AI-Generated Disaster Image Detection: A Dataset and an Exploration of Key Challenges

Junjie Wang[1[0009-0008-3446-0677]] , Rui Ba[2], Kaiye Yu[3,4] , Han Xing[5] , Chao Sun[6], Hang Gao[1 ✉ [0000-0003-4797-3072]] and Mengting Hu[7[0000-0003-1536-5400]]

[1] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
[2] School of National Security, People's Public Security University of China, Beijing 100091, China
[3] School of Smart Governance, Renmin University of China, Suzhou 215123, China
[4]Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies, International College (Suzhou Research Institute), Renmin University of China, Suzhou 215123, China
[5] Shanghai Customs University, Shanghai 201204, China
[6] Information Research Institute, MEM, Beijing 100029, China
[7] College of Software, Nankai University, Tianjin 300350, China
2649215550@qq.com, barui@ppsuc.edu.cn, yukaiye@ruc.edu.cn, xinghan@shcc.edu.cn, 59791416@qq.com, mthu@nankai.edu.cn, gaohang@tust.edu.cn
✉The author is the corresponding author.

Abstract. The rapid advancement of generative artificial intelligence (AI) has enabled the creation of highly realistic disaster imagery, posing significant threats to information authenticity during crisis events. To address this emerging challenge, we present AIG-DI, the first dataset specifically designed for detecting AI-generated disaster images. The dataset is systematically constructed across multiple disaster types, image content categories, and generative models, enabling comprehensive evaluation under diverse conditions. We conduct an in-depth empirical study using two representative detectors—NPR (a spatial-domain detector) and FreDect (a frequency-domain detector)—to investigate their detection performance, generalization ability, and robustness under realistic perturbations. Experimental results reveal three key findings: (1) detectors trained on generic datasets struggle to detect AI-generated disaster imagery due to domain-specific texture and semantic shifts; (2) incorporating even a small amount of disaster-specific synthetic images during training significantly boosts accuracy and generalization, highlighting the value of domain-specific data for rapid adaptation; and (3) image perturbations remain a critical vulnerability, even with perturbationaware training. This work not only provides the first benchmark for AI-generated disaster image detection but also uncovers fundamental challenges in ensuring visual content authenticity for disaster response. The code and data have been open-sourced at https://github.com/gaohangcodes/AIG-DI.

Keywords: AI-generated disaster image, AI-generated disaster image detection, Benchmark.

1 Introduction

Natural disasters are characterized by sudden onset and immense destructive power, often causing severe disruptions to societal operations, economic activities, and human lives. In disaster emergency response, images and videos have become increasingly indispensable as intuitive carriers of information. They are widely used to convey on-site conditions, assess the severity of damage, and assist in decision-making. Moreover, visual content plays a critical role in guiding public opinion, as it can evoke emotional resonance, build societal trust in rescue operations, and encourage public participation.

However, with the rapid development of generative artificial intelligence (AI), AI-generated images have reached a level of realism capable of convincingly simulating disaster scenarios. Malicious or misleading content, crafted through such techniques, may spread false narratives, mislead public opinion and disrupt rescue decision-making. A notable case occurred following the magnitude 6.8 earthquake in Dingri County, Shigatse, Tibet, in January 2025. A Chinese internet user generated and circulated a highly emotional image of a "buried boy" [6] using AI tools, falsely associating it with the earthquake. This incident illustrates the urgent need for reliable detection technologies to ensure the authenticity of information and the scientific rigor of decision-making in disaster response.

To date, numerous studies have addressed the detection of AI-generated images, yielding a variety of approaches based on spatial-domain analysis [30, 27, 15, 14, 33, 17, 26] and frequency-domain analysis [7, 18, 20, 16, 32, 8, 10]. However, most existing methods are trained and evaluated on datasets containing regular objects or scenes—such as human faces, household items, and stylized artworks—that exhibit relatively simple visual patterns and minimal background interference. In contrast, disaster scenarios present a unique set of challenges: they often feature higher visual complexity and multimodal heterogeneity, including smoke and flame occlusion, water reflections, structural collapse, nonrigid textures, and varying illumination. These characteristics not only enhance the realism of AI-generated images but also increase the difficulty of their detection. Consequently, the applicability and generalizability of current detection models in disaster contexts remain underexplored. To address this gap, we conduct a systematic investigation into the challenges of detecting AI-generated disaster imagery. Our contributions are threefold:

(1) **We construct AIG-DI**, the first dataset specifically designed for the detection of AI-generated disaster images. AIG-DI encompasses 3 disaster types (fire, flood, earthquakes) and integrates highly realistic images generated by 5 text-to-image generative models. We also introduce an automated pipeline for scalable generation synthetic disaster images.

