SYNTHETIC MEDIA ATTRIBUTION

1 **Divya Appapogu, Daniel Delijani**
{divsp}, {delijani}@bu.edu

1 Introduction

In recent years, we have witnessed an explosive growth in the generation of synthetic media, driven by the increasing accessibility of sophisticated generative models. Remarkable models like DALL · E3 [1], DALL · E2[2], Stable Diffusion[3], ImageGen[4], and a host of others [5] [6] [7] have revolutionized various scientific fields and daily life applications. However, the democratization of these powerful synthetic media tools has also given rise to a significant threat. This threat is particularly pronounced in contexts like the propagation of fake news, the proliferation of deceptive posts on social media, and within digital environments with limited content oversight. These synthetic images and videos, produced with remarkable realism, can be deployed across diverse domains, contributing to mass misinformation, societal panic, and eroding the line between reality and fiction.

To address the malicious use of such media, it is crucial to develop techniques for detecting and attributing falsified images or text to the specific generator or tool used in their creation. Such attribution is critical for combating the spread of misinformation, ensuring accountability for those responsible for generating content, recognizing the creative contributions of artists, and tracing the lineage of the underlying generative model and many other potential issues that could arise.

Our research focuses on the task of attributing a given image to the generator that produced it, without access to information about the manipulation techniques applied. This task involves successfully identifying the generator used which involves image-based models StyleGAN2 [5], StyleGAN3 [6], Taming-Transformers [8], Latent Diffusion [9], LSGM [10], and CLIP-Guided Diffusion [11] for image generation. This work is part of the SemEval competition conducted by DARPA.

By enhancing our understanding of the capabilities and limitations of these generative models, we aim to mitigate the risks associated with the misuse of synthetic media, preserving the integrity of digital information in an era where the boundaries between fact and fiction have become increasingly porous.

To tackle this, we've developed an end-to-end multi-head classifier with two distinct heads. This classifier takes an input image and produces an output vector representing the probabilities of the image being generated by different generators. We implemented this model using a ResNet-18 architecture, where the final layer acts as a classification head for generating the vector. Here we use ResNet-18 as the backbone for the classifier, with dual classification layers and one of these layers encourages misclassification of subclasses, aiming to obscure subclass information within the model.

To achieve this, we employ an adversarial strategy at runtime. A subclass classifier, such as ImageNet, is used to generate adversarial versions of test dataset images. These adversarial images are then input into a generator attribution model. The primary idea is to perturb certain image features shared between the generator attribution model and ImageNet using adversarial attacks driven by ImageNet. The goal is to prevent the generator attribution model from relying on subclass information for its predictions. Our approach aims to enhance model privacy and reduce the dependence on subclass-related features, potentially leading to more robust and subclass-agnostic model performance.

2 Previous Work

The existing body of research has made significant strides in addressing the problem of attributing synthetic media to their generators, particularly in the context of deepfakes and GAN-generated images. These studies have introduced various innovative approaches and techniques to tackle the challenge of source attribution, but there remain noteworthy gaps and emerging needs in this field.

For instance Brandon et al [12] addresses the challenge of attributing GAN-generated images to their sources by redefining the problem as a series of binary classification tasks. Transfer learning is leveraged to adapt forgery detection networks to multiple independent attribution problems. The semi-decentralized modular design proposed in this study demonstrates effectiveness in solving attribution tasks efficiently, with class activation mapping aiding model interpretation. While these models perform competitively, they exhibit susceptibility to type II errors in the presence of image perturbations.

To combat the evolving capabilities of deep generative models in producing highly realistic deepfakes, Yang he et al [13] introduce a novel detection approach. Rather than relying on frequency artifacts, this method focuses on re-synthesizing testing images to extract visual cues for detection. The flexibility of the re-synthesis procedure, encompassing super-resolution, denoising, and colorization, contributes to improved effectiveness, cross-GAN generalization, and robustness against perturbations.

Michael et al [14] tackle the problem of attributing synthetic images to specific GAN generators in a white-box setting by inverting the generation process. The approach simultaneously determines whether a generator produced an image and

