Pretrained Image Encoders of Vision-Language and Multimodal Models for Synthetic Facial Image Detection

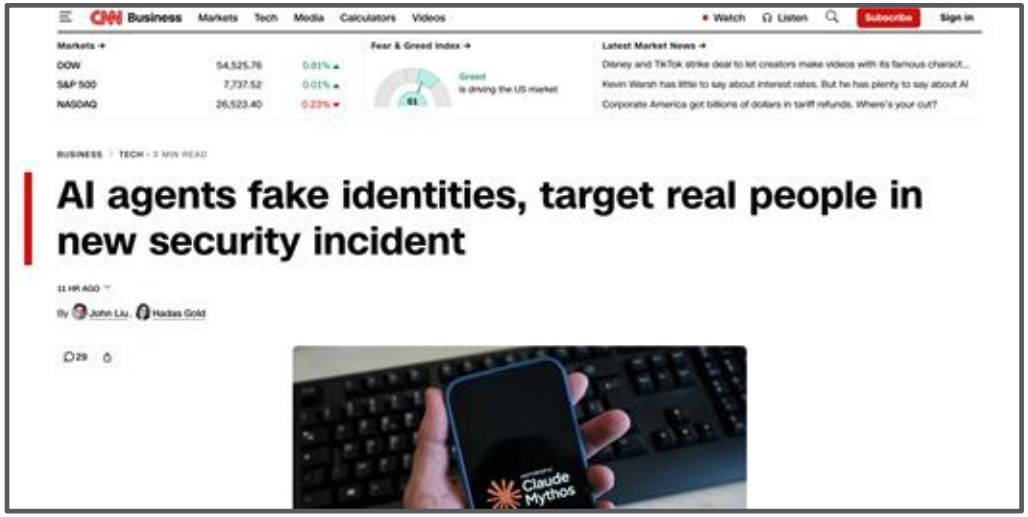
Liyue Fan and Joseph Roberson

Privacy Lab

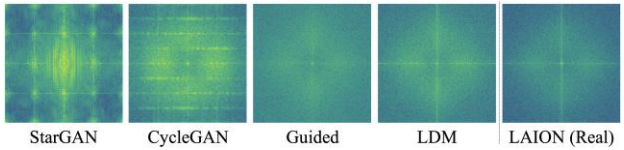
University of North Carolina at Charlotte

Motivation

- Detecting synthetically generated facial images have important implications
  - Regulate use of synthetic data and prevent potential misuse
  - Advance synthetic data generation methods, from encoder/decoder and GAN, to diffusion models



Challenges and SOTA



- **Generalizability**: supervised classifiers trained to detect certain generation methods may fail to detect unseen methods
  - Latching on artifacts of generation methods for training
  - Pre-trained CLIP ViT may enable generalization from GANs to autoregressive and diffusion models [Ojha et al. 2023]

Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018)
Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS). pp. 1–6. IEEE (2019)
Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)

- **Low-data regime**: large amounts of training samples may be hard to obtain in practice, e.g., new generation methods
- **Robustness**: attackers may attempt to evade detection by post-processing synthetic samples

Vision-Language and Multimodal Models

- **Potential**: VLM and MM delivered promising results in complex tasks and a variety of domains, e.g., medical imaging

Bommasani, R.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
Han, T., Adams, L.C., Nebelung, S., Kather, J.N., Bressem, K.K., Truhn, D.: Multimodal large language models are generalist medical image interpreters. medRxiv pp. 2023–12 (2023)

- **Hypothesis**: they may provide a powerful feature space for accurate and generalizable synthetic image detection

- **Intuition**:

  - VLM and MM trained on large amounts of public domain data, not to recognize synthetic images

  - May capture semantic information in real images by leveraging multiple modalities

