Light2Lie: Detecting Deepfake Images Using Physical Reflectance Laws

Kavita Kumari, Sasha Behrouzi, Alessandro Pegoraro, Ahmad-Reza Sadeghi,

NDSS 2026

TECHNISCHE UNIVERSITÄT DARMSTADT

Deepfake Images: From Synthetic Media to Personalized Attack Surface

A teen died after being blackmailed with A.I.-generated nudes. His family is fighting for change

AI Images are Causing Havoc for People Affected by Hurricane Helene

BUSINESS

EU privacy investigation targets Musk's Grok chatbot over sexualized deepfake images

AI 'slop' is transforming social media - and a backlash is brewing

Viral scam: French woman duped by AI Brad Pitt love scheme faces cyberbullying

UNICEF calls for criminalization of AI content depicting child sex abuse

By Jasper Ward

February 4, 2026 6:20 PM GMT+1 · Updated February 4, 2026

Deepfake Definition

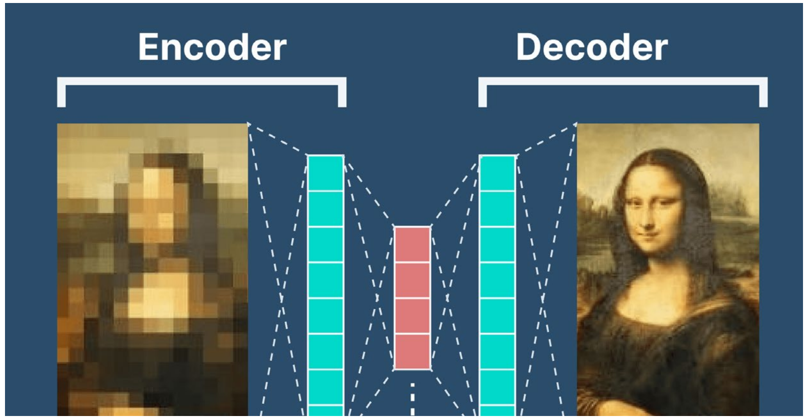
- A deepfake is any synthetic media, generated using AI, that is designed to imitate or fabricate human behavior or identity with the intent to mislead or deceive.

- The evolution of deepfake generation:

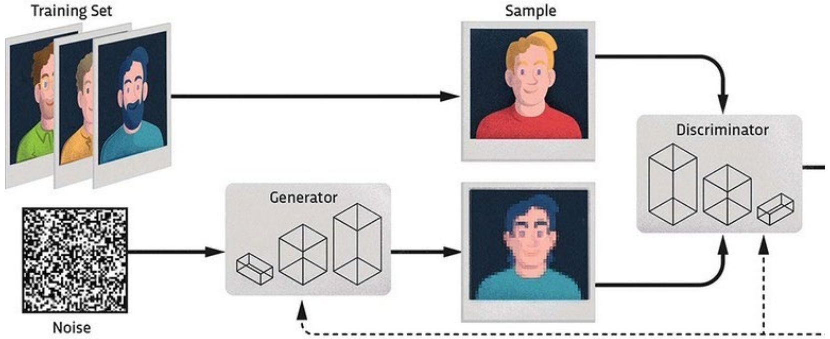
→ Autoencoders

→ GANs

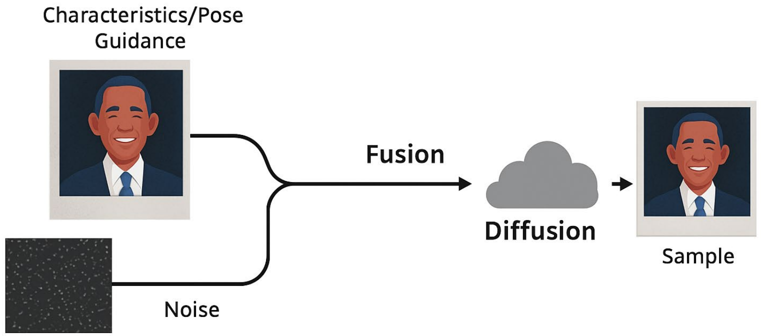
→ Diffusion + Fusion



Autoencoder



GAN (Generative Adversarial Network)