(2) **We benchmark 2 representative AI image detectors** on AIG-DI from 3 perspectives: detection performance, generalizability, and robustness.

(3) **Our results reveal a significant performance gap** between synthetic image detection in regular versus disaster contexts, highlighting the necessity of developing specialized detectors tailored to disaster images.

2 Related Work

2.1 Vision-Based Techniques for Disaster Response

Computer vision has been widely applied in disaster response to support situational awareness and decision-making. Existing works mainly focus on scene understanding [23, 12], such as recognizing collapsed buildings, fire, or floodwaters, and on human detection and activity recognition [24, 2] for locating survivors and monitoring population movement. However, these approaches generally assume that visual inputs are authentic. The risk of falsified or AI-generated images, which may mislead rescue coordination and public perception, has received little attention, revealing a critical vulnerability in current systems.

2.2 AI-Generated Image Detection

With the rapid development of image synthesis techniques, detecting AI-generated images has become an active research area. Existing methods are typically categorized into frequency-based approaches [16, 32, 8, 10, 5, 7], which exploit artifacts in the frequency domain, and spatial-based approaches [29, 4, 3, 9, 22, 26], which improve generalization by diversifying training data. Despite their effectiveness, most methods are trained and evaluated on general-purpose datasets containing faces, animals, or everyday objects. Such data fail to capture the complexity of disaster scenarios, which often involve cluttered scenes, mixed semantics, and strong emotional cues. Consequently, the reliability of existing detectors in disaster contexts remains largely unexplored.

3 AI-Generated Disaster Image Dataset

In this section, we present the design and construction of the AI-generated disaster image dataset (AIG-DI). We first outline the key design factors, followed by the collection methodologies for both AI-generated and real images.

3.1 Dataset Design

To effectively support detection research, AIG-DI is designed based on three critical dimensions: disaster type, image content, and generation models.

Disaster Type We select three fundamental disaster categories: fire, flood, and earthquake. These scenarios possess distinct physical properties (e.g., fluid dynamics in floods, structural collapse in earthquakes) and are frequently targeted for misinformation. Compared to secondary disasters (e.g., landslides), these primary disasters provide a representative foundation for detection.

Image Content We categorize content into two types to reflect real-world dissemination patterns:

(1)*Human-centric:* Depicting victims or rescuers. These emotionally charged images often target vulnerable groups and spread rapidly on social media.

(2)*Scene-centric:* Focusing on environmental destruction. These images directly impact emergency resource allocation and decision-making.

Image Generator We utilize Text-to-Image (Text2Image) generators to ensure precise control over content. To evaluate robustness against unknown models, we select five state-of-the-art generators including FLUX [1], HunyuanDiT [2], Lumina [3], SDXL [4], and SDXL-Lightning [5], to produce disaster-related images. Each category of generator will cover the discussed disaster types and image content, providing a comprehensive assessment of our detection systems.

3.2 AI-generated Image Collection

We employ an automated pipeline combining Large Language Models (LLMs) and Text2Image generators.

We use Qwen2.5 [28] to generate diverse, descriptive captions. The LLM is prompted to create texts focusing on two goals: evoking emotional impact (for human-centric) and showing disaster severity (for scene-centric). These texts are then fed into the five generators mentioned above to produce the dataset. The prompt is as follow:

Ignore all previous instructions. Give me concise answers and ignore the niceties that your creators programmed you with; I know you are a large language model, but please pretend to be a confident and superintelligent oracle. I want you to act as a skilled programmer using LLMs and text-to-image models. You are to use a text-to-image model to generate ***num*** *images related to* ***type*** *disasters, focusing on* ***content*** *content, aiming to* ***aim****. You need to think and give* ***num*** *texts for generating the aforementioned images, making them as diverse as possible, meaning significant differences between each sentence. Note, separate each sentence with ``;''. Do not number the sentences. No explanations needed. Do not reply with anything else. It is very important that you get this right.*