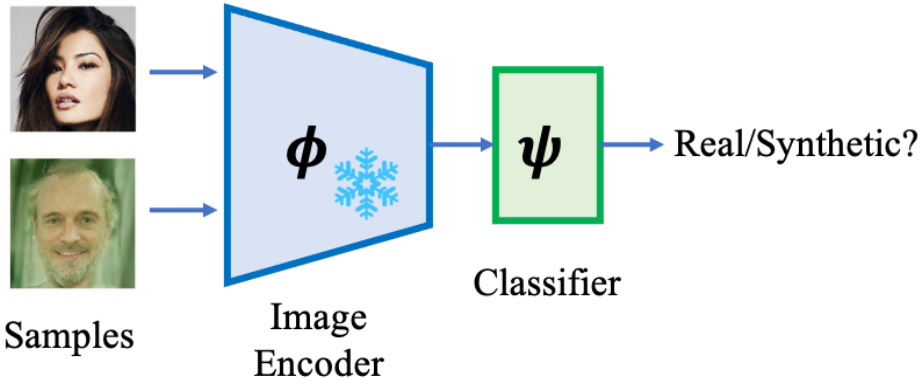
Contributions

- Developed synthetic facial image detection on pretrained image encoders of VLM and MM models
  - prior work: leveraged VLM and MM for natural language explanations
- Adopted a range of architectures, including CLIP ViT, Flamingo, and ImageBind
  - prior work: only considered CLIP ViT
- Aimed at distinguishing subtle differences between various diffusion model-based generation methods
  - prior work: differentiated between different types of models, such as GANs vs. LDM

Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., et al.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models. arXiv preprint arXiv:2410.09732 (2024)
Yu,P.,Fei,J.,Gao,H.,Feng,X.,Xia,Z.,Chang,C.H.:Unlockingthecapabilitiesof large vision-language models for generalizable and explainable deepfake detection. arXiv preprint arXiv:2503.14853 (2025)
Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)

Approach Overview

- Leverage pretrained image encoders (ϕ) in VLM and MM models and train only a classifier (ψ) for image embeddings



- **Classifier**: a single linear layer with sigmoid activation
- **Image Encoder**: CLIP ViT (B/32, B/16, L/14), Idefics-9B/Flamingo, ImageBind

Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. Advances in Neural Information Processing Systems 36, 71683–71702 (2023)

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180– 15190 (2023)

Experiment Settings

- **Datasets**: DiffusionFace, 12 subsets, including real, 5 unconditional diffusion methods, 6 conditional diffusion methods
  - 30000 samples per subset, 80-20 train/test split randomly

Chen, Z., Sun, K., Zhou, Z., Lin, X., Sun, X., Cao, L., Ji, R.: Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. arXiv preprint arXiv:2403.18471 (2024)

- **Baselines**:
  - CNNDetection, ResNet-50 fully fine-tuned for each task
  - CLIP ViT L/14, frozen image encoder

Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF confer- ence on computer vision and pattern recognition. pp. 8695–8704 (2020)
Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24480–24489 (2023)

- **Implementation**: Adam optimizer, batch size 64, learning rate 0.001, early stopping patience 5
  - https://github.com/fan-group/multimodal-syndet

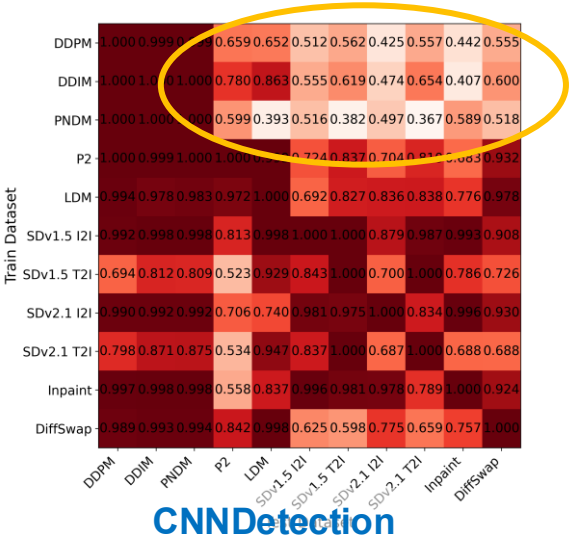
Results – Method Attribution

- Both supervised baseline (CNNDetection) and pretrained VLM and MM image encoders perform quite well
  - CNNDetection and ImageBind image encoder are top performers; CLIP ViT L/14 and Flamingo perform comparably
  - Pretrained image encoders are more efficient in training
  - CLIP ViT B/32, B/16 underperform the larger L/14 model

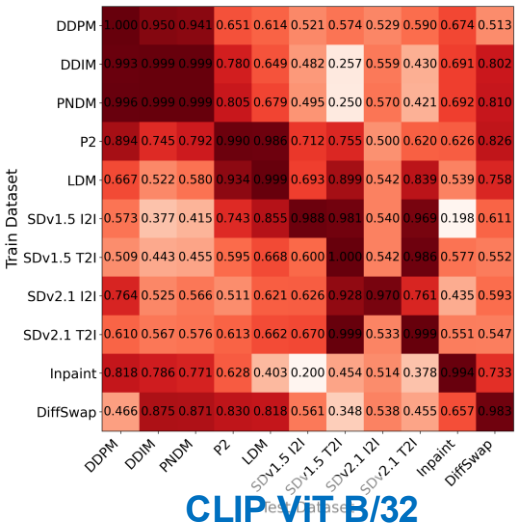
Table 1: Generation Method Attribution Evaluation: best performance in each row in boldface; second best performance in each row underlined.