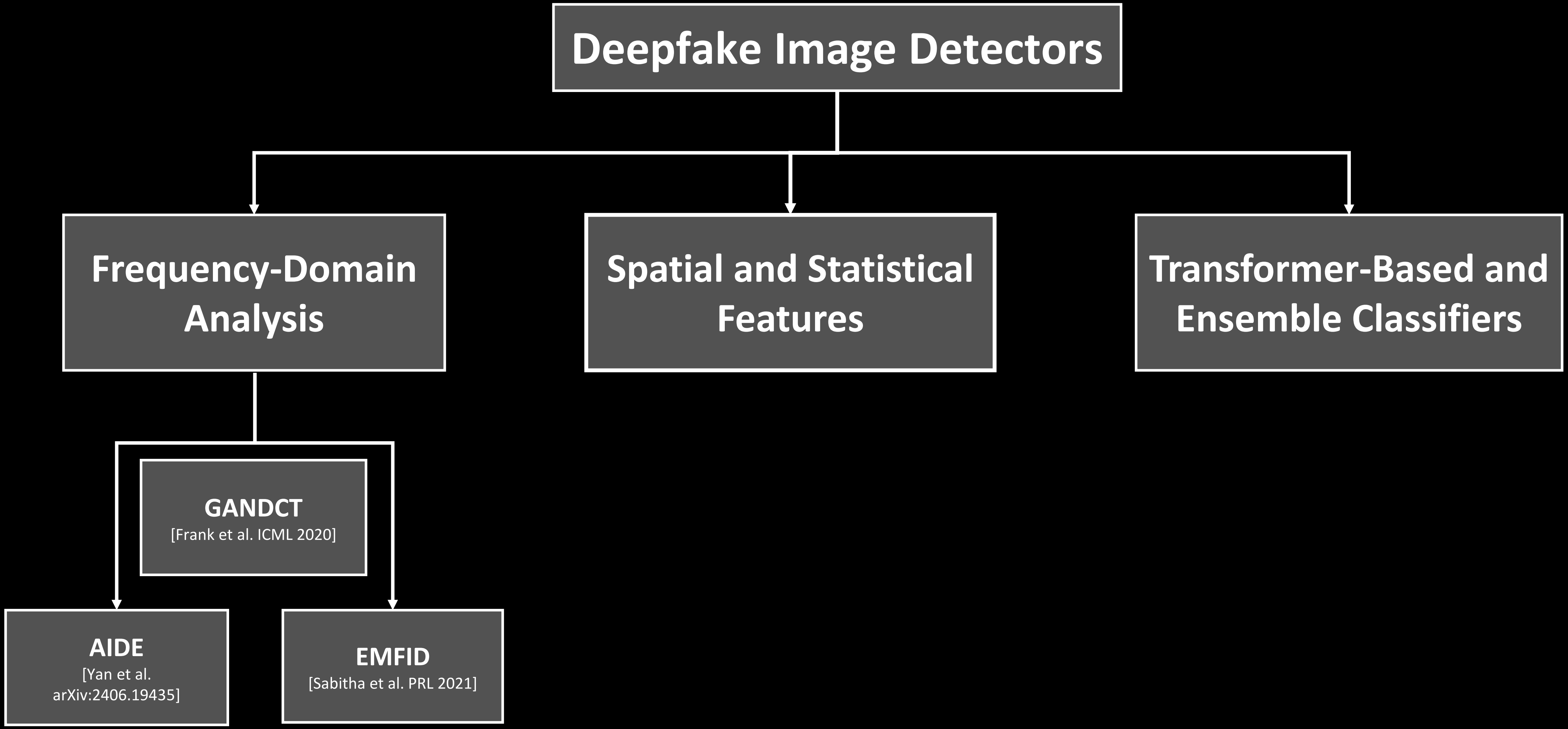


Fusion + Diffusion

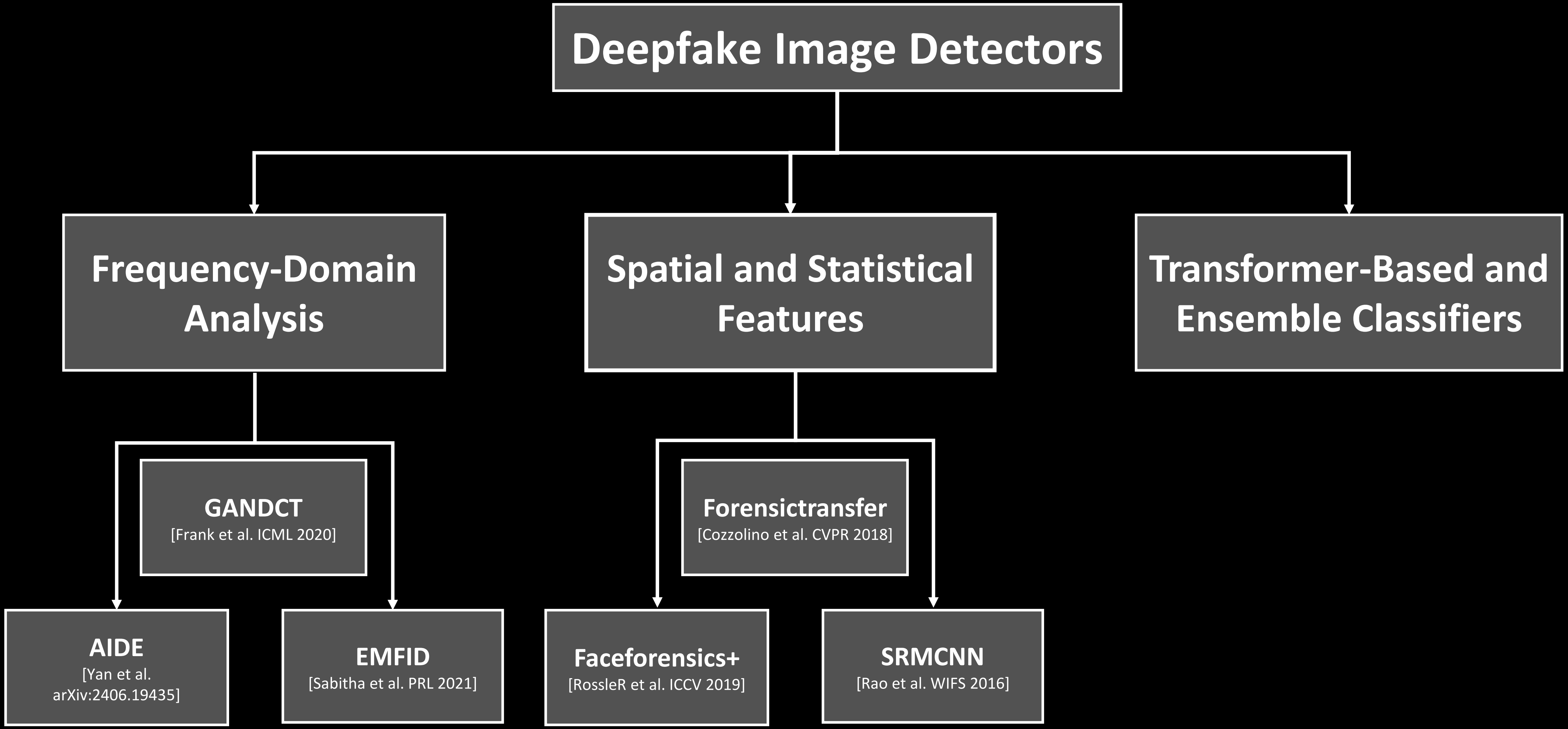
Categorization of Image Detectors

- **Deepfake Image Detectors**
  - **Frequency-Domain Analysis**
  - **Spatial and Statistical Features**
  - **Transformer-Based and Ensemble Classifiers**

