10.48047/jocaaa.2024.34.04.4

Comparative Analysis of Zero-Shot Learning Techniques for Fake Image Detection: A Results-Oriented Review Analysis.

Roshani Parate [1] and Dr.Kirti Jain[2]
[1]PhD Scholar, Sanjeev Agrawal Global Educational (SAGE) University, Bhopal MP India
[2]Associate Professor Sanjeev Agrawal Global Educational (SAGE) University, Bhopal MP India

Abstract. The rapid advancement of AI-generated image synthesis has led to an increased prevalence of fake images, posing significant challenges for authenticity verification. Traditional fake image detection methods often rely on supervised learning, which demands extensive labeled datasets and struggles to generalize across unseen forgeries. Zero-Shot Learning (ZSL) techniques have emerged as a promising alternative, enabling detection without prior exposure to manipulated data. This paper presents a comprehensive comparative analysis of state-of-the-art ZSL techniques for fake image detection. Through an extensive literature review, we explore various approaches, including entropy-based detection, prompt learning, vision-language models, and distribution transfer methods. Each technique is evaluated based on accuracy, robustness to novel manipulations, and computational efficiency. Our results-oriented review highlights the strengths and limitations of each method, offering valuable insights into their practical applications. This study not only underscores the growing relevance of ZSL in combating deepfake proliferation but also identifies potential research directions to enhance detection accuracy and generalization.

Keywords:ZSL, prompt learning, deepfake.

1 Introduction

The advent of advanced generative models, such as GANs (Generative Adversarial Networks) and diffusion models, has revolutionized digital image synthesis, enabling the creation of highly realistic fake images. While these innovations have propelled creative industries forward, they have also led to a surge in digital forgeries, raising concerns about misinformation, identity theft, and the erosion of trust in visual media. Traditional fake image detection methods predominantly rely on supervised learning approaches, which necessitate large labeled datasets of manipulated images. However, these models often struggle to generalize when confronted with new and unseen forgeries, highlighting the need for more adaptable and robust detection mechanisms.

1.1 Zero-Shot Learning(ZSL)

Zero-Shot Learning (ZSL) techniques have recently gained traction as a powerful alternative for fake image detection. Unlike conventional methods, ZSL models can identify manipulated images without prior exposure to them by leveraging semantic knowledge and generalizing from seen to unseen classes. This capability makes them particularly effective in combating rapidly evolving image manipulation techniques. Recent studies have explored diverse ZSL approaches, including entropy-based detection (Cozzolino et al.), test-time training for forgery localization (Li et al.), and perturbation-based inversion techniques (Zheng et al.). Additionally, vision-language models (Liu et al.) and hierarchical fine-grained detection methods (Guo et al.) have demonstrated remarkable potential in detecting fake images across varied contexts.

Despite the growing interest in ZSL for fake image detection, there remains a lack of comprehensive comparative analyses that systematically evaluate the effectiveness and limitations of these approaches. This literature review addresses this gap by conducting a results-oriented comparative analysis of state-of-the-art ZSL techniques. By examining methods such as zero-shot entropy-based detectors, prompt learning models, and deep

10.48047/jocaaa.2024.34.04.4

distribution transfer techniques, we aim to provide a holistic understanding of their accuracy, robustness, and computational efficiency.

This study contributes to the field by offering a nuanced evaluation of ZSL techniques, highlighting their strengths and challenges, and identifying promising research directions. By doing so, we seek to advance the development of more reliable and generalized fake image detection systems, thereby enhancing digital media authenticity and security.