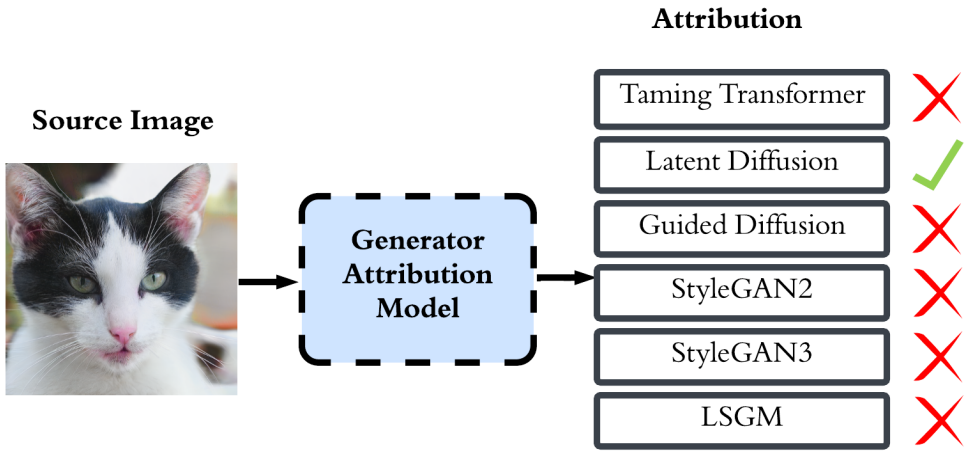


Figure 1: **Synthetic Media Attribution**

recovers an input that closely matches the synthetic image, facilitating source attribution. Ning Yu et al [15] explore learning GAN fingerprints for image attribution and classifying images as real or GAN-generated. They reveal that GANs carry distinct model fingerprints in their generated images, supporting image attribution. Even minor differences in GAN training result in different fingerprints, enabling fine-grained model authentication. The learned fingerprints consistently outperform baselines across various setups.

In this work Sheng Yu et al [16] investigate the possibility of creating a universal detector to distinguish real images from those generated by CNNs, regardless of architecture or dataset used. They collect a dataset of fake images generated by various CNN-based models and demonstrates that a standard image classifier trained on one specific CNN generator can generalize well to unseen architectures and datasets, suggesting common systematic flaws in CNN-generated images. Ning Yu et al [17]introduce artificial fingerprints into generative models during training, which then appear in the generated deepfakes. The approach remains effective across various generative models, exhibits robustness against perturbations, and outperforms state-of-the-art baselines in deepfake detection and attribution, closing the responsibility loop in generative model usage.

This [18] study demonstrates that each GAN leaves specific fingerprints in the images it generates, akin to real-world cameras marking images with traces of their patterns. These fingerprints offer valuable assets for forensic analyses and could aid in source identification. Addressing the challenge of open-world attribution, this [19] research presents an iterative algorithm for discovering images generated by previously unseen GANs. The algorithm exploits the distinct fingerprints left by GANs on their generated images, achieving high accuracy in discovering new GANs and generalizing to unseen real datasets.

In summary, while existing research has laid a solid foundation for synthetic media attribution, there is a need for further advancements in terms of robustness and adaptability to open-world scenarios. The ability to attribute synthetic media to their generators in open-world scenarios, where novel sources like diffusion models may emerge continuously, is a significant requirement and till date there has been very little research addressing the same. Our research endeavors to develop methods that can discover and attribute images generated by previously unseen generators, contributing to a more comprehensive solution.

3 Approach

We construct an end-to-end multi-head classifier C, comprising two distinct heads: C_1 and C_2, that takes an image I as input and produces an output vector O of dimension 1×6. This output vector O represents the probabilities of the image I being generated by various generators, and the probabilities sum to 1. To implement our model, we employ a base ResNet-18 architecture, with the final layer serving as a classification head that generates a 6-dimensional vector.

Mathematically, this can be represented as follows:

$$M : I \rightarrow O, \quad O \in \mathbb{R}^{1\times 6}$$

Where:

- C represents our multi-head classifier model.
- I represents the input image.
- O represents the output vector containing the probabilities of I being generated by specific generators.
- $\mathbb{R}^{1\times 6}$ denotes a 1x6-dimensional real vector space.

Our approach involves the utilization of a ResNet-18 architecture represented as R as the backbone with dual classification layers. Specifically, for C_2, is designed to promote misclassification of subclasses.

$$L_{C_2} = \sum [C_2(x) \neq \text{True Subclass}],$$

where I is an image, and True Subclass is the correct subclass label.