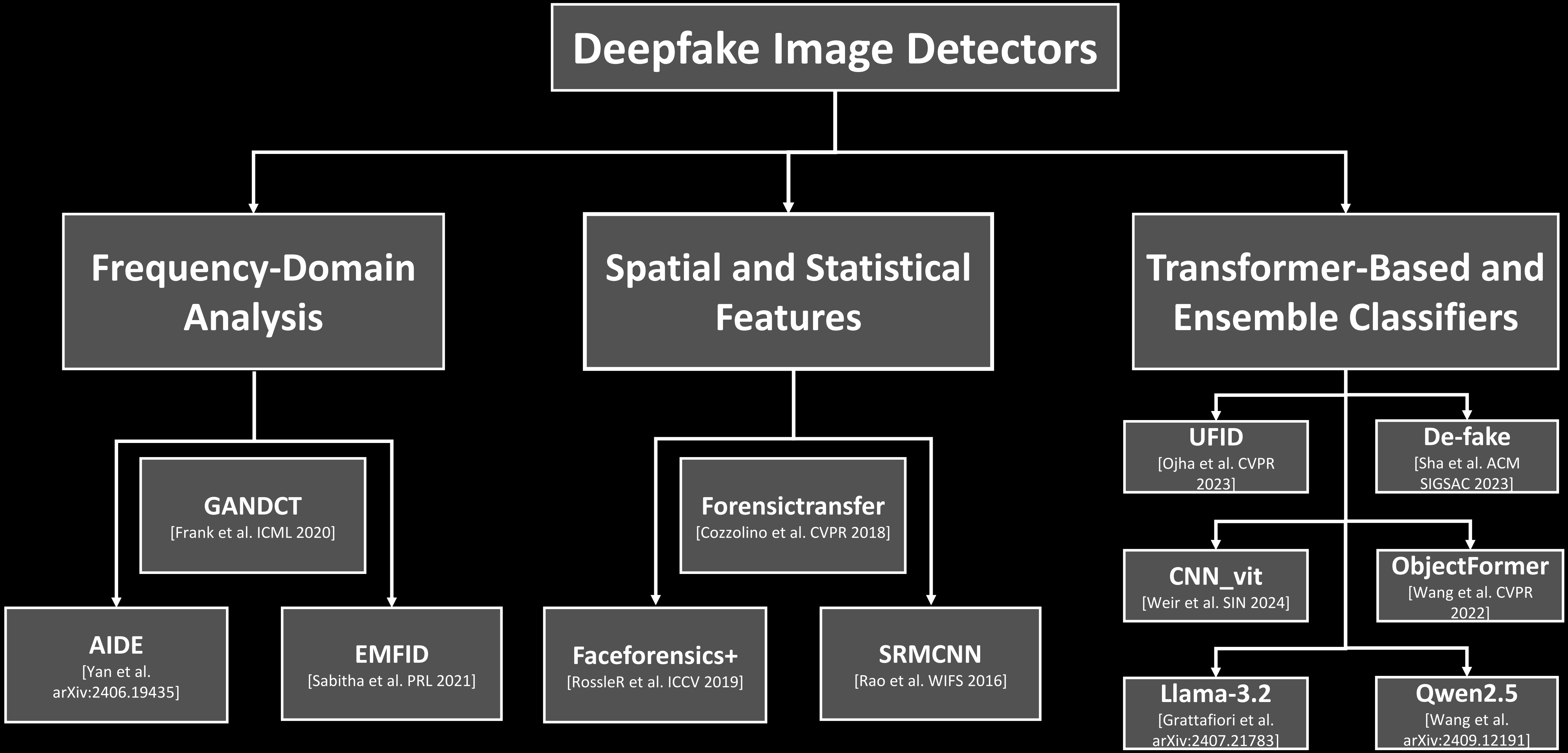
Categorization of Image Detectors



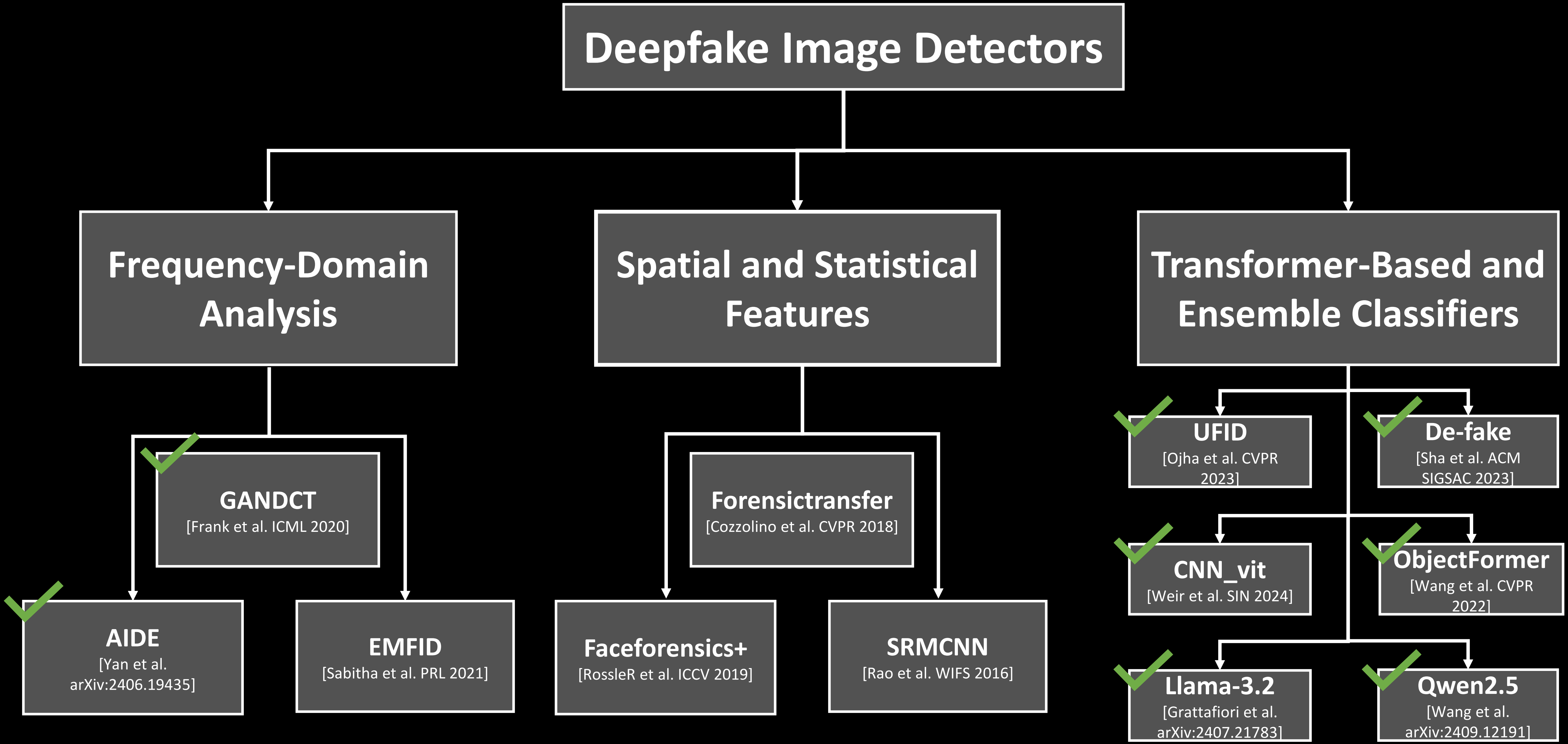
Categorization of Image Detectors



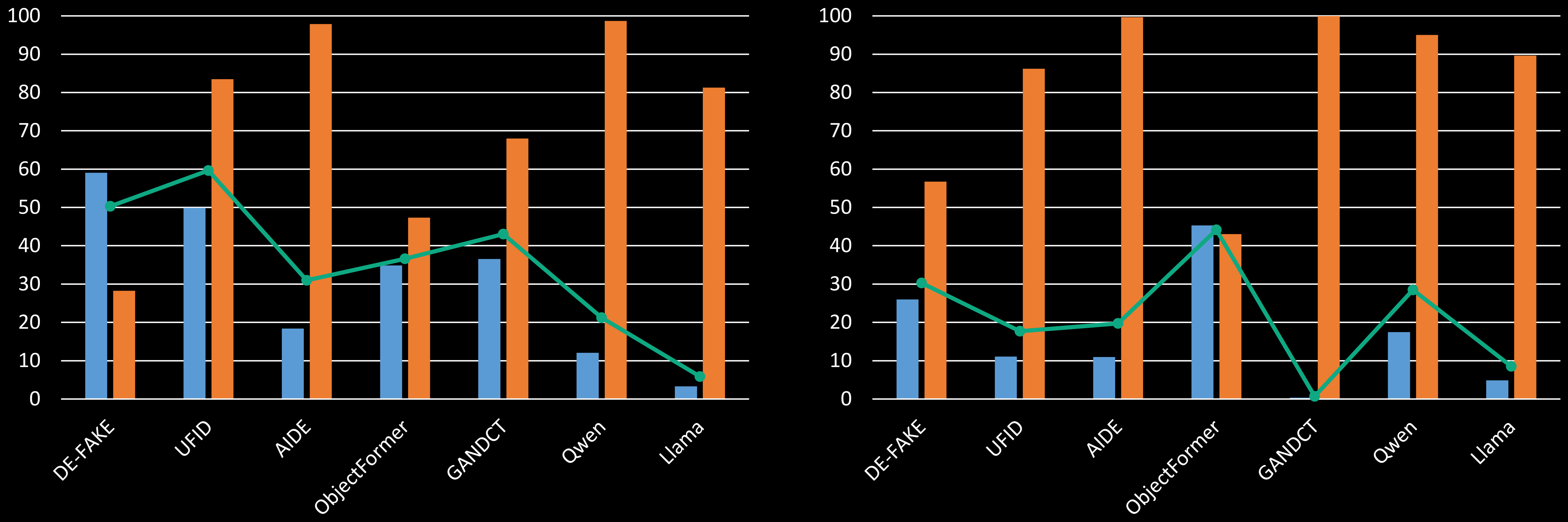
Categorization of Image Detectors



Categorization of Image Detectors



Evaluation of Existing Detectors



$TPR = \frac{TP}{TP + FN}$ $TNR = \frac{TN}{TN + FP}$ $F1 = \frac{2\,TP}{2TP + FP + FN}$

