INFORMATION
Volume 29, Number 1, pp.139-145

ISSN 1343-4500 eISSN 1344-8994

Visual Identification of AI-Generated Disaster Images for Fake News Verification

Ziqi Chen*, Jiacheng Li**, Masato Noto**

* *Field of Electrical, Electronics and Information Engineering,*
Graduate School of Engineering, Kanagawa University, Yokohama, Japan
E-mail: r202570164ri@jindai.jp

** *Department of Applied Systems and Mathematics, Kanagawa University, Yokohama, Japan*
E-mail: {lijiacheng, noto}@kanagawa-u.ac.jp

Abstract

Recent advances in generative artificial intelligence (AI) have accelerated the spread of AI-generated images on social platforms, particularly in disaster-related misinformation. Such visual content can mislead the public and amplify social panic. However, traditional image forensic techniques show limited robustness when facing diverse generative models and post-processing operations.

This study proposes a lightweight detection approach that combines CLIP visual embedding with a logistic regression classifier to distinguish real disaster images from AI-generated ones. Experiments on a disaster-domain dataset demonstrate that the proposed method achieves approximately 98% accuracy and shows clear linear separability in feature visualization. The results indicate that CLIP features are highly discriminative even under small-sample conditions, providing a practical foundation for multimodal fake news verification.

Keywords : AI-Generated Images, CLIP, Fake News Verification, Logistic Regression, Disaster Dataset

1 Introduction

The rapid development of generative AI, particularly diffusion models such as Stable Diffusion and Midjourney, has resulted in a growing volume of AI-generated images circulating on social platforms and news media. Unlike textual misinformation, visual misinformation tends to spread faster and provoke stronger emotional reactions, especially during disaster events, where photorealistic images can easily mislead the public and amplify social anxiety . International fact-checking organizations have repeatedly emphasized that disaster-related visual content is among the most problematic categories of misinformation due to its realism and emotional intensity [1][2]. These concerns highlight the need for reliable detection of AI-generated disaster images as part of fake-news verification.

Traditional image forensic techniques rely on pixel- or frequency-domain features, including noise inconsistency analysis and artifact inspection [3][4]. However, previous work has shown that these handcrafted cues lack robustness across generative models and degrade significantly under common post-processing operations such as compression or scaling [5][6]. To overcome these limitations, recent studies have shifted toward feature-based detection using pretrained models. In particular, CLIP has demonstrated strong cross-domain transferability and robustness, with Cozzolino et al. reporting over 90% AUC even when trained on small (1k) datasets and under compression or geometric distortions [8]. At the same time, lightweight linear models such as logistic regression have been shown to offer stable and computationally efficient inference in resource-constrained scenarios, making them suitable for practical security or verification tasks [9]. These findings suggest that high-level semantic features extracted by CLIP may already encode discriminative differences between real and synthetic imagery, allowing effective classification with simple linear decision boundaries.

Although previous studies have demonstrated the effectiveness of CLIP embeddings and lightweight linear classifiers, several gaps remain unaddressed. Existing CLIP-based detectors are typically evaluated on general or multi-domain datasets, and it is unclear whether their conclusions hold in disaster-specific scenarios, where visual patterns tend to be more concentrated and emotionally charged. Moreover, prior work has not fully examined the stability of lightweight classifiers under realistic perturbations such as social-media compression or noise contamination. These limitations suggest that the practical viability of CLIP-based lightweight detection methods in disaster-related misinformation remains insufficiently explored.

Motivated by these observations, this study investigates whether high-level CLIP visual features, combined with a logistic regression classifier, can provide reliable discrimination between real and AI-generated disaster images even under limited data conditions. Our approach evaluates both class separability and robustness by analyzing model behavior under clean, Gaussian-noisy, and JPEG-compressed inputs. By focusing on semantic embeddings and a simple linear decision boundary, this work aims to fill the methodological gap between general CLIP-based detectors and real-world disaster-oriented fake-news verification, providing an efficient and deployable baseline model.

2 Methodology

2.1 Dataset Construction

The real disaster images in this study were collected from open-source fire datasets such as the Kaggle Fire Dataset, resulting in approximately 1,000 real samples. AI-generated images were produced using text-to-image diffusion models. Stable Diffusion v1.5 was first employed with a unified news-style prompt, but the outputs showed structural and texture distortions in complex scenes. To improve realism, Realistic Vision v5 and Photorealistic_XL_v1.0 were adopted, yielding more consistent building geometry, object shapes, and lighting.

Prompts were refined with "realistic" "photo" and "news photo" and sampling parameters were tuned to increase diversity while maintaining a news-photography style. All generated images were manually screened, and low-quality or semantically inconsistent samples were removed. The final dataset consists of 1,500 synthetic and about 1,500 real images, resized to 512×512 and split into training and test sets with a 7:3 ratio.

