

AI-Generated Image Detection Using Semantic Feature

Hau-Ching Chen*, Yu-Pin Shen*, Yi-Hung Yeh*, Jiann-Liang Chen*

*Department of Electrical Engineering, National Taiwan University of Science and Technology Taipei, Taiwan

M11207522@mail.ntust.edu.tw, M11207508@gapps.ntust.edu.tw, M11207510@mail.ntust.edu.tw,
lchen@mail.ntust.edu.tw

Abstract— In recent years, the rapid advancement of deep learning has brought generative models to the forefront of technological innovation, enabling the creation of highly realistic images from simple text inputs. However, this progress has also introduced new challenges in distinguishing real from synthetic content, a concern particularly evident in the misuse of such technologies for malicious purposes. This paper investigates the generalization capabilities of the CLIP Vision Transformer (ViT) model for fake image detection, leveraging its semantic understanding of images. We employ a cross-validation approach using the GenImage dataset to evaluate the performance of CLIP ViT against traditional models like ResNet-50 and Swin-T. Results demonstrate that CLIP ViT exhibits superior generalization ability on unseen datasets, outperforming other models by focusing on high-level semantic features rather than dataset-specific patterns. This study highlights the potential of CLIP ViT in advancing the detection of generative model outputs by capturing underlying semantic relationships, suggesting a paradigm shift towards more robust and generalized detection techniques.

Keywords— Deep Learning, CLIP Vision Transformer, Fake Image Detection, Semantic Understanding, Cross-Validation

I. INTRODUCTION

In recent years, deep learning technology has experienced rapid growth, and generative models have emerged as one of the most prominent advancements. These models are capable of producing highly realistic images and generating objects, landscapes, and even artistic styles from simple text inputs. They have the potential to disrupt industries like entertainment and design. However, this technology has also sparked considerable controversy and fear. In Taiwan, there have been instances where deepfake technology was exploited for fraudulent activities, leveraging the reputation of public figures to deceive the public, causing significant damage to the individuals' reputations and fostering widespread distrust.

If proactive measures are not taken to address similar issues, this could have profound implications for society. Research [1] has indicated that the average human's ability to distinguish fake images is only about 60%, whereas the most advanced detection models can achieve up to 90% accuracy. This highlights that while the human eye struggles to catch subtle errors in highly realistic fake images, a well-trained deep-learning model can identify them. However, given the

continuous evolution and improvement of generative models, developing more powerful and accurate detection technologies is becoming increasingly urgent.

Training a model to distinguish between real and fake images differs fundamentally from traditional detection models. Conventional detection models typically require only the collection of a dataset related to the target subject to complete training. However, for fake detection models, it is practically challenging to gather all fake images generated by various models, due to the vast amount of data involved, the fact that some models remain proprietary, or the emergence of new generative models in the future. Additionally, different generative models may possess distinct synthesis characteristics, which makes it difficult for fake detection models trained using traditional methods to generalize effectively. These models often utilize convolution networks, which excel at recognizing patterns, even the intrinsic ones, cannot hence handle fake images generated by previously unseen models, rendering them unsuitable for the complex, dynamic real-world environment.

This paper proposed a cross-validation approach to evaluate the capability of a model to generalize, wherein the detection model is trained on a dataset generated by a specific model and then validated on datasets from other generative models. This method aims to simulate the real-world scenario where the detection model encounters fake images it has never seen, thereby assessing the model's generalization performance in recognizing fake content.

This paper will focus on validating the use of CLIP [2] as the backbone for fake image detection. Due to its large-scale pairing of real images and text, CLIP [2] demonstrates exceptional capabilities at extracting semantic meaning from images. We propose that this capability might mean the classification head we attached after CLIP ViT can find and classify images by the semantic meanings, a more complex idea than the simple intrinsic pattern, and thus able to generalize itself to future generation models.