Limitations of Deepfake Detectors

Current AI detection methods:

- Accuracy is domain dependent

  [Rossler et al., ICCV 2019] [Luo et al., CVPR 2021]

- Limited adaptability and generalization

  [Hao et al., ICCV 2021] [Liang et al., CVPR 2022]

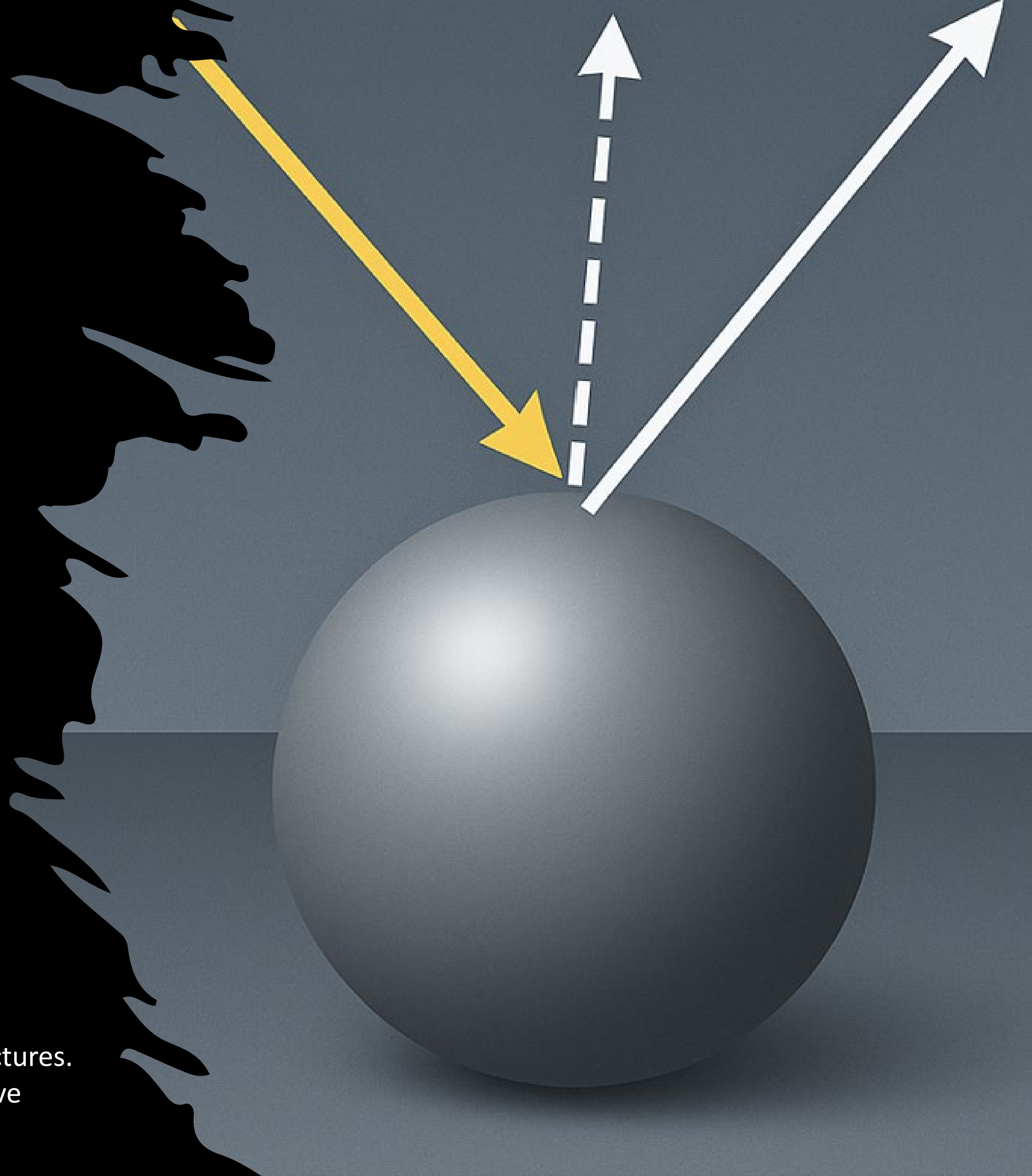
- Lacking comprehensive dataset evaluation