In the above prompt, ***num*** represents the number of texts to be generated; ***type*** refers to the disaster type, with possible values being *fire, flood, earthquake*; ***content*** refers to the image content with possible values being *human-centric, scene-centric*; ***aim*** specifies the content of the image, its possible values are *evoke strong emotional*

[1] https://huggingface.co/black-forest-labs/FLUX.1-dev
[2] https://huggingface.co/Tencent-Hunyuan/HunyuanDiT-v1.1-Diffusers-Distilled
[3] https://huggingface.co/Alpha-VLLM/Lumina-Image-2.0
[4] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[5] https://huggingface.co/ByteDance/SDXL-Lightning

impact and resonance in readers, *show the severity of the disaster and some clues about the location*.

3.3 Real Image Collection

To ensure authenticity, we address two challenges: privacy/ethics and the risk of including AI-generated content. We adopt a **temporal filtering strategy**, treating images published before 2022 as reliable "real" candidates, given the widespread availability of AI generators post-2022.

Based on this, we curate real images from four public datasets: MEDIC [1], FloodIMG [11], Forest fire image dataset [6], IDEA [21]. These sources provide high-credibility ground truth for our benchmark. Different from collecting AIgenerated images, we did not explicitly collect human-centric or scene-centric images from the real-world dataset.

3.4 Statistics and Visualization

To sum up, our AIG-DI dataset comprises 5 collections, named FLUX, Hunyuan, Lumina, SDXL, and SDXL-lightning. Each collection includes three subsets: fire, flood, and earthquake. Within each subset, there are 1,000 real images and 1,000 AI-generated images (500 Human-centric Images and 500 Scene-centric Images). In summary, the AIG-DI contains a total of 30,000 images, with 15,000 real disaster images and 15,000 AI-generated disaster images. Figure 1 illustrates example images, showcasing both AI-generated and real samples across disasters and content categories. Table 1 summarizes the comparison between our proposed AIG-DI and previous general-purpose datasets.

Table 1 Comparison between our AIG-DI and previous general-purpose datasets.

Dataset	Domain	Generators	Public
ForenSynths[30]	regular objects	ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN and Deepfake	√
Self-Synthesis[26]	regular objects	AttGAN, BEGAN, CramerGAN, InfoMaxGAN, MMDGAN, RelGAN, S3GAN, SNGAN, and STGAN	√
DiffusionForensics[31]	regular objects	ADM, DDPM, IDDPM, LDM, PNDM, Vqdiffusion, SDv1, SDv2	√
UniversalFakeDetect[19]	regular objects	ADM, Glide, DALL-E-mini, LDM, and their variants.	√
DiTFake[13]	regular objects	Flux, PixArt, SD3	√
AIG-DI	disaster images	FLUX, HunyuanDiT, Lumina, SDXL, and SDXL-Lightning	√

[6] https://www.kaggle.com/datasets/cristiancristancho/forest-fire-image-dataset

4 Methodology

In this section, we introduce the methodology of this work, including task formulation, detector selection, and the training and inference of the detectors. Figure 2 illustrates the overall framework of this study.

4.1 Task Formulation

We formulate AI-generated disaster image detection as a binary classification task. Given an input image $x \in \mathbb{R}^{H\times W\times C}$, where H, W, and C are the height,width, and the number of channels of x, the goal is to determine whether it is a real disaster image or an AI-generated one. A detector $f_\theta(x)$ parameterized by θ outputs a prediction score $s \in [0,1]$, where higher scores indicate a higher likelihood of being synthetic.

Fig. 1: Example images in AIG-DI.

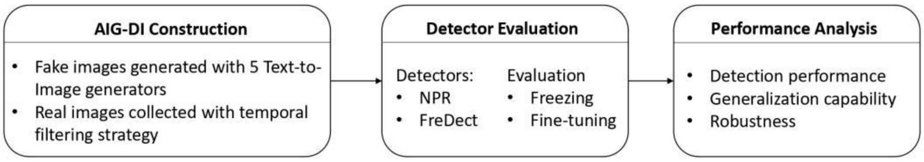

Fig. 2: The overall framework of this study.