Metric	CNNDetection	CLIP ViT			Flamingo	ImageBind
		B/32	B/16	L/14		
Top-1 Acc	**0.978** ± 0.000	0.893 ± 0.002	0.917 ± 0.001	0.918 ± 0.000	0.954 ± 0.001	<u>0.957</u> ± 0.000
Top-2 Acc	<u>0.996</u> ± 0.000	0.990 ± 0.001	0.995 ± 0.000	<u>0.996</u> ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000
Top-3 Acc	<u>0.999</u> ± 0.000	0.998 ± 0.000	<u>0.999</u> ± 0.000	<u>0.999</u> ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000
AUC	**1.000** ± 0.000	0.993 ± 0.000	0.995 ± 0.000	0.995 ± 0.000	0.997 ± 0.000	<u>0.998</u> ± 0.000

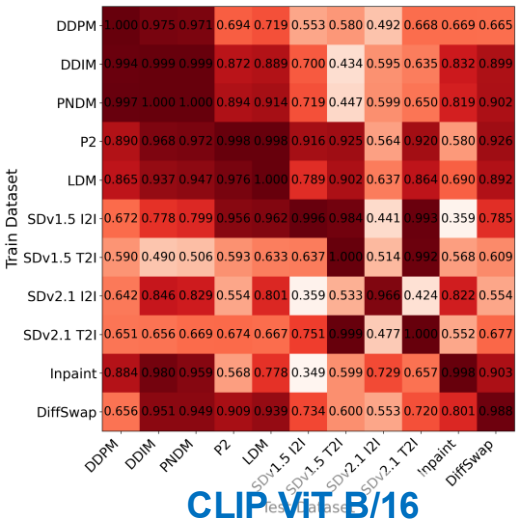
Results – Unseen Method Detection



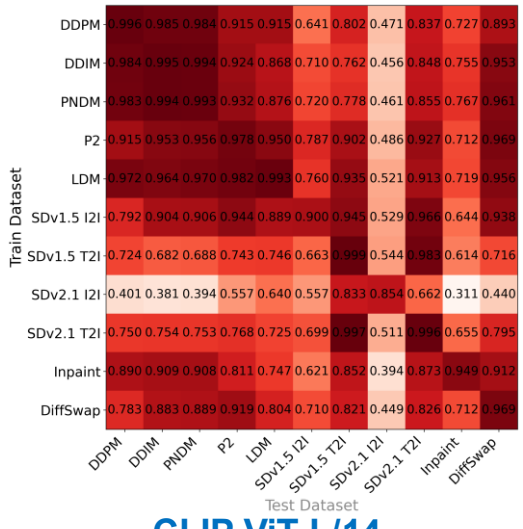
CNNDetection



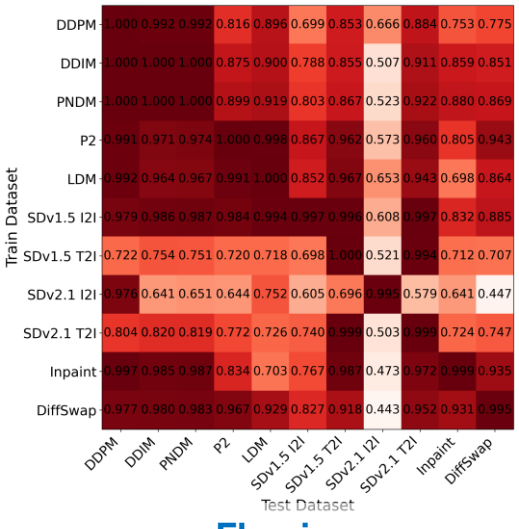
CLIP ViT B/32



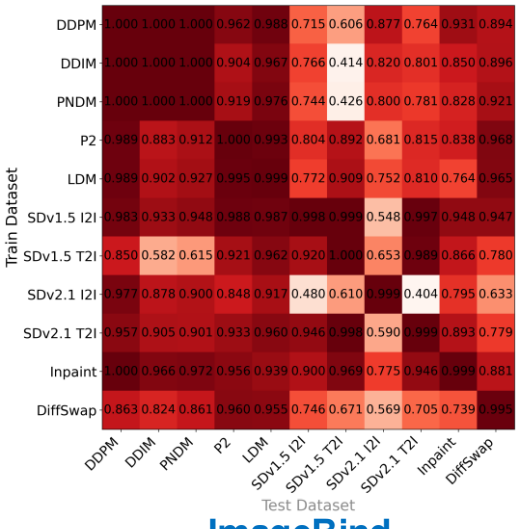
CLIP ViT B/16



CLIP ViT L/14



Flamingo



ImageBind

Results – Robustness

- Post-processing test samples lead to reduced performance
  - CNNDetection and pretrained Flamingo encoder are top performers; note that CNNDetection image encoder is fully updated with data augmentation

Table 2: Generation Method Attribution Robustness : best performance in each row in boldface; second best performance in each row underlined.

