victim artist

## Model training



## Image generation



a) Style transfer

b) Cloak optimization

Shan, Shawn, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. "Glaze: Protecting artists from style mimicry by text-to-image models." arXiv preprint arXiv:2302.04222 (2023).

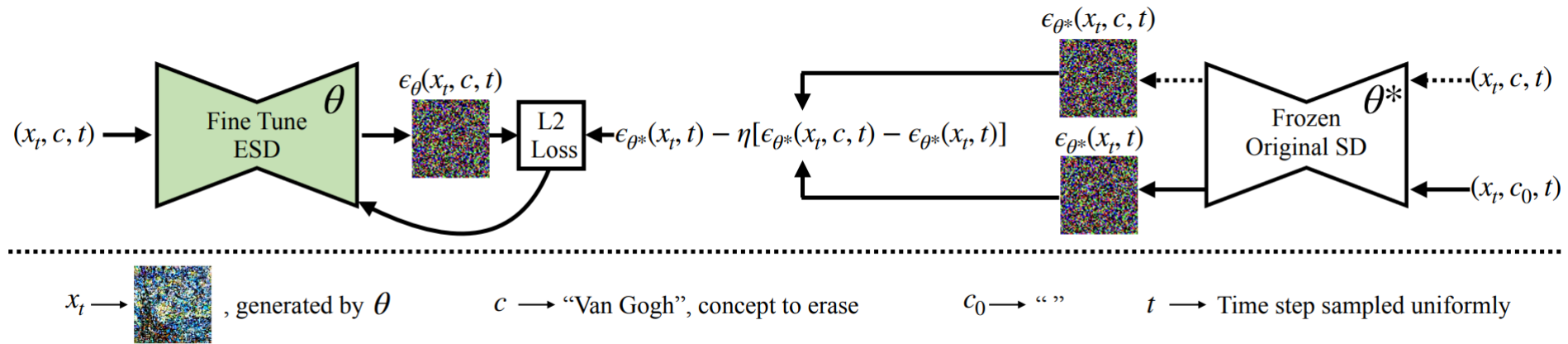
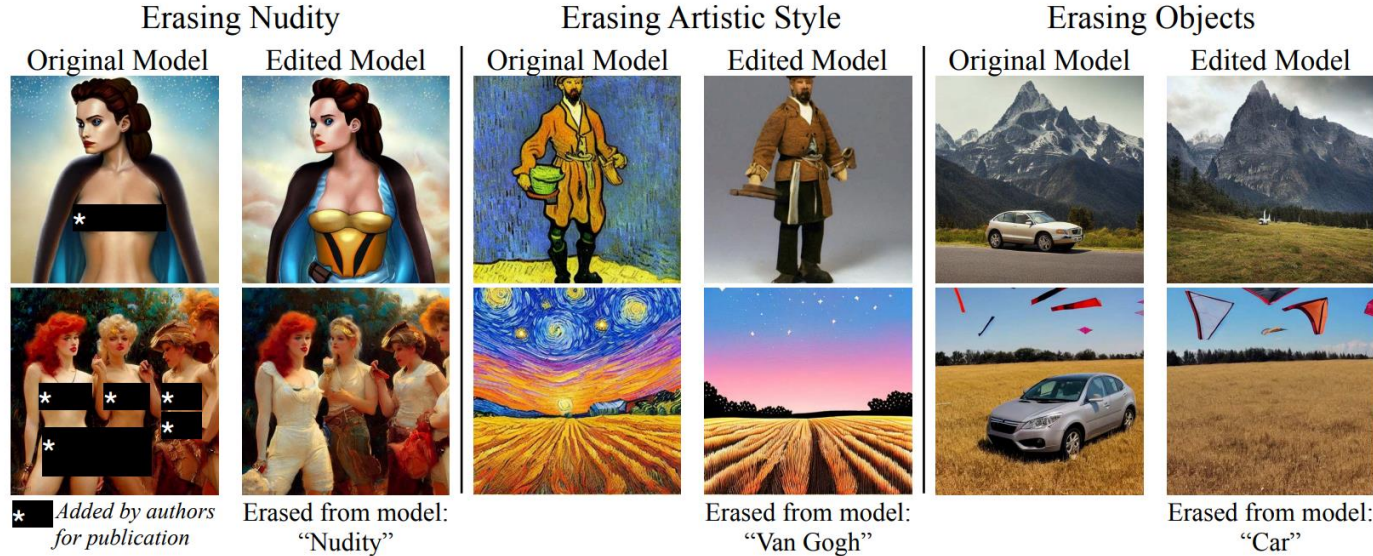


# Takeways

- The evolution of the deepfake technologies (VAE, GAN, Diffusion models, etc) is fast and requires more ethical consideration for it.
- Educate the public to less rely on the videos as the evidence.
- The release of powerful AI models should be careful.



# Safe Deployment of Diffusion Model



Gandikota, Rohit, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. "Erasing concepts from diffusion models." *arXiv preprint arXiv:2303.07345* (2023).

# Thank You!

# Any Questions?

Jun-Cheng Chen

Research Center for Information Technology Innovation, Academia Sinica

[pullpull@citi.sinica.edu.tw](mailto:pullpull@citi.sinica.edu.tw)



# 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] Rössler, 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>