  [Zi et al., ACM MM 2020] [Li et al., CVPR 2020]

Our Approach: Light2Lie

Intuition and Hypothesis

- AI images often mimic texture but fail at realistic light reflections.
- Real reflections follow physical laws based on surface properties.
- We utilize Blinn's microfacet theory[1] to model surface reflections.
- Analyzing light behavior reveals if an image breaks physical rules.
- Inconsistent highlights expose synthetic origins



[1] Blinn, J. F. (1977, July). Models of light reflection for computer synthesized pictures. In Proceedings of the 4th annual conference on Computer graphics and interactive techniques (pp. 192-198).

Light2Lie: Motivation

Modeling each pixel as a microfacet (tiny planar surfaces)

Light2Lie: Motivation

Modeling each pixel as a microfacet (tiny planar surfaces)

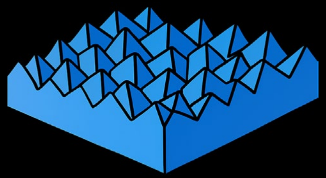
↓

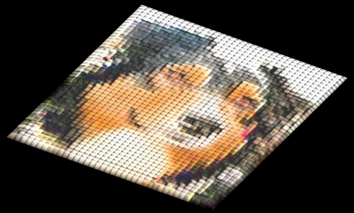
To determine the intensity & placement of highlights on a surface

Light2Lie: Motivation

Modeling each pixel as a microfacet (tiny planar surfaces)

To determine the intensity & placement of highlights on a surface

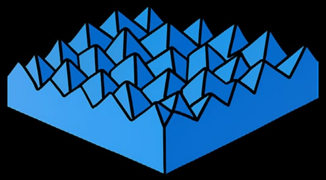


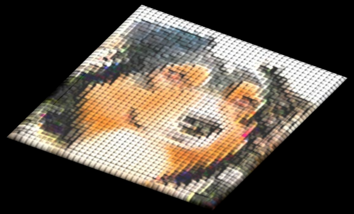


Light2Lie: Motivation

Modeling each pixel as a microfacet (tiny planar surfaces)

To determine the intensity & placement of highlights on a surface





Approximates the physical models

Light2Lie: Motivation

Modeling each pixel as a microfacet (tiny planar surfaces)

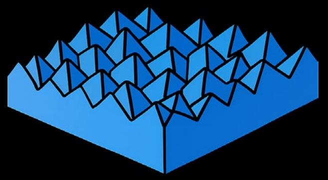
↓

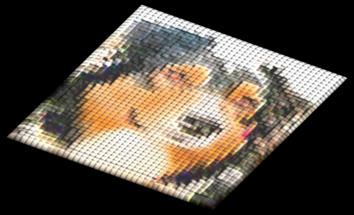
To determine the intensity & placement of highlights on a surface

Approximates the physical models

↓

To extract the surface geometry & determine the reflectivity of the input image samples

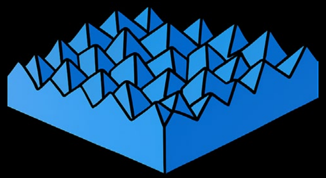


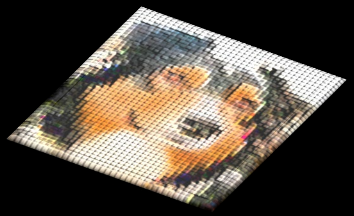


Light2Lie: Motivation

Modeling each pixel as a microfacet (tiny planar surfaces)

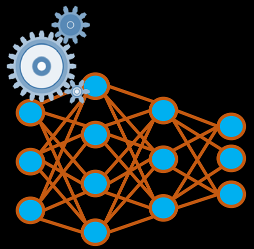
To determine the intensity & placement of highlights on a surface

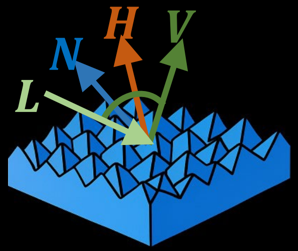




Approximates the physical models

To extract the surface geometry & determine the reflectivity of the input image samples





Light2Lie: Motivation

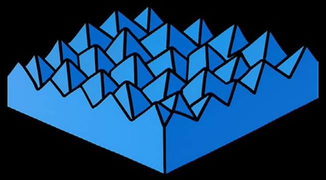
Modeling each pixel as a microfacet (tiny planar surfaces)

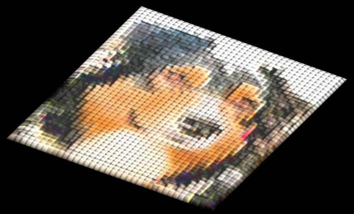
To determine the intensity & placement of highlights on a surface