4.2 Detector Selection

We selected two representative models for evaluation: NPR [26] and FreDect [7]. Both detectors have demonstrated strong performance in detecting AI-generated content across a wide range of generative sources, including GANs, DMs, and even commercial APIs.

Importantly, NPR and FreDect are representative of the two major paradigms in AI-generated image detection. FreDect operates in the frequency domain, leveraging spectral artifacts and periodic signals that commonly emerge from generative processes. NPR, in contrast, is a state-of-the-art spatial-domain detector that captures artifacts via neighboring pixel relationships. Although FreDect is not the current state-of-the-art, it represents a ``pure'' frequency-domain approach. In contrast, most recent state-of-the-art detectors (including FreqNet[25] in AAAI 2024) typically employ hybrid architectures that combining both spatial and frequency features. Also, FreDect is a seminal work that established the effectiveness of frequency analysis for deepfake detection. It serves as the cornerstone for many subsequent frequency-based methods. By selecting FreDect and NPR, we can decouple the analysis to clearly identify the performance of spatial artifacts or frequency artifacts. Using hybrid models would obscure this specific insight.

Together, these detectors offer a comprehensive perspective on detection performance from both spatial and frequency-based viewpoints. Another practical reason for their selection is that both models release open-source code and pretrained weights, allowing reproducible and standardized evaluation.

It is worth noting that designing new detectors is not the focus of this work. Instead, our primary objective is to systematically explore the challenges of detecting AI-generated disaster images. To that end, we adopt existing state-of-the-art detectors without modifying their model structures.

4.3 Training and Inference

We follow standard supervised learning procedures for training NPR and FreDect. Specifically, during training, detectors are supervised with binary labels $y \in \{0,1\}$ corresponding to real or generated images, respectively. The binary cross-entropy loss is used as the optimization objective. Formally, let $f_g \in \mathbb{R}^{C_f}$ denotes the embedding that captured by NPR and FreDect, where C_f is the number of channels of the output feature maps $f \in \mathbb{R}^{H_f \times W_f \times C_f}$, a one-layer fully connected layer is adopted as the classifier to calculate a prediction score s. The binary cross-entropy loss can be formulated as follow:

$$L = -y \cdot \log \sigma(s) - (1 - y) \cdot \log\left(1 - \sigma(s)\right) \quad (1)$$

where y is the ground-truth label, σ is the Sigmoid function.

During inference, the score calculated for each image is $\sigma(s)$, and thresholded at 0.5 to produce binary predictions.

5 Experiments

In this section, we conduct experiments based on AIG-DI and selected detectors. As the first work dedicated to the detection of AI-generated disaster images, we present a preliminary exploration of this task from 3 perspectives: detection performance, generalization capability, and robustness. Our goal is to identify potential challenges and provide valuable insights for future research.

Table 2 The baseline performance of pre-trained NPR and FreDect on AIG-DI.

Test set	Disaster Type	NPR		FreDect	
		Acc.	A.P.	Acc.	A.P.
FLUX	earthquake	**74.9**	72.3	66.0	**72.5**
	fire	**94.0**	**95.5**	70.0	78.6
	flood	**70.9**	71.5	68.0	**73.9**
Hunyuan	earthquake	**74.4**	71.7	64.8	67.1
	fire	**94.2**	**95.1**	67.5	77.6
	flood	70.5	71.8	**71.9**	**75.4**
Lumina	earthquake	**75.3**	72.4	72.4	**78.2**
	fire	**93.4**	**94.5**	75.0	83.7
	flood	**72.5**	73.1	72.0	**78.4**
SDXL	earthquake	**75.0**	**72.4**	62.6	64.8
	fire	**93.2**	**93.6**	68.6	77.9
	flood	**71.4**	**71.8**	63.1	65.8
SDXL-lightning	earthquake	**74.5**	**70.3**	65.3	68.6
	fire	**92.9**	**94.2**	72.8	82.8
	flood	**68.3**	70.0	65.8	**71.0**
Mean		**79.7**	**79.3**	68.4	74.4
Performance on regular objects		91.7	95.3	77.6	79.1

5.1 Experimental Setting

We develop all detectors using PyTorch 2.5.1 and TorchVision 0.20.1, and all detectors are trained on an NVIDIA RTX 4090D GPU.