II. RELATED WORK

A. Generation Models

GAN [7] (Generative Adversarial Networks), as one of the earliest generative models, has seen significant advancements in recent years. It operates by pitting a Generator against a

Discriminator in an adversarial process to produce synthetic images that are increasingly indistinguishable from real data. Over time, numerous improvements have been made to enhance the performance of GAN [7]. For instance, ProGAN [9] (Progressive Growing of GANs) introduces a training strategy that progressively increases the number of layers in both the generator and discriminator, starting from low-resolution images and gradually raising the model complexity as the resolution increases. This progressive training method ensures more stable training and leads to a substantial improvement in image quality.

Similarly, BigGAN [15] leverages larger model parameters and more extensive training datasets to generate higher-quality images. It also employs class-conditional generation, allowing the model to control the class of the generated images, thereby enhancing both the diversity and quality of the output. These advancements have significantly elevated the generative capabilities of GAN [7] models.

On the other hand, Diffusion [16] Models have impressed many with their outstanding performance by framing image generation as a denoising process. In the reverse process, Gaussian noise is gradually reconstructed into a complete image, producing visuals that closely resemble real data. Among these models, Latent Diffusion Models [17] (LDM) stand out by shifting the denoising process from the pixel space to the latent space, a compressed way of representing the original image. This shift significantly reduces computational costs, increases generation speed, and maintains high-quality results. Building upon LDM [17], the widely recognized Stable Diffusion has gained immense popularity for its simplicity and ability to produce high-quality images, making it a favorite among users. GLIDE [18] (Guided Language-to-Image Diffusion for Generation and Editing) incorporates text descriptions into the diffusion process, allowing images to gradually align with the content of the text during the denoising process.

B. Detection Methods

Currently, numerous studies focus on identifying synthetic images. For instance, [3] concentrates on detecting local artifacts and texture consistency within images, while [4] analyzes pixel-level patterns, learning common combinations of adjacent pixels in real images to identify forgeries. Studies like [5], [6], and [8] utilize the unique frequency-domain features of GANs [7], [9] to detect synthetic images. However, these approaches focus on low-level abstractions to identify image flaws, which are susceptible to being overcome by more advanced generative models. Moreover, GAN-specific detection models struggle to generalize to generation models based on Diffusion [10]. Even several studies, such as [11] and [12], aim to resolve frequency issues in GANs and have achieved notable results. Therefore, to gain an advantage in the ongoing competition between generative models and detection models, it is crucial to focus on higher-level abstractions that are more difficult for models to learn. Semantic-level detection of synthetic images is necessary. Studies like [13] and [14] leverage the rich semantic analysis capabilities of large-scale pre-trained visual-language models, such as CLIP [2], to

identify synthetic images, achieving outstanding detection and generalization performance.

C. Image Dataset

GenImage dataset, a million-scale fake image dataset, which ensures diversity by using real images from the ImageNet-1k dataset as a real image source and generates an equal number of fake images for each class. Each generated image corresponds to a real image in terms of class and quantity, maintaining class balance across the dataset. Importantly, the fake images are produced using eight different generative models—MidJourney, SD V1.4, SD V1.5, ADM [20], GLIDE [18], Wukong VQDM [19], and BigGAN [15]—with equal distribution of images across these methods. This approach mimics real-world scenarios and enhances the dataset's ability to reflect a wide range of generative techniques.

III. METHOD

In this experiment, we utilize the GenImage dataset for both training and validation. Given that CLIP's pre-training strategy aligns its Vision Transformer (ViT) output tokens with natural language representations, we expect the output tokens to capture the semantic meanings of the input images. To classify the images based on these extracted semantic features, we add a fully connected layer to the 768-dimensional token produced by the ViT model. This layer uses a sigmoid activation function to perform binary classification, with the classification threshold set to 0.5.

Each model will be trained on a single subset of the GenImage dataset and evaluated on all other subsets after each epoch to assess its generalization and classification ability.