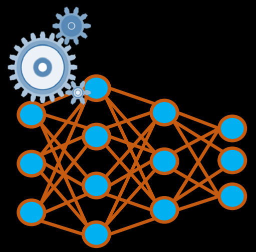
Approximates the physical models

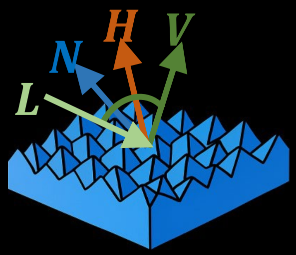
To extract the surface geometry & determine the reflectivity of the input image samples

Incorporates compositional bias





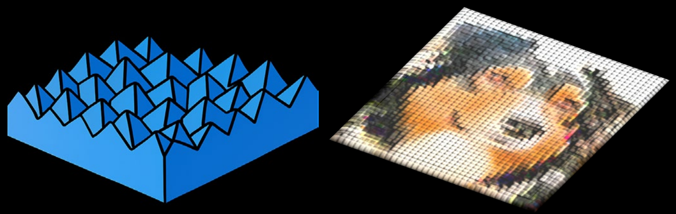




Light2Lie: Motivation

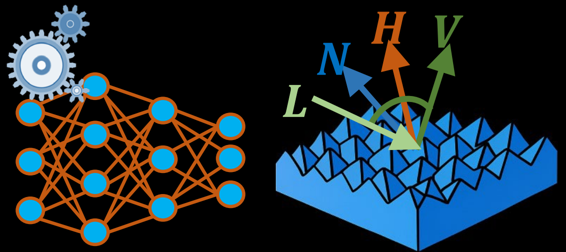
Modeling each pixel as a microfacet (tiny planar surfaces)

To determine the intensity & placement of highlights on a surface



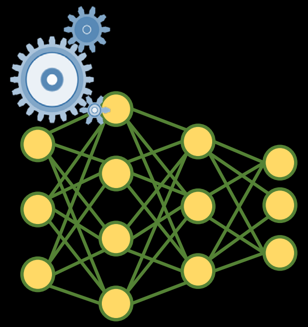
Approximates the physical models

To extract the surface geometry & determine the reflectivity of the input image samples