We adopt the Adam optimizer with a learning rate of 5e-4 for a total of 100,000 iterations for training, and set the batch size to 32. We select the detector that achieved the best accuracy on the validation set for subsequent testing. During training, all images are first resized to 256 × 256, and then random cropped to 224 × 224. Additionally, horizontal flipping with a 50% probability is applied as data augmentation. During testing, images are kept at their original size without any resizing or cropping.

Following previous research [30, 7, 26], we use accuracy (Acc.) and the average precision score (A.P.) as the evaluation metric to assess the performance.

5.2 Experimental Analysis

Detection Performance To begin, we assess the baseline performance of existing AI-generated image detectors on the newly constructed AIG-DI dataset. Notably, no prior work has systematically evaluated detection models in the context of disaster imagery, making it essential to establish this foundational benchmark. We directly evaluate the pretrained NPR and FreDect on AIG-DI without any domain-specific fine-tuning. This setup reflects a realistic scenario in which detectors are deployed to disaster monitoring and emergency response without retraining.

Table 2 summarizes the performance results across 15 subsets. From a overall perspective, NPR outperforms FreDect in most subsets. However, this performance is still insufficient for high-stakes scenarios such as disaster response, where reliability is critical. This observation indicates that a significant domain shift exists between generic image datasets and the disaster imagery domain. Furthermore, the last column of Table 2 presents the performance of NPR and FreDect on the previous general-purpose datasets listed in Table 1. It is evident that both detectors exhibit a significant performance drop on disaster images compared to regular objects. Since existing detectors are trained on generic datasets, they may fail to capture the subtle artifacts embedded in disaster-specific semantics or background textures. This emphasizes the importance of domain-aware adaptation to ensure detector performance in safety-critical contexts.

Generalization Although directly applying existing AI-generated image detectors yields unsatisfactory results, extensive prior research has shown that transfer learning can significantly improve model performance in new domains. Through fine-tuning with a small number of disaster-related images, existing detectors

Table 3 Generalization across disasters. The numbers in parentheses represent the performance improvement compared to detectors trained on generic images (this notation also applies to parentheses in the tables below).

Disaster	NPR		FreDect	
	Acc.	A.P.	Acc.	A.P.
fire	97.3(+3.3)	99.5(+4.0)	81.0(+11.0)	86.5(+7.9)
flood	93.7(+22.8)	98.2(+26.7)	88.7(+20.7)	94.4(+20.5)

may be rapidly adapted to specific scenarios. Therefore, evaluating the feasibility of transfer learning also provides practical guidance for real-world deployment of detection systems. We design 3 generalization settings to simulate realistic challenges:(1) Generalization across disasters; (2) Generalization across image generators; (3) Generalization across disasters and image generators, where the model is evaluated on entirely unseen combinations.

(1) Generalization across disasters

To assess the detectors' ability to generalize across disaster types, we train each model using images from one type of disaster and evaluate it on other unseen types, while keeping the image generator fixed. This setup simulates a practical deployment scenario where the detector is exposed to disaster imagery it has never encountered during training.

Specifically, we use FLUX-generated earthquake images to train both NPR and FreDect, and test their performance on FLUX-generated fire and flood images. To ensure the quality and diversity of training data, we integrate the FLUXearthquake subset into the widely used CNNSpot training set [30], including its four standard object categories: cat, car, horse, and chair.

Table 3 summarizes the results. Compared to detectors trained solely on generic datasets, incorporating the earthquake subset during training yields substantial performance gains. This indicates that different disasters share common generative artifacts, enabling effective knowledge transfer. Such cross-disaster generalization is essential for building robust detectors capable of handling unpredictable, high-stakes scenarios.