2.2 Feature Extraction (CLIP Visual Embedding)

The CLIP ViT-B/32 model is used to extract visual embeddings. CLIP learns a joint image–text representation through contrastive training and provides strong transferability for downstream recognition tasks . Each image is processed by the CLIP visual encoder to obtain a 512-dimensional embedding, which is normalized and then used as input to the classifier. These global semantic features enable effective discrimination between real and AI-generated disaster images.This study uses the CLIP ViT-B/32 model to extract image feature vectors. CLIP learns a shared embedding space through contrastive learning between images and text, and its visual encoder has been proven to possess strong transferability in large-scale semantic alignment tasks [7][8].

Each image is input into the CLIP visual encoder to obtain a 512-dimensional embedding vector, which is then normalized and fed into the classifier for training. Since disaster images generally exhibit relatively consistent composition and color tone, the global features extracted by CLIP can effectively capture semantic differences, thereby distinguishing between real and AI-generated images.

2.3 Classifier Selection: Logistic Regression

Logistic regression is adopted as a lightweight and stable classifier in this study. Compared with support vector machines (SVMs) or deep neural networks (DNNs), logistic regression offers a simpler model structure and a more stable optimization process, especially when the features are linearly or approximately linearly separable [10][11]. The model takes the 512-dimensional CLIP embedding xas input and outputs the probability that an image belongs to the AI-generated class, defined as

$$y = \sigma\left(\sum_{j=1}^{d} w_j x_j + b\right) \tag{1}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Here, w and b denote the weight vector and bias term, respectively. The model output $y \in (0,1)$ is interpreted using a threshold of 0.5, and the parameters are trained using the standard cross-entropy loss.

A further advantage of logistic regression is its ease of deployment. The model contains few parameters, enables fast inference on CPUs or mobile devices, and avoids the computational and memory overhead associated with kernel-based SVMs or DNNs. This makes it suitable for practical fake-news verification scenarios where efficiency and lightweight operation are essential.

3 Experiments and Results

3.1 Experimental Setup

All experiments were conducted in a Windows environment equipped with an NVIDIA RTX 2060 GPU. All implementations were based on the PyTorch framework, and the CLIP ViT-B/32 model (pretrained weights: laion2b_s34b_b79k) was used to extract 512-dimensional visual embeddings.

The dataset was divided into training and test sets with a 7:3 ratio, and Stratified 5-Fold Cross Validation was applied to evaluate the robustness of the classifier. The logistic regression model was optimized using the LBFGS algorithm with a regularization parameter of C = 1.0 and a maximum of 5000 iterations. All experiments were repeated five times under the same settings, and the average value was reported as the final result.

To further verify the robustness of the model, two types of noise were introduced during the testing phase:

(a) JPEG compression noise (compression quality = 50)

(b) Gaussian noise (standard deviation $\sigma = 5$)

Under identical settings, CLIP features were extracted from both clean and noise-added images for classification.

3.2 Experimental Results

On the dataset containing 1,500 real and 1,500 AI-generated disaster images, the proposed CLIP + Logistic Regression model achieved an accuracy of 98% on the test set.

Table 1: Performance of the CLIP + Logistic Regression model on the test set.

Classification	**Precision**	**Recall**	**F1-score**
Real	0.99	0.97	0.98
AI	0.97	0.99	0.98
Accuracy		**0.982**	

The PCA visualization (Figure 1) shows two clearly separated clusters, indicating strong linear separability of CLIP embeddings. The confusion matrix (Figure 2) further confirms that the model maintains high precision and recall for both classes.

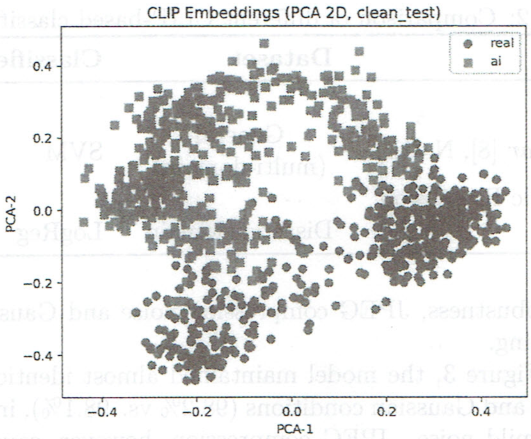

Figure 1: PCA scatter plot

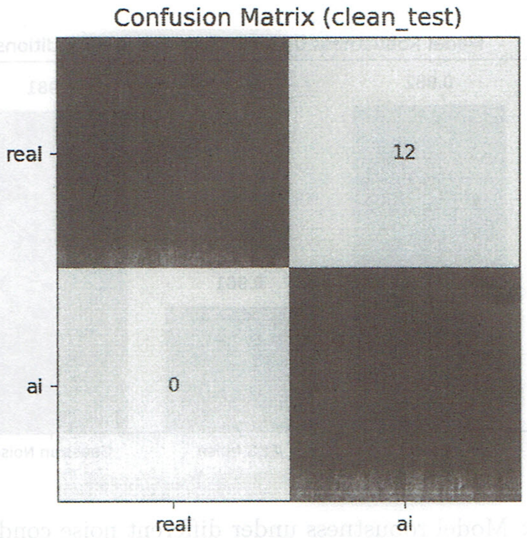

Figure 2: Confusion matrix

Following the experimental design of Cozzolino et al. [8], who showed that CLIP-based detectors trained on small datasets (1k) can approach the performance of 10k-scale training, this study adopted a similar configuration and achieved 98% accuracy on a domain-specific disaster dataset. This suggests that CLIP features remain highly discriminative even with limited data, likely because disaster images exhibit concentrated scene characteristics that enhance separability.