Setting/Metric		CNNDetection	CLIP ViT L/14	Flamingo	ImageBind
Resize (/16)	Top-1 Acc	**0.267** ± 0.012	0.200 ± 0.002	0.200 ± 0.001	0.098 ± 0.000
	Top-2 Acc	**0.455** ± 0.023	0.356 ± 0.002	0.378 ± 0.001	0.209 ± 0.002
	Top-3 Acc	**0.596** ± 0.026	0.473 ± 0.002	0.495 ± 0.001	0.327 ± 0.004
	AUC	**0.764** ± 0.016	0.695 ± 0.002	0.711 ± 0.001	0.648 ± 0.001
JPEG ($q = 10$)	Top-1 Acc	**0.267** ± 0.012	0.207 ± 0.002	0.235 ± 0.001	0.108 ± 0.000
	Top-2 Acc	**0.455** ± 0.023	0.343 ± 0.003	0.378 ± 0.001	0.233 ± 0.003
	Top-3 Acc	**0.595** ± 0.031	0.464 ± 0.002	0.495 ± 0.001	0.343 ± 0.002
	AUC	**0.763** ± 0.016	0.700 ± 0.003	0.711 ± 0.001	0.647 ± 0.001
Gaussian Blur ($r = 22$)	Top-1 Acc	0.225 ± 0.030	0.209 ± 0.000	**0.272** ± 0.002	0.098 ± 0.001
	Top-2 Acc	**0.402** ± 0.041	0.338 ± 0.000	**0.402** ± 0.002	0.240 ± 0.003
	Top-3 Acc	**0.545** ± 0.042	0.435 ± 0.001	0.514 ± 0.001	0.365 ± 0.000
	AUC	**0.729** ± 0.025	0.700 ± 0.002	0.725 ± 0.002	0.653 ± 0.001

Results – Sample Efficiency

- VLM and MM encoders enable accurate classification with limited training samples
  - 5 samples per class will yield >0.94 test AUC with the pre-trained ImageBind encoder

Table 3: Generation Method Attribution (Test AUC) with Limited Training Samples: best performance in each row in boldface; second best performance in each row underlined.

#samples per class	CNNDetection	CLIP ViT L/14	Flamingo	ImageBind
5	0.786 ± 0.015	$\underline{0.808} \pm 0.008$	0.803 ± 0.027	$\mathbf{0.944} \pm 0.012$
10	0.855 ± 0.005	$\underline{0.861} \pm 0.009$	0.854 ± 0.010	$\mathbf{0.961} \pm 0.004$
15	$\underline{0.883} \pm 0.011$	0.880 ± 0.004	0.879 ± 0.015	$\mathbf{0.970} \pm 0.003$
30	$\underline{0.924} \pm 0.005$	0.908 ± 0.001	0.911 ± 0.003	$\mathbf{0.978} \pm 0.004$

Discussion

- Pretrained image encoders in vision-language and multimodal models enable robust and generalizable synthetic facial image detection, reducing fine-tuning needs.

- The ImageBind encoder is highly efficient in the number of training samples, providing accurate classification with as low as 5 samples per generation method.

- Within the same model family, CLIP ViT L/14 outperforms B/32 and B/16 models in method attribution and unseen detection.

Future work

- Conduct data augmentation for training the classification head with pretrained image encoders to achieve improved robustness
- Increase the number of runs for the study of sample efficiency, due to randomly selected training samples
- Evaluate the inclusion of multiple real data sources as well as additional generation methods (such as DALL-E and Midjourney)
- Further study the representation space VLM and MM foundation models, e.g., multi-modality privacy and memorization risks

Acknowledgement

- Co-author: Joseph Roberson
- External collaborator: Dr. Dan Li at Vanderbilt University
- Anonymous reviewers for helpful feedback and comments

- Funding: National Science Foundation grant CNS-2027114. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors
- Infrastructure: College of Computing and Informatics for providing GPU server rack and server management

- Contact: liyue.fan@charlotte.edu