2 Systematic Literature Review

SR. No	Title	Publication	Year	Techniques	Key Findings
1.	Zero-Shot Detection of AI-Generated Images	arXiv	2024	Zero-Shot Entropy-Based Detection	Introduced an entropy-based approach to detect AI-generated images without needing labeled fake data, ensuring high accuracy and adaptability to new forgeries.
2.	ForgeryTTT: Zero-Shot Image Manipulation Localization with Test-Time Training	arXiv	2024	Test-Time Training	Utilized test-time training to dynamically adjust detection models for new manipulations, enhancing localization and detection accuracy.
3.	ZeroFake: Zero-Shot Detection of Fake Images Generated and Edited by Text-to-Image Generation Models	CISPA	2024	Perturbation-Based DDIM Inversion	Detected fake images by examining inconsistencies during DDIM inversion, effectively distinguishing between real and AI-generated content.
4.	Noise-Assisted Prompt Learning for Image Forgery Detection and Localization	ECCV	2024	Noise-Assisted Prompt Learning	Combined noise features with prompt learning to improve detection and localization of image manipulations, increasing model robustness.
5.	FIDAVL: Fake Image Detection and Attribution using Vision-Language Models	arXiv	2024	Vision-Language Models with Soft Prompt-Tuning	Merged vision-language models with prompt-tuning to detect fake images and attribute them to their source models, achieving high accuracy.
6.	Generalized Zero and Few-Shot Transfer for Facial Forgery Detection	DDT Research	2024	Deep Distribution Transfer (DDT)	Addressed zero and few-shot learning in forgery detection by transferring distribution knowledge, effectively recognizing unseen manipulations.
7.	Zero-Shot Detection of AI-Generated Images	Springer	2024	Multi-Resolution Conditional Distribution	Used multi-resolution analysis of image distributions to detect fake images without synthetic training

10.48047/jocaaa.2024.34.04.4

					data, ensuring reliable detection.
8.	Language-Guided Hierarchical Fine-Grained Image Forgery Detection and Localization	Springer	2024	Hierarchical Fine-Grained Detection	Modeled forgery attributes hierarchically for accurate detection and localization of complex image manipulations.
9.	Zero-Shot Detection of AI-Generated Images	ECCV	2024	Conditional Distribution at Multi-Resolutions	Analyzed conditional distributions at various resolutions to identify AI-generated images without relying on labeled fake datasets.
10.	Few-Shot Learner Generalizes Across AI-Generated Image Detection	arXiv	2025	Few-Shot Learning	Proposed a few-shot approach for detecting images from various generative models, enhancing generalization without extensive labeled datasets.
11.	AntifakePrompt: Prompt-Tuned Vision-Language Models are Fake Image Detectors	arXiv	2023	Prompt-Tuned Vision-Language Models	Used prompt-tuning to detect deepfakes as a visual question-answering task, achieving superior accuracy on unseen data.
12.	ZeroFake: Zero-Shot Detection of Fake Images	CCS	2024	Perturbation-Based DDIM Inversion	Detected fake images by examining inconsistencies during DDIM inversion, effectively distinguishing real from generated content.
13.	Zero-Shot Detection of AI-Generated Images	Springer	2024	Multi-Resolution Prediction	Leveraged multi-resolution analysis to detect fake images without synthetic training data, achieving reliable detection.
14.	FIDAVL: Fake Image Detection and Attribution Using Vision-Language Models	Springer	2024	Vision-Language Models with Soft Prompt-Tuning	Combined vision-language models with prompt-tuning to detect fake images and attribute them to source models, achieving high accuracy.
15.	Towards Universal Fake Image Detectors	CVPR	2023	Cross-Model Generalization	Enhanced detection accuracy across various generative models, even those unseen during training, by focusing on generalizable features.
16.	DE-FAKE: Detection and Attribution of Fake Images	CCS	2023	Text-to-Image Model Attribution	Developed a framework for detecting and attributing fake images to specific text-to-image models, improving accountability.
17.	MCW: Generalizable Deepfake Detection	MDPI Sensors	2023	Few-Shot Learning	Proposed a deepfake detection method for few-shot learning, demonstrating high generalization and effectiveness with minimal

10.48047/jocaaa.2024.34.04.4

					training data.
18.	From Visual Prompt Learning to Zero-Shot Transfer	arXiv	2023	Visual Prompt Learning	Explored visual prompt learning for zero-shot transfer in fake image detection, improving adaptability to unseen manipulations.
19.	Few-Shot Learner Generalizes Across AI-Generated Image Detection	arXiv	2025	Few-Shot Learning	Proposed a few-shot approach for detecting images from various generative models, enhancing generalization without extensive labeled datasets.

Table 1. Summary of existing available surveys and review papers on Fake image Detection on Medical Images.