Table 2: Comparison of different CLIP-based classifiers.

Model	Dataset	Classifier	Result
CLIP + SVM (*Raising the Bar* [8], N=1k)	General (multi-domain)	SVM	0.90
CLIP + Logistic Regression (Ours)	Disaster-domain	LogReg	**0.98**

To evaluate robustness, JPEG compression noise and Gaussian noise were added during testing.

As shown in Figure 3, the model maintained almost identical performance between the clean and Gaussian conditions (98.2% vs. 98.1%), indicating strong stability against mild noise. JPEG compression, however, caused a larger reduction to 90.1%. A comparison between Gaussian and JPEG further shows that compression artifacts are significantly more disruptive than random noise. Overall, the model exhibits practical robustness, while performance under compression remains the main limitation.

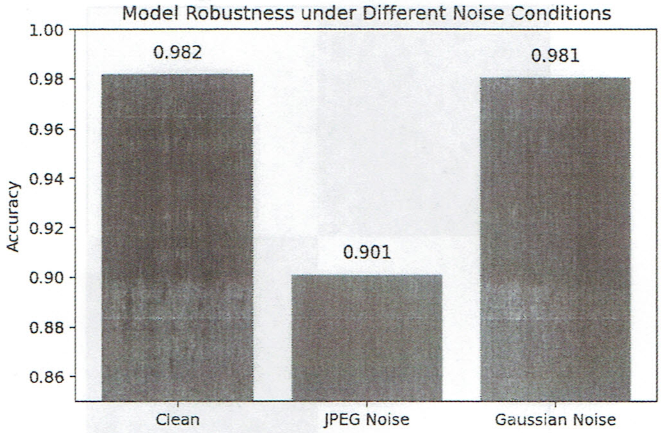

Figure 3: Model robustness under different noise conditions.

4 Concluding Remarks

The experimental results show that CLIP features provide strong linear separability for disaster-related image detection. Even with limited data, the logistic regression classifier achieves high accuracy, confirming the effectiveness of lightweight linear models compared with more complex approaches such as SVMs or DNNs. The PCA visualization further illustrates that CLIP embeddings naturally form two well-separated clusters with only a small number of boundary errors, indicating that the model captures semantic-level differences rather than relying solely on low-level artifacts.

In addition, the comparison with the CLIP-based detector of Cozzolino et al. suggests that domain-specific disaster imagery enhances feature consistency.

This allows CLIP embeddings to remain highly discriminative even under small-sample conditions, contributing to the strong performance observed in this study.

Overall, the results demonstrate that the proposed CLIP + Logistic Regression method is both effective and stable for identifying AI-generated disaster images. Future work will include expanding the dataset, evaluating cross-disaster generalizability, and incorporating textual information toward a unified multimodal fake-news detection framework.

References

[1] International Fact-Checking Network, "State of the Fact-Checkers Report 2024," Poynter Institute, Mar. 2025.

[2] U.S. Department of Homeland Security, "Countering false information on social media in disasters and emergencies," Mar. 2018.

[3] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, Nov. 2020, pp. 3247–3258.

[4] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR)*, San Jose, CA, USA, 2019, pp. 506–511.

[5] Z. Meng, B. Peng, J. Dong, T. Tan, and H. Cheng, "Artifact feature purification for cross-domain detection of AI-generated images," *Comput. Vis. Image Underst.*, vol. 247, 2024, p. 104078.

[6] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 24480–24489.

[7] W. Wu, Z. Sun, Y. Song, J. Wang, and W. Ouyang, "Transferring vision-language models for visual recognition: A classifier perspective," *Int. J. Comput. Vis.*, vol. 132, no. 2, 2024, pp. 392–409.

[8] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the bar of AI-generated image detection with CLIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Seattle, WA, USA, 2024, pp. 4356–4366.

[9] K. Rahman, M. S. Islam, and M. M. Rahman, "Cognitive lightweight logistic regression-based IDS for IoT networks," *Security Commun. Netw.*, vol. 2023, Article ID 7690322.

[10] A. Agrawal, S. Barratt, and S. Boyd, "Learning convex optimization models," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 8, Aug. 2021, pp. 1355–1364.

[11] G.-X. Yuan, K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "A comparison of optimization methods and software for l_1-regularized linear classification," *J. Mach. Learn. Res.*, vol. 11, 2010, pp. 3183–3234.