IV. VALIDATION

The accuracy score for all other models in the above tables were directly taken from GenImage paper, with the score of our tested model attached in the last row. The training parameter thus might be different, especially the trained iteration. However, keep in mind, we only trained these models for one full epoch from the same data, our training result could only be worse than the original ones.

In Table 1, a notable difference between ResNet-50 and CLIP ViT was observed in internal cross-validation results. While ResNet-50 followed the typical pattern of highest performance on the dataset it was trained on, CLIP ViT, in contrast, showed higher performance on unseen datasets in several cases. This behavior challenges the usual understanding of model generalization, prompting further investigation into the underlying mechanisms.

Table 2 presents a cross-validation comparison between models trained on various datasets. Notably, while CLIP ViT's performance showed a decline when detecting fake images generated by all three variants of Stable Diffusion (SD V1.4, SD V1.5, and Wukong)—which are based on the same generative method—the model excelled when classifying images from other generative models. Despite the expected similarity among the Stable Diffusion variants, CLIP ViT demonstrated a strong ability to generalize across different

TABLE 1. INTERNAL CROSS-VAL COMPARISON BETWEEN CLIP ViT AND RESNET-50.

			Tested on								Avg Acc. (%)
			ADM	BigGAN	GLIDE	Midjourney	SD V1.4	SD V1.5	VQDM	Wukong	
Trained on	ADM	CLIP	87.8	88.8	75.6	65.8	78.4	77.4	89.8	81.4	80.6
		ResNet	99.0	88.3	97.1	58.6	53.1	53.2	61.5	53.0	70.5
	BigGAN	CLIP	61.2	96.1	70.7	61.3	59.5	60.6	83.7	61.6	69.3
		ResNet	50.6	99.9	68.4	49.9	49.9	49.9	50.6	49.9	58.6
	GLIDE	CLIP	66.2	86.3	96.1	82.3	84.9	84.6	78.3	82.8	82.7
		ResNet	56.0	74.0	99.9	50.7	50.0	50.1	51.0	50.3	60.3
	Midjourney	CLIP	53.3	73.6	69.9	62.5	57.7	57.0	49.4	54.0	59.7
		ResNet	64.1	50.1	78.9	98.8	76.4	76.9	52.5	71.4	71.1
	SD V1.4	CLIP	56.1	77.5	75.8	79.3	81.0	80.1	51.7	70.8	71.5
		ResNet	53.5	52.0	61.9	54.9	99.9	99.7	56.6	98.2	72.1
	SD V1.5	CLIP	55.8	75.9	77.4	82.0	81.1	81.4	51.1	71.0	72.0
		ResNet	52.7	51.3	60.1	54.4	99.8	99.9	56.9	98.5	71.7
	VQDM	CLIP	73.6	79.4	62.7	72.1	74.0	73.9	86.5	78.3	75.1
		ResNet	50.7	66.8	60.1	50.1	50.0	50.0	99.9	50.2	59.7
	Wukong	CLIP	55.8	80.1	78.7	73.7	78.4	78.4	57.8	78.6	72.7
		ResNet	51.4	50.9	58.3	54.5	99.7	99.6	58.7	99.9	71.6

TABLE 2. CROSS-VAL ACCURACY SCORE COMPARISON WHEN TRAINING ON STABLE DIFFUSION V1.4 DATA.

Method	Tested on								Avg Acc. (%)
	ADM	BigGAN	GLIDE	Midjourney	SD V1.4	SD V1.5	VQDM	Wukong	
ResNet-50	53.5	52	61.9	54.9	99.9	99.7	56.6	98.2	72.1
DeiT-S	49.8	53.5	58.1	55.6	99.9	99.8	56.9	98.9	71.6
Swin-T	49.8	57.6	67.6	62.1	99.9	99.8	62.3	99.1	74.8
CNNSpot	50.1	46.8	39.8	52.8	96.3	95.9	53.4	78.6	64.2
Spec	49.7	49.8	49.8	52	99.4	99.2	55.6	94.8	68.8
F3Net	49.9	49.9	50	50.1	99.9	99.9	49.9	99.9	68.7
GramNet	50.3	51.7	54.6	54.2	99.2	99.1	50.8	98.9	69.9
CLIP	56.1	77.5	75.8	79.3	81.0	80.1	51.7	70.8	71.5