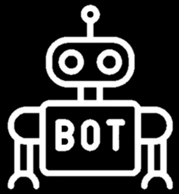
Incorporates compositional bias

To penalize mismatch of energies by the model



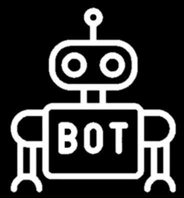
Light2Lie: Modeling the surfaces

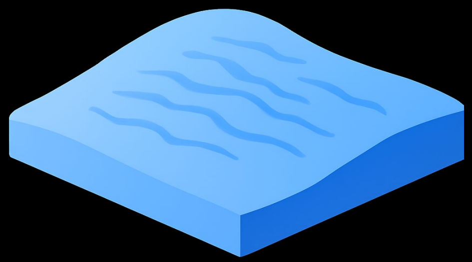
Input Images



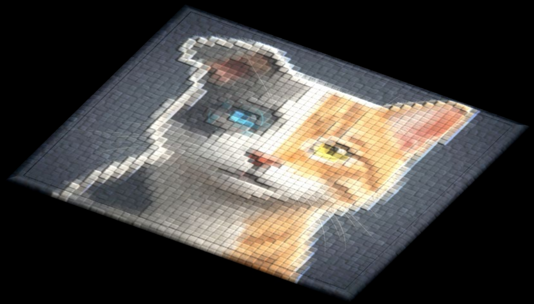
Light2Lie: Modeling the surfaces

Input Images





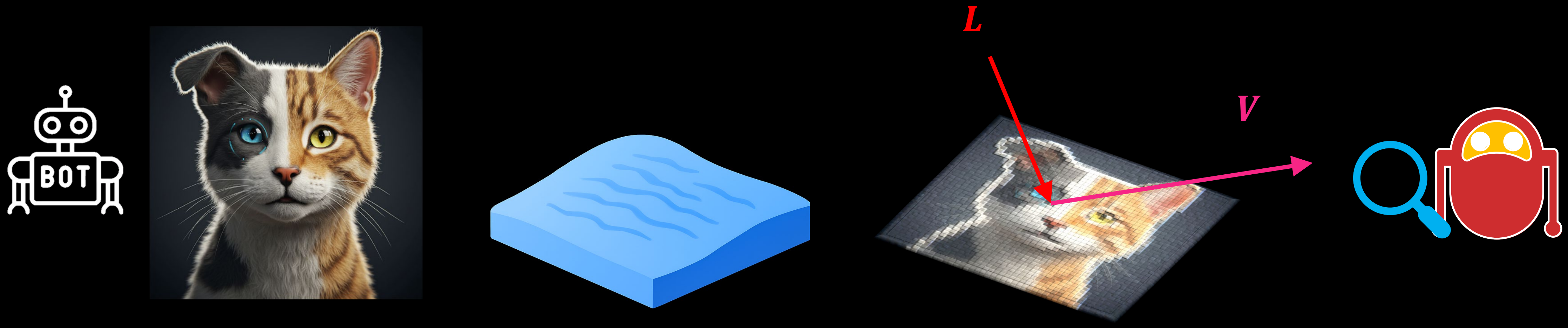
Microfacets



Light2Lie: Modeling the surfaces

Input Images

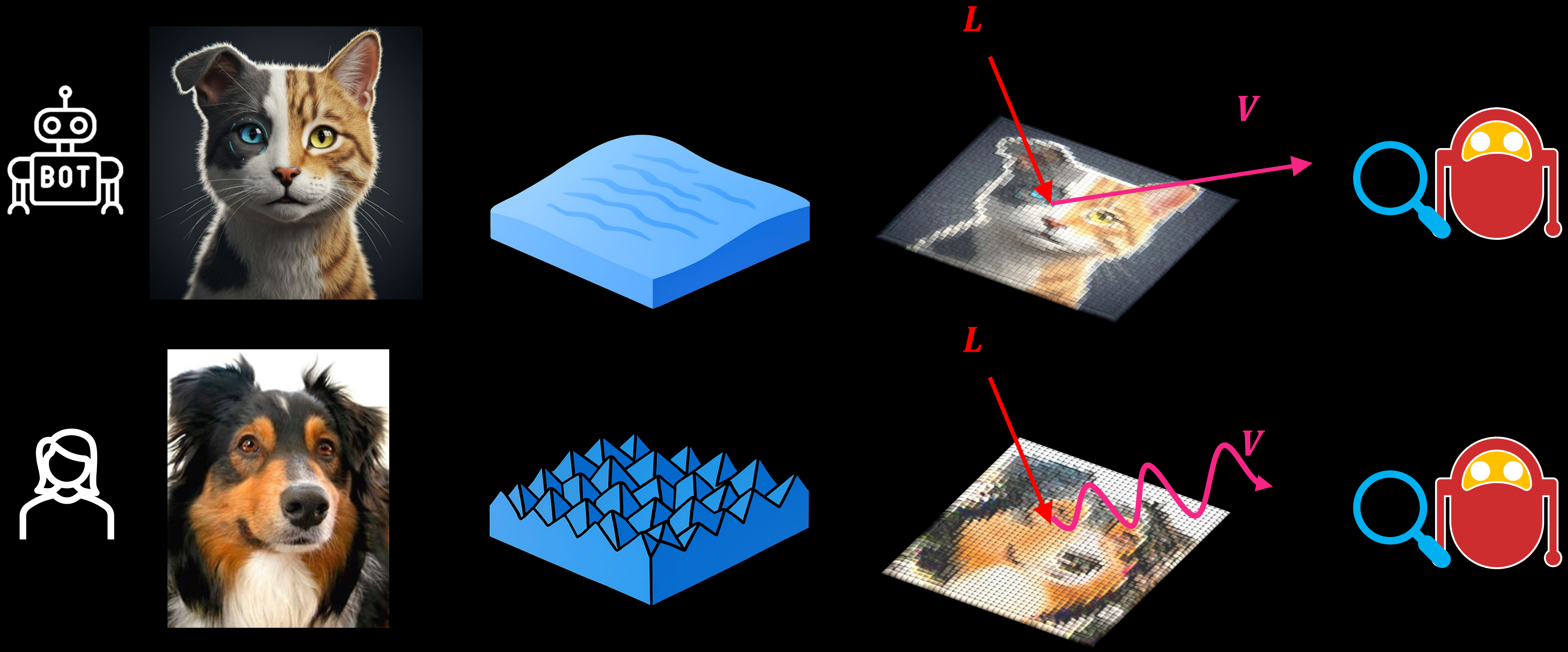
Microfacets



Light2Lie: Modeling the surfaces

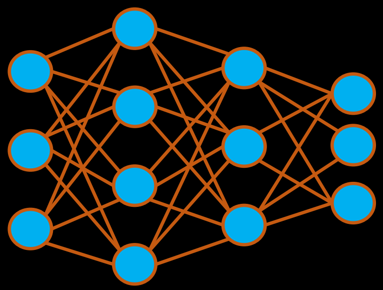
Input Images

Microfacets



Core Components

Base Reflectance



Embedding

Vector $\mathcal{E}(\vec{x})$

$$\boldsymbol{F_0} = B_s(\mathcal{E}(\vec{x}))$$

$$L_s = Binay\_cross\_entropy$$

$$W : argmin_W L_s(W)$$

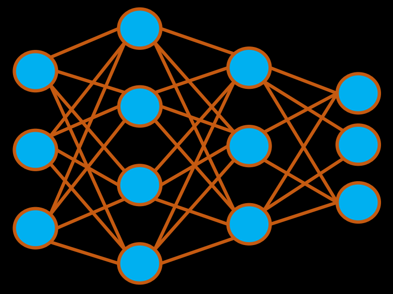
Core Components

Embedding $\mathcal{E}(\vec{x}) \coloneqq [e_1, e_2, \ldots, e_k]$

Embedding

Vector $\mathcal{E}(\vec{x})$

Base Reflectance



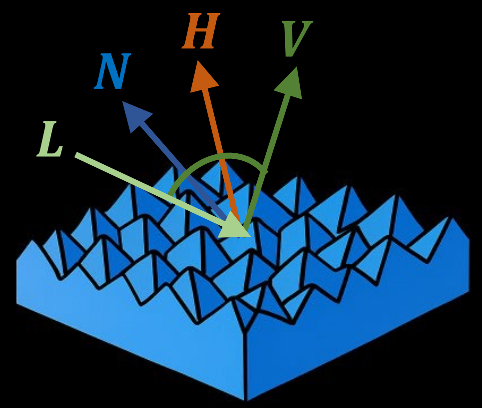
$$\boldsymbol{F_0} = B_s(\mathcal{E}(\vec{x}))$$