Table 1: Data Generated using the generative models in the problem statement

	StyleGAN2	StyleGAN3	Latent-Diffusion	Taming-Transformers	LSGM	Guided Diffusion
1	FFHQ	Afhqv2	FFHQ	FFHQ	CelebAHQ	CelebAHQ
	LSUN-Car	FFHQ	CelebAHQ	CelebAHQ	-	LSUN-Bed
	LSUN-Cat	MetFaces	LSUN-Bed	Image-Net Cats	-	LSUN-Cat
	LSUN-Church	-	LSUN-Churches	-	-	LSUN-Horse
	LSUN-Horse	-	-	-	-	-

Table 2: Quantity of Data Generated using the generative models in the problem statement column and row datasets are the same as table 1 (numbers are approximate)

	StyleGAN2	StyleGAN3	Latent-Diffusion	Taming-Transformers	LSGM	Guided Diffusion
1	5000	5000	5000	5000	10000	5000
	1000	5000	3000	3000	-	1000
	1000	2000	1000	5000	-	1000
	1000	-	1000	-	-	1000
	1000	-	-	-	-	-

The high-level idea is to blind the model of the subclass information of images at runtime using an adversarial approach. To achieve this, a subclass classifier, such as ImageNet, is employed to generate adversarial versions of test dataset images. These adversarial images are then fed into a generator attribution model. The hypothesis is that certain image features enable both the generator attribution model and ImageNet to predict subclasses, and by perturbing these features using adversarial attacks based on ImageNet, we can prevent the generator attribution model from leveraging subclass information for its predictions. This approach aims to enhance model privacy and reduce the reliance on subclass-related features, potentially leading to more robust and subclass-agnostic model performance.

3.1 Datasets and Experiments

To train our attribution model, we meticulously curate our own dataset, encompassing all six problem domains highlighted in our problem statement. Within this dataset, we generate images spanning various object categories, as delineated in Table 1. Our dataset curation is deliberate, aiming to encompass a diverse array of object categories. This diversity serves a specific purpose: allowing our model to discern the defining features that distinguish one generator model from another, rather than merely categorizing images based on their depicted objects. As an illustration, consider the case of LSUN car data, which is exclusive to StyleGAN2. In this context, our objective is clear: our attribution model must accurately attribute images to the appropriate generator model, even when presented with images of cars that were not generated by StyleGAN2 but by alternative models. In essence, our goal is to equip the attribution model with the ability to make precise associations between images and their respective generator models, irrespective of the object category depicted in each image.

With the dataset we assembled through our multi-headed adversarial model, we have achieved commendable performance across all six generators participating in the Semafor competition. The performance results are graphically depicted in Figure 2 for each of these generators. The evaluation metric considers the log likelihood ratio (LLR) scores of each model. This metric assesses the accuracy of our predictions, distinguishing between those we got right and those that appear to be incorrect and as shown in the graphs is depicted by happy box. Here three out of six generators hit the happy box for our model including taming transformers, LSGM and StyleGAN3. The other three models also were on par with the other submissions from other teams.

4 Conclusion

In conclusion, our research focuses on developing an end-to-end multi-head classifier that attributes images to specific generators. This approach has significant implications for addressing the misuse of synthetic media, providing accountability for content generation, and preserving digital information integrity.

References

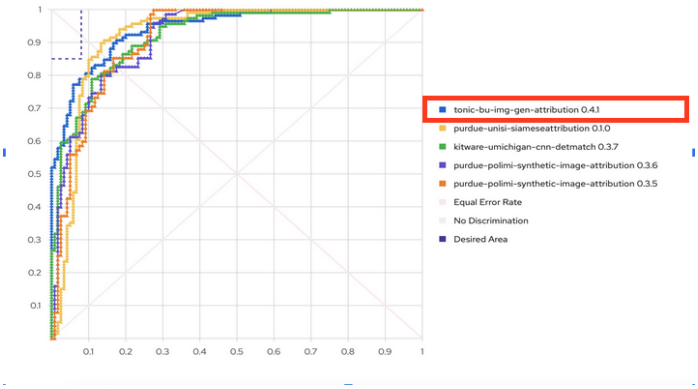
[1] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

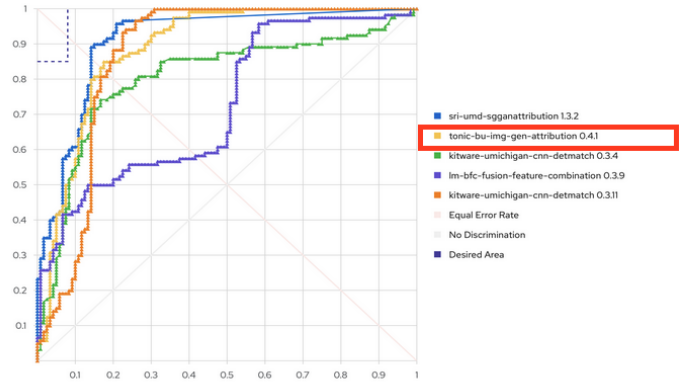
[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[4] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