datasets. This capability is particularly significant when compared to Swin-T, the top-performing model overall, as CLIP ViT outperformed Swin-T specifically on data from non-Stable Diffusion models. These results suggest that CLIP ViT is less dependent on specific model artifacts and more capable of recognizing the underlying semantics of fake images, a characteristic crucial for generalization across unseen data. As further shown in Table 3, CLIP ViT outperformed all other models referenced in the GenImage paper, particularly in terms of its ability to generalize based on learned semantic meanings rather than relying on dataset-specific patterns.

The ability of CLIP ViT to outperform on unseen datasets suggests that it is not merely learning to identify dataset-specific artifacts, but is instead learning a more abstract, generalized representation of what constitutes a fake image. We hypothesize that this is indicative of CLIP ViT's capacity to extract semantic meaning from images—a higher-level understanding that transcends individual dataset characteristics and can generalize to new types of generative models.

Although direct evidence of semantic feature extraction (e.g., attention maps) is beyond the scope of this study, the observed performance pattern aligns with CLIP's known ability to capture semantic relationships. Future work may explore this hypothesis further through targeted experiments such as adversarial attacks or feature visualization techniques.

In summary, the performance of CLIP ViT, particularly its ability to outperform on unseen datasets, provides strong evidence that the model is learning higher-level semantic features. This marks a potential shift in fake image detection, where the ability to generalize across diverse datasets may hinge on a model's capacity to understand the underlying semantics of what makes an image fake.

V. CONCLUSION AND FEATURE WORKS

In this paper, we demonstrated that CLIP ViT holds significant potential for detecting AI-generated fake images, particularly toward those created by models the detection network has not encountered during training. Our results suggest that CLIP ViT is capable of learning higher-level

TABLE 3. CROSS-VAL ACCURACY SCORE OVERALL COMPARISON.

Method	Trained on								Avg Acc. (%)
	ADM	BigGAN	GLIDE	Midjourney	SD V1.4	SD V1.5	VQDM	Wukong	
ResNet-50	59.7	66.6	73.1	59	72.3	72.4	60.9	71.4	66.9
DeiT-S	59.5	66.3	71.1	60.7	74.2	74.2	61.7	73.1	67.6
Swin-T	61.3	69.5	76.9	61.7	76	76.1	65.8	75.1	70.3
CNNSpot	57	56.6	57.1	58.2	70.3	70.2	56.7	67.7	61.7
Spec	57.9	64.3	65.4	56.7	72.4	72.3	61.7	70.3	65.1
F3Net	66.5	56.5	57.8	55.1	73.1	73.1	62.1	72.3	64.6
GramNet	58.7	61.2	65.3	58.1	72.8	72.7	57.8	71.3	64.7
CLIP	80.6	69.3	82.7	59.7	71.5	72.0	75.1	72.7	72.9

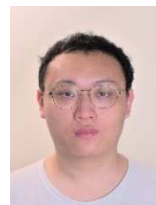
semantic features, which allows it to generalize across different generative models more effectively than traditional architectures such as ResNet-50. Unlike models that perform well only on the datasets they were trained on, CLIP ViT exhibited a stronger capacity to detect fake images in previously unseen data, providing compelling evidence that the model is capturing abstract concepts inherent to fake images.

This behavior contrasts with the dataset-specific performance of ResNet-50, which suggests that our approach, using CLIP ViT, moves beyond the identification of superficial artifacts and towards understanding the semantic structure of fake content. This result supports our hypothesis that the ability to extract these higher-level features is crucial for robust generalization in fake image detection.