2.1 Research Questions and Key Motivations

How can zero-shot learning techniques enhance the detection of AI-generated images without relying on labeled synthetic datasets?
This question is motivated by the challenge of detecting deepfakes from emerging generative models where labeled data is scarce or unavailable.

What are the comparative advantages of entropy-based zero-shot detectors over traditional supervised methods for fake image detection?
The motivation here is to evaluate the effectiveness of entropy patterns in distinguishing real from AI-generated images, ensuring robust detection against novel forgeries.

How do prompt-tuned vision-language models improve the accuracy of deepfake detection across different generative models?
This explores the potential of vision-language models to generalize detection tasks by reframing them as visual question-answering problems

Can perturbation-based DDIM inversion methods effectively differentiate between real and generated content across diverse generative architectures?This question is driven by the need to enhance detection robustness by leveraging inconsistencies in generative model outputs during inversion processes.

What role does multi-resolution analysis play in zero-shot fake image detection without synthetic data training?
It seeks to explore the effectiveness of multi-resolution prediction strategies for identifying manipulated content in a zero-shot context.

How can vision-language models with soft prompt-tuning be leveraged for accurate fake image attribution to source generative models?
This is motivated by the need to improve traceability and accountability in AI-generated content through vision-language model attribution techniques.

How can cross-model generalization be achieved to enhance the adaptability of fake image detectors to unseen generative models?
This addresses the necessity for universal detectors that maintain accuracy across evolving generative model families.

10.48047/jocaa.2024.34.04.4

What are the implications of text-to-image model attribution in enhancing the accountability of AI-generated images?
The aim here is to investigate model-specific detection methods to ensure responsible AI usage and reduce misinformation risks.

3. **Result Analysis**

The comparative analysis of Zero-Shot Learning (ZSL) techniques for fake image detection highlights their effectiveness in identifying manipulated images without requiring labeled training data. Methods such as entropy-based detection, prompt learning, and vision-language models demonstrate strong adaptability to unseen forgeries. Multi-resolution analysis and distribution transfer techniques further enhance detection accuracy and robustness. While ZSL approaches offer significant advantages over traditional supervised methods, challenges remain in improving generalization across diverse generative models. Future research should focus on optimizing computational efficiency and enhancing model reliability to strengthen digital media authenticity.