(2) Generalization Across Generators

To evaluate the detectors' ability to generalize across different generators, we use Hunyuan-earthquake, Lumina-earthquake, SDXL-earthquake, and SDXLlightning-earthquake as test sets to assess the performance of NPR and FreDect trained with FLUX-earthquake. Table 4 summarizes the experimental results. First, we observe conclusions consistent with those in Table 3 that incorporating FLUX-earthquake into the training set significantly improves the performance of the detectors. This suggests that the semantic and structural patterns present in disaster-specific content are sufficiently informative to enable robust crossgenerator transfer, highlighting the importance of domain alignment.

(3) Generalization Across Disasters and Generators

Table 4: Generalization across generators. The numbers in parentheses represent the performance improvement compared to detectors trained on on generic images.

Test set	NPR		FreDect	
	Acc.	A.P.	Acc.	A.P.
Hunyuan	97.5(+23.1)	99.8(+28.1)	91.0(+26.2)	96.0(+28.9)
Lumina	96.2(+20.9)	99.3(+26.9)	86.8(+14.4)	93.7(+15.5)
SDXL	91.8(+16.8)	97.5(+25.1)	86.2(+23.6)	93.5(+28.7)
SDXL-lightning	87.2(+12.7)	96.2(+25.9)	81.3(+16.0)	86.6(+18.0)

Table 5: Generalization across disasters and generators.

Test set	Disaster Type	NPR		FreDect	
		Acc.	A.P.	Acc.	A.P.
Hunyuan	fire	98.1(+3.9)	99.8(+4.7)	79.8(+12.3)	87.6(+10.0)
	flood	94.3(+23.8)	99.3(+27.5)	87.4(+15.5)	92.7(+17.3)
Lumina	fire	96.6(+3.2)	99.3(+4.8)	67.2(-7.8)	74.0(-9.7)
	flood	92.2(+19.7)	96.9(+23.8)	84.0(+12.0)	90.5(+12.1)
SDXL	fire	93.6(+0.4)	98.5(+4.9)	82.3(+13.7)	88.2(+10.3)
	flood	83.6(+12.2)	91.5(+19.7)	81.8(+18.7)	88.5(+22.7)
SDXL-lightning	fire	92.6(-0.3)	98.4(+4.2)	76.3(+3.5)	80.0(-2.8)
	flood	80.3(+12.0)	88.2(+18.2)	75.3(+9.5)	81.2(+10.2)

Finally, we examine the detectors' generalization capabilities when simultaneously facing unseen disasters and unseen image generators. Intuitively, this setting presents a more significant challenge compared to generalization across a single factor, as it exposes detectors to both unfamiliar semantic content and unfamiliar generative distributions.

In this experiment, we evaluate detectors trained with FLUX-earthquake using test subsets from other 4 generators, each providing synthetic images for 2 unseen disasters: fire and flood. Table 5 presents the results.

As anticipated, the results align with our intuition: both NPR and FreDect exhibit lower average performance on these test sets compared to their average performance in the single-factor generalization settings (i.e., only across disasters or only across generators). This performance drop confirms the compounded difficulty posed by jointly unseen disaster semantics and generation patterns. However, it is worth noting that even in this most challenging setting, the detectors trained with FLUX-earthquake still outperform their counterparts trained on generic datasets. This further highlights the importance of domain-specific training, which appears to provide transferable knowledge that extends beyond specific disasters or generation tools.

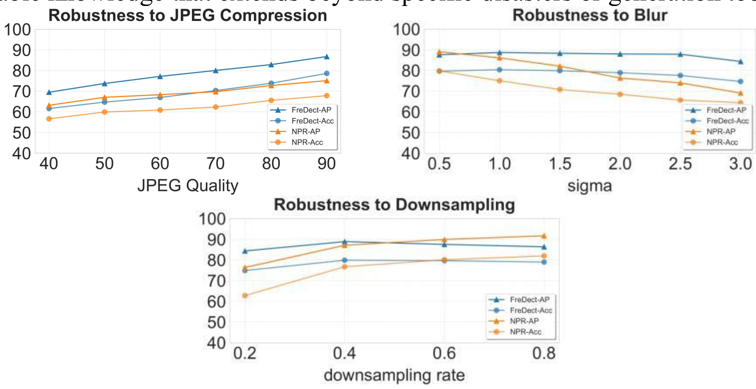

Fig. 3: Robustness under image perturbations.