$$L_s = Binay\_cross\_entropy$$

$$W : argmin_W L_s(W)$$

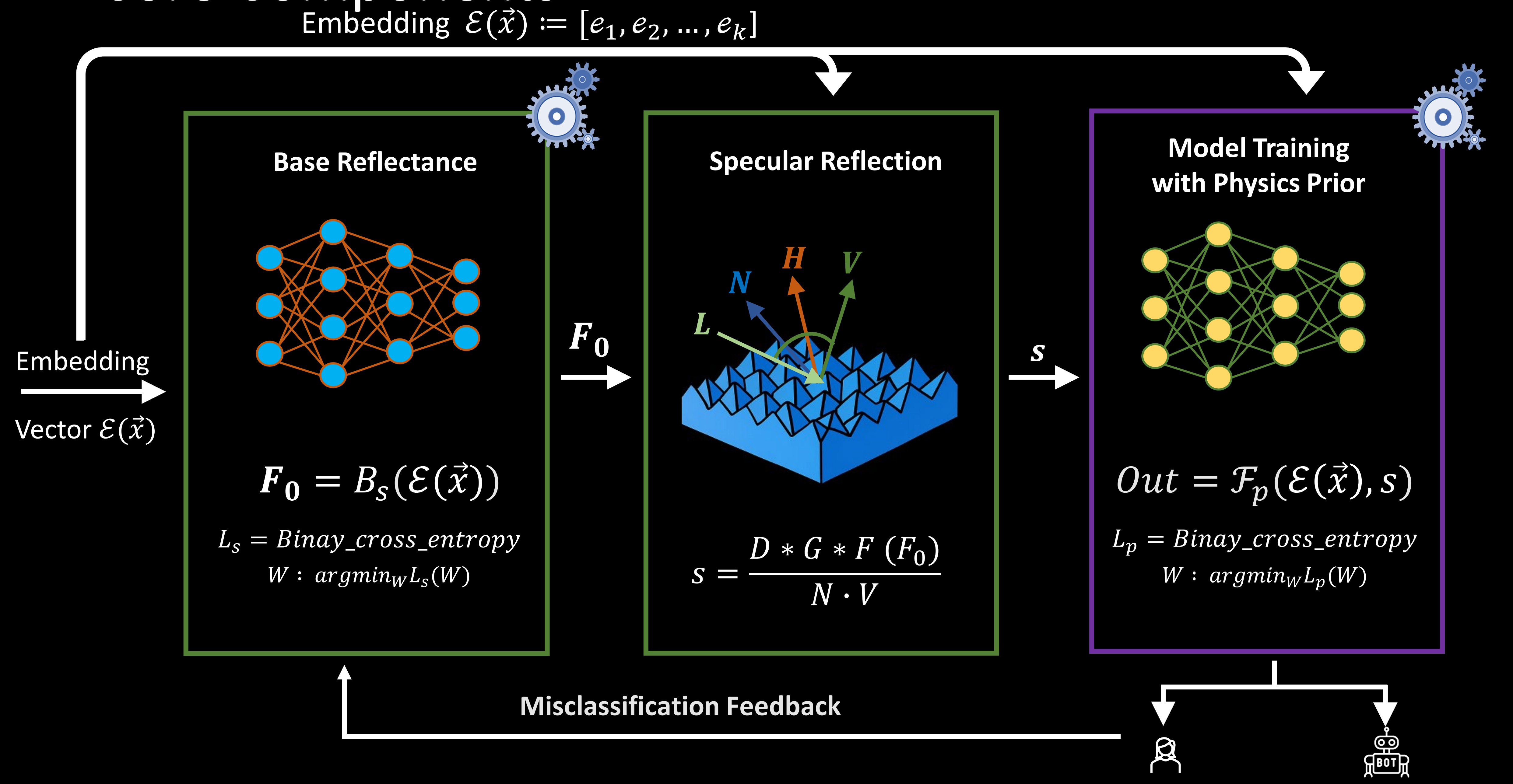
$\boldsymbol{F_0}$

Specular Reflection



$$s = \frac{D * G * F\ (F_0)}{N \cdot V}$$

Core Components



Evaluation

Evaluation

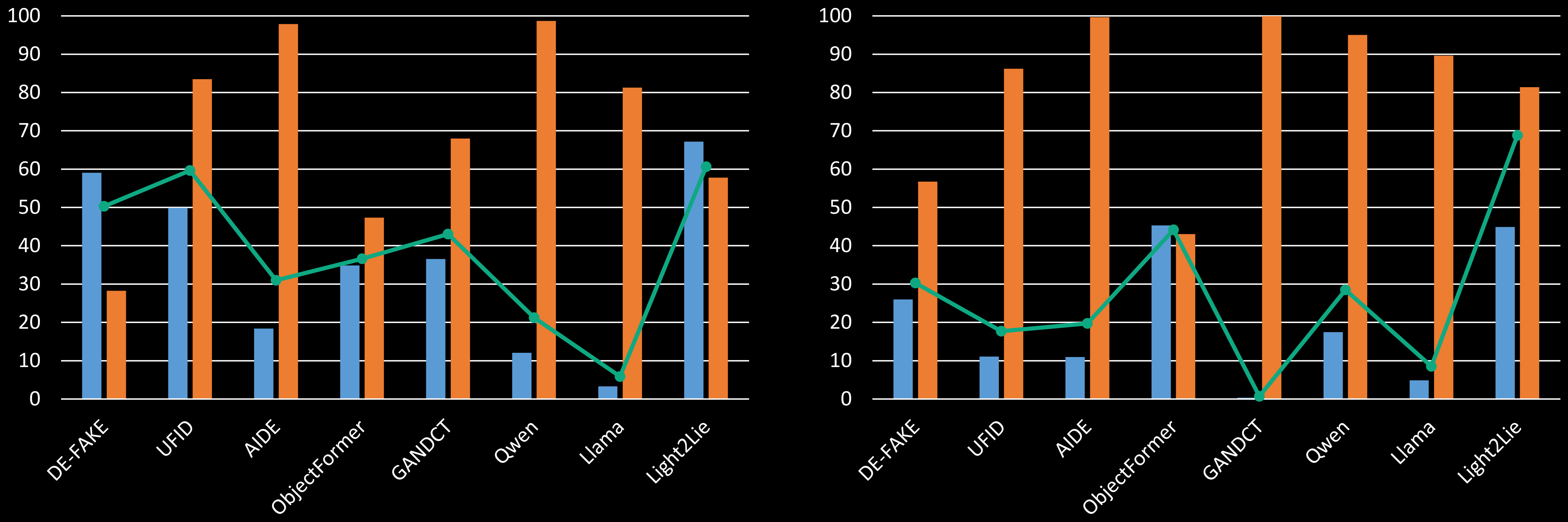
Dataset statistics for GAN-based, Diffusion-based approaches, and Genuine images

	Generation Approach	# of Samples
Diffusion	• DALL·E 2 • Stable Diffusion • DreamStudio	• 27,072 • 50,048 • 32,768
GAN	• StyleGAN • CIFAKE	• 7,040 • 60,096
Real	• LAION	• 6,358

Existing Works tested for Generalized Evaluation

- DE-FAKE
- Universal Fake Image Detector (UFID)
- AIDE
- Objectformer
- GANDCT
- Llama-3.2-11B-Vision-Instruct
- Qwen2.5-VL-72B-Instruct

Evaluation of Existing Detectors



$TPR = \frac{TP}{TP + FN}$ $TNR = \frac{TN}{TN + FP}$ $F1 = \frac{2\,TP}{2TP + FP + FN}$

Conclusion

- Deepfake images are a real threat to modern societies

- We employed Physics Augmented Intelligence
  - improves modelling
  - Allows for generalization

- We addressed existing detectors' limitations

- We proposed Light2Lie
  - Utilized Reflectance Laws to detect deepfakes
  - Better generalization performance to new approaches