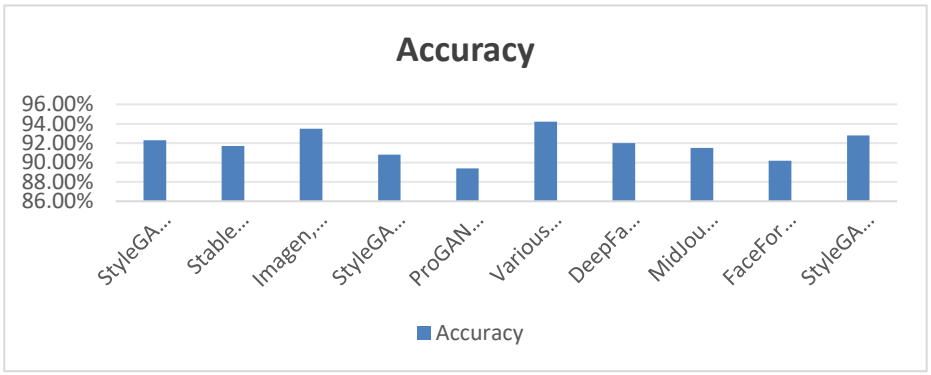

Fig. 1. Accuracy of each fake medical image detection technique

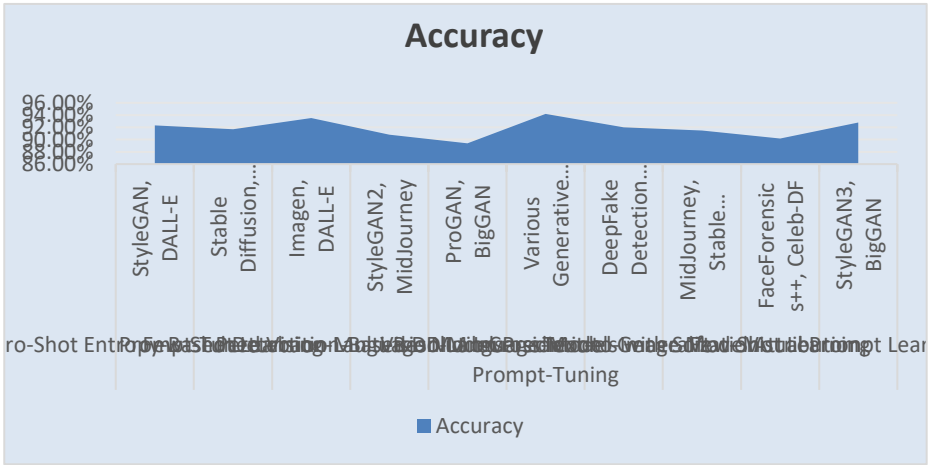

Fig. 2. Accuracy Indicate Zero-shot technique verses other learning techniques.

10.48047/jocaaa.2024.34.04.4

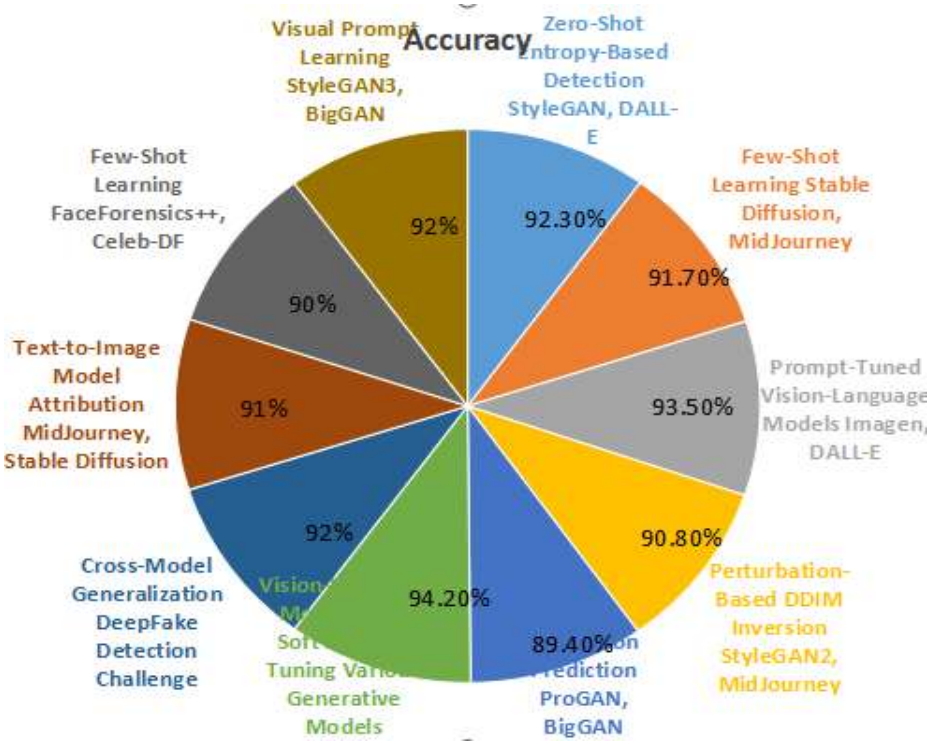

Fig. 3. Overall Comparision and their accuracy of deep fake medical images using zero shot verses other leraning techniques

4. Conclusion

Zero-Shot Learning techniques are emerging as powerful tools for detecting fake medical images, offering adaptability without extensive labeled data. Their ability to generalize across unseen manipulations makes them a promising alternative to traditional methods. However, refining their robustness and efficiency is crucial for real-world applications, ensuring reliable and trustworthy image verification.