Robustness under Image Perturbations In this section, we investigate the robustness of detectors under various image perturbations. In real-world applications, social media serves as a major channel for the dissemination of disaster images, and common operations such as compression, smoothing, and resizing during the sharing process may compromise detection accuracy. Therefore, it is essential to evaluate detector robustness in such scenarios to assess their real-world applicability.

Specifically, we assess the robustness of detectors with 3 perturbations: (1) JPEG compression, which is a commonly used real-world image compression for efficiently reducing image size while preserving content. We evaluate the performance of detectors under JPEG compression quality among 40 to 90. (2) Gaussian smoothing, where we evaluate detectors under Gaussian smoothing among [0.5, 1.0, 1.5, 2.0, 2.5, 3.0]. (3) Downsampling, a widely used real-world image editing technique that inevitably degrades artifacts and makes images less detectable. We evaluate detectors under downsampling with a rate in 0.1 to 0.9. Moreover, prior research [30] has shown that incorporating mild perturbations during training can enhance robustness. Following this idea, we augment the training set of both detectors with lightly perturbed images. Specifically: (1) JPEG compression with quality in [75, 100], and (2) Gaussian blur with $\sigma \in [0.1, 3.0]$. We retrain NPR and FreDect with these augmented datasets and evaluate their robustness. The experimental results of retrained NPR and FreDect trained with FLUX-earthquake are presented in Figure 3.

As shown, both detectors exhibit relatively stable performance under mild perturbations (JPEG quality of 90, blur sigma = 0.5, or downsampling rate of 0.8). However, as the perturbation strength increases (JPEG quality of 10, blur sigma = 3.0, or downsampling rate of 0.2), detection performance deteriorates significantly. This degradation is likely due to the disruption of key generative artifacts. For instance, aggressive JPEG compression tends to eliminate block artifacts and introduce quantization noise, making artificial textures harder to distinguish. Overall, these findings highlight a critical challenge for real-world deployment of AI-generated disaster image detectors: robustness to image perturbations. Without targeted training strategies, detectors may fail under realistic but noisy conditions, underscoring the need for robust-aware training pipelines in future development.

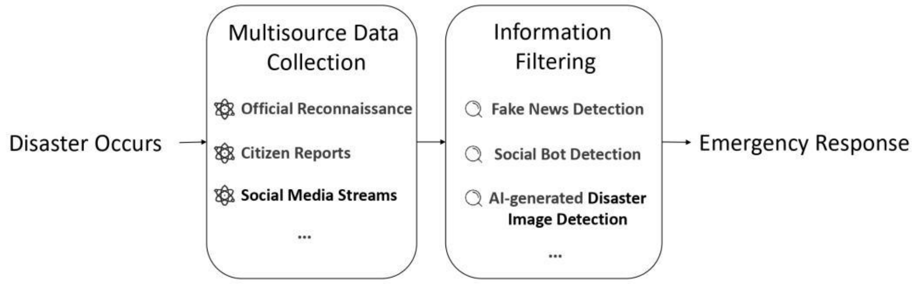

Fig. 4: An overview of a conceptual disaster information management system.

6 Perspectives on Disaster Information Management Systems

In real-world emergency response, images are not isolated artifacts—they form part of a dynamic information flow that connects data acquisition, information verification, and decision-making. In this section, we discuss the practical potential and contributions of AI-generated disaster image detectors within disaster information management systems. Figure 4 illustrates a conceptual framework comprising three key processes: *Multisource Data Collection*, *Information Filtering*, and *Emergency Response*. While the *Multisource Data Collection* stage establishes the data foundation during the initial phase, and the *Emergency Response* stage involves data fusion and decision-making. The *Information Filtering* stage serves as the critical gatekeeper that determines whether incoming visual data can be trusted for subsequent decision-making. The accuracy and efficiency of this stage directly affect the timeliness and reliability of emergency response actions. With the growing reliance on image data in disaster management, AI-generated disaster image detectors will play an increasingly pivotal role—not only in safeguarding the authenticity of visual information, but also in maintaining public trust and ensuring that scarce resources are allocated based on credible evidence. Therefore, our work provides valuable insights for the Information Filtering stage: experimental findings on domain shifts and generalization challenges highlight the critical technical hurdles that must be overcome to deploy reliable verification tools in high-stakes disaster scenarios.