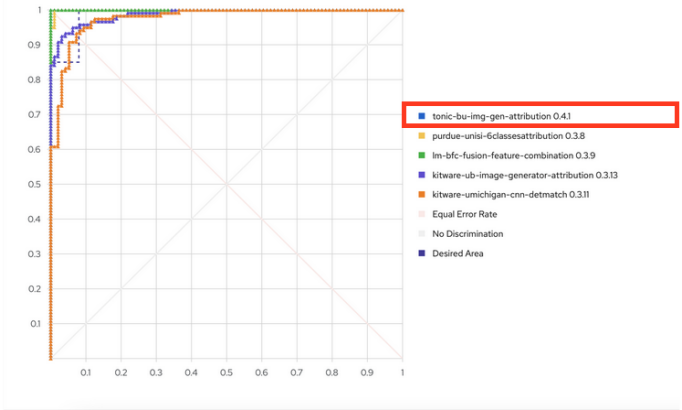
[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.



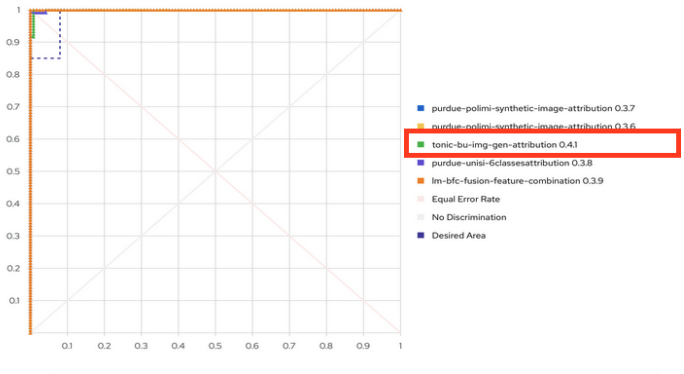
(a) Guided Diffusion score chart: semafor competition



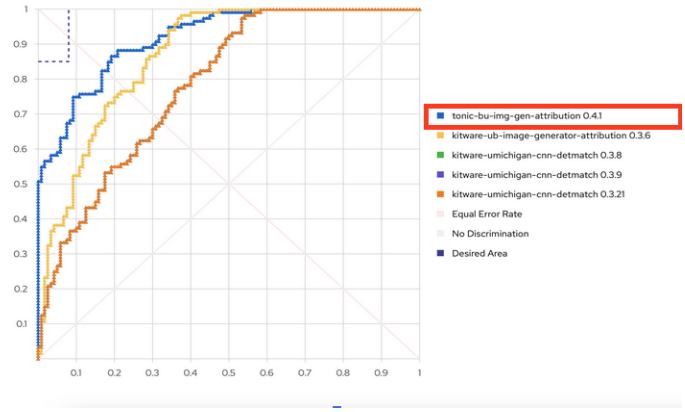
(b) Latent Diffusion score chart: semafor competition



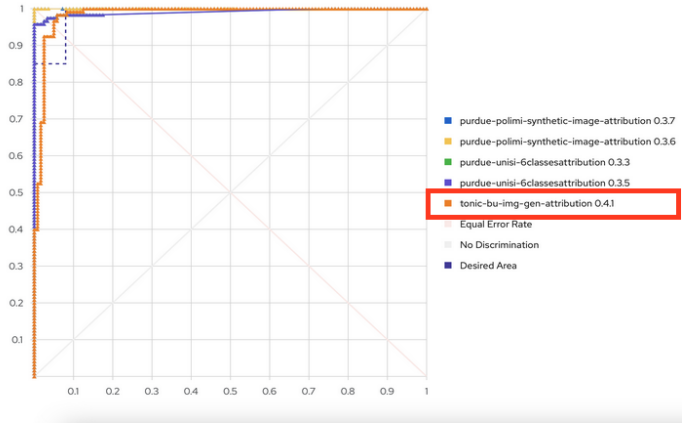
(c) LSGM score chart: semafor competition



(d) Taming Transformers score chart: semafor competition



(e) StyleGAN2 score chart: semafor competition



(f) StyleGAN3 score chart: semafor competition

Figure 2: Semafor Generatir Attribution Competition scores

[6] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

[7] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023.

[8] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

[9] Andreas Blattmann, Robin Rombach, Kaan Oktay, and Björn Ommer. Retrieval-augmented diffusion models, 2022.

[10] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[11] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[12] Brandon B. G. Khoo, Chern Hong Lim, and Raphael C. W. Phan. Transferable class-modelling for decentralized source attribution of gan-generated images, 2022.

[13] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis, 2021.

[14] Michael Albright and Scott McCloskey. Source generator attribution via inversion, 2019.

[15] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints, 2019.

[16] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020.

[17] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data, 2022.

[18] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints?, 2018.

[19] Sharath Girish, Saksham Suri, Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world gan generated images, 2021.