Our future work will aim to deepen our understanding of CLIP ViT's ability to extract semantic features. We plan to explore more sophisticated methods, such as adversarial attacks, to test the model's robustness and further validate its generalization capabilities. Additionally, investigating the inner workings of the model through attention map visualization and feature analysis will help uncover how these semantic meanings are represented and utilized during classification. We believe this line of research will contribute to the development of more resilient detection systems capable of adapting to rapidly evolving generative models.

REFERENCES

- [1] Z. Lu, D. Huang, L. Bai, J. Qu, X. Liu, and W. Ouyang, "Seeing is not always believing: A quantitative study on human perception of ai-generated images," *NeurIPS*, 2023.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, "Learning transferable visual models from natural language supervision," *Proceedings of the International Conference on Machine Learning*, pp. 8748-8763, PMLR, 2021.
- [3] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. S. Manjunath, "Detecting GAN generated fake images using co-occurrence matrices," *arXiv preprint arXiv:1903.06836*, 2019.
- [4] L. Chai, D. Bau, S. N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," *Proceedings of the Computer Vision-ECCV 2020: 16th European Conference*, 2020.
- [5] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," *Proceedings of the International Conference on Machine Learning*, pp. 3247-3258, PMLR, 2020.
- [6] X. Zhang, S. Karaman, and S. F. Chang, "Detecting and simulating artifacts in gan fake images," *Proceedings of the 2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1-6, 2019.
- [7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, ... and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, iss. 11, pp. 139-144, 2020.
- [8] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695-8704, 2020.
- [9] T. Karras, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, pp. 6840-6851, 2020.
- [11] Z. Li, P. Xia, X. Rui, and B. Li, "Exploring the effect of high-frequency components in gans training," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, iss. 5, pp. 1-22, 2023.
- [12] S. J. Moon, C. Kim, and G. M. Park, "WaGI: wavelet-based GAN inversion for preserving high-frequency image details," *arXiv preprint arXiv:2210.09655*, 2022.
- [13] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480-24489, 2023.
- [14] M. Zhu, H. Chen, M. Huang, W. Li, H. Hu, J. Hu, and Y. Wang, "Gendit: Towards good generalizations for ai-generated image detection," *arXiv preprint arXiv:2312.08880*, 2023.
- [15] A. Brock, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
- [16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684-10695, 2022.
- [17] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, ... and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.
- [18] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, ... and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696-10706, 2022.
- [19] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780-8794, 2021.



Hau-Ching Chen was born in Taiwan in 1999. He received a B.S. degree in Electronic Engineering in 2023. He is currently pursuing an M.S. degree in Electrical Engineering at the National Taiwan University of Science and Technology (NTUST), Taipei. His main research interests include Deep Learning and Computer Vision.



Yu-Pin Shen was born in Taiwan in 1999. He received a B.S. degree in Electronic Engineering in 2022. He is currently pursuing an M.S. degree in Electrical Engineering at the National Taiwan University of Science and Technology (NTUST), Taipei. His main research interests include Artificial Intelligence, the Internet of Things (IoT), AI Image Detection, and Cybersecurity.



Yi-Hung Yeh was born in Taiwan in 2000. He received a B.S. degree in Electronic Engineering in 2022. He is currently pursuing an M.S. degree in Electrical Engineering at the National Taiwan University of Science and Technology (NTUST), Taipei. His main research interests include Artificial Intelligence, Human-Computer Interaction and Computer Vision.



Jiann-Liang Chen (Senior Member, IEEE) was born in Taiwan on December 15, 1963. He received a Ph.D. in Electrical Engineering from the National Taiwan University, Taipei, Taiwan, in 1989. Since August 1997, he has been with the Department of Computer Science and Information Engineering of National Dong Hwa University, where he is a professor and Vice Dean of Science and Engineering College. Prof. Chen joins the Department of Electrical Engineering, National Taiwan University of Science and Technology, as a Distinguished Professor and Dean. His research interests are cellular mobility management, cybersecurity, personal communication systems, and the Internet of Things (IoT).