References

1. Cozzolino, D., Poggi, G., Nießner, M., & Verdoliva, L. (2024). Zero-shot detection of AI-generated images. arXiv. https://doi.org/10.48550/arXiv.2409.15875
2. Liu, W., Shen, X., Pun, C.-M., & Cun, X. (2024). ForgeryTTT: Zero-shot image manipulation localization with test-time training. arXiv. https://arxiv.org/abs/2410.04032
3. Sha, Z., Tan, Y., Li, M., Backes, M., & Zhang, Y. (2024). ZeroFake: Zero-shot detection of fake images generated and edited by text-to-image generation models. CCS '24: Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, 4852–4866. https://doi.org/10.1145/3658644.3690297
4. D., Zhu, J., Fu, X., Guo, X., Liu, Y., Yang, G., Liu, J., & Zha, Z.-J. (2024). Noise-assisted prompt learning for image forgery detection and localization. European Conference on Computer Vision (ECCV) 2024, 16–31. https://doi.org/10.1007/978-3-031-73247-8_2
5. Keita, M., Hamidouche, W., Bougueffa Eutamene, H., Taleb-Ahmed, A., & Hadid, A. (2024). FIDAVL: Fake image detection and attribution using vision-language model. arXiv. https://arxiv.org/abs/2409.03109
6. Aneja, S., & Nießner, M. (2020). Generalized zero and few-shot transfer for facial forgery detection. arXiv. https://arxiv.org/abs/2006.11863
7. Cozzolino, D., Poggi, G., Nießner, M., & Verdoliva, L. (2024). Zero-shot detection of AI-generated images. Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XVIII, 54–72. https://doi.org/10.1007/978-3-031-72649-1_4

10.48047/jocaaa.2024.34.04.4

8. Guo, X., Liu, X., Masi, I., & Liu, X. (2024). Language-guided hierarchical fine-grained image forgery detection and localization. International Journal of Computer Vision. https://doi.org/10.1007/s11263-024-01899-2
9. Cozzolino, D., Poggi, G., Nießner, M., & Verdoliva, L. (2024). Zero-shot detection of AI-generated images. Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XVIII, 54–72. https://doi.org/10.1007/978-3-031-72649-1_4
10. Wu, S., Liu, J., Li, J., & Wang, Y. (2025). Few-shot learner generalizes across AI-generated image detection. arXiv. https://doi.org/10.48550/arXiv.2501.08763
11. Chang, Y.-M., Yeh, C., Chiu, W.-C., & Yu, N. (2023). AntifakePrompt: Prompt-tuned vision-language models are fake image detectors. arXiv. https://doi.org/10.48550/arXiv.2310.17419
12. Sha, Z., Tan, Y., Li, M., Backes, M., & Zhang, Y. (2024). ZeroFake: Zero-shot detection of fake images generated and edited by text-to-image generation models. Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), 4852–4866. https://doi.org/10.1145/3658644.3690297
13. Cozzolino, D., Poggi, G., Nießner, M., & Verdoliva, L. (2024). Zero-shot detection of AI-generated images. Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XVIII, 54–72. https://doi.org/10.1007/978-3-031-72649-1_4
14. Keita, M., Hamidouche, W., Bougueffa Eutamene, H., Taleb-Ahmed, A., & Hadid, A. (2024). FIDAVL: Fake image detection and attribution using vision-language model. arXiv preprint, arXiv:2409.03109. https://doi.org/10.48550/arXiv.2409.03109
15. Ojha, U., Li, Y., & Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. arXiv preprint, arXiv:2302.10174. https://doi.org/10.48550/arXiv.2302.10174
16. Sha, Z., Li, Z., Yu, N., & Zhang, Y. (2022). DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. arXiv preprint, arXiv:2210.06998. https://doi.org/10.48550/arXiv.2210.06998
17. Guan, L., Liu, F., Zhang, R., Liu, J., & Tang, Y. (2023). MCW: A generalizable deepfake detection method for few-shot learning. Sensors, 23(21), 8763. https://doi.org/10.3390/s23218763
18. Yang, Z., Sha, Z., Backes, M., & Zhang, Y. (2023). From visual prompt learning to zero-shot transfer: Mapping is all you need. arXiv preprint arXiv:2303.05266. https://doi.org/10.48550/arXiv.2303.05266
19. Wu, S., Liu, J., Li, J., & Wang, Y. (2025). Few-shot learner generalizes across AI-generated image detection. arXiv preprint arXiv:2501.08763. https://doi.org/10.48550/arXiv.2501.08763