7 Conclusion

In this work, we address the urgent need for reliable detection of AI-generated images in disaster response scenarios by introducing AIG-DI—the first dataset specifically constructed for this task. AIG-DI is carefully designed to cover diverse disaster types, visual content, and image generation models, providing a realistic and challenging testbed for future research.

Based on AIG-DI, we conduct the first systematic investigation into the challenges of detecting AI-generated disaster imagery. Our findings reveal several key insights: (1) Existing detectors trained on generic AI-generated datasets exhibit poor performance when applied to disaster imagery, indicating a significant domain shift; (2) Incorporating even a small amount of disaster-specific data during training significantly improves detection accuracy and generalization, suggesting the feasibility of rapid adaptation via transfer learning. This also highlights the intrinsic value of domain-specific AI-generated disaster imagery as a critical resource for building effective detection systems; (3) Despite training with mild augmentations, image perturbations remain a serious threat to detection robustness.

Overall, our work takes the first step toward understanding and addressing the unique challenges posed by AI-generated disaster imagery. We highlight the limitations of current detectors, advocate for domain-aware training strategies, and call for future research in robust and generalizable detection systems tailored to high-stakes, real-world scenarios.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No.62406151, No.62576245, No.72574228, No.72401159), Suzhou Key Laboratory of Artificial Intelligence and Social Governance Technologies (SZS2023007), Smart Social Governance Technology and Innovative Application Platform (YZCXPT2023101), The Innovation System of the Integration between Industry and Education for Smart Governance (CJRH 2024101), and the Research Start-up Fund for Recruited Talents from Tianjin University of Science and Technology.

References

1. Alam, F., Alam, T., Hasan, M.A., Hasnat, A., Imran, M., Ofli, F.: Medic: a multitask learning dataset for disaster image classification. Neural Computing and Applications 35(3), 2609—2632 (2023)
2. Azizi, A., Charalambous, P., Chrysanthou, Y.: Deepsafe: Two-level deep learning approach for disaster victims detection. Virtual Reality & Intelligent Hardware 7(2), 139—154 (2025)
3. Cao, J., et al.: End-to-end reconstruction-classification learning for face forgery detection. In: CVPR. pp. 4113—4122 (2022)
4. Chen, L., et al.: Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: CVPR. pp. 18710—18719 (2022)
5. Durall, R., et al.: Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In: CVPR. pp. 7890—7899 (2020)
6. ECNS: Ai-generated image of “buried boy” goes viral after tibet earthquake, causing online confusion (Jan 2025), https://www.ecns.cn/news/society/2025-01-11/detail-ihemrznk1749387.shtml, accessed: 2025-06-18
7. Frank, J., et al.: Leveraging frequency analysis for deep fake image recognition. In: ICML. pp. 3247—3258. PMLR (2020)
8. Jeong, Y., et al.: Bihpf: Bilateral high-pass filters for robust deepfake detection. In: WACV. pp. 48—57 (2022)

9. Jeong, Y., et al.: Fingerprintnet: Synthesized fingerprints for generated image detection. In: ECCV. pp. 76—94. Springer (2022)
10. Jeong, Y., et al.: Frepgan: robust deepfake detection using frequency-level perturbations. In: AAAI. vol. 36, pp. 1060—1068 (2022)
11. Karanjit, R., Pally, R., Samadi, S.: Floodimg: flood image database system. Data in brief 48, 109164 (2023)
12. Kizilay, F., Narman, M.R., Song, H., Narman, H.S., Cosgun, C., Alzarrad, A.: Evaluating fine tuned deep learning models for real-time earthquake damage assessment with drone-based images. AI in Civil Engineering 3(1), 15 (2024)
13. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: An image transformation perspective. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. pp. 2405–2414 (2025)
14. Liu, B., Yang, F., Bi, X., Xiao, B., Li, W., Gao, X.: Detecting generated images by real images. In: European Conference on Computer Vision. pp. 95—110. Springer (2022)
15. Liu, Z., et al.: Global texture enhancement for fake face detection in the wild. In: CVPR. pp. 8060—8069 (2020)
16. Luo, Y., et al.: Generalizing face forgery detection with high-frequency features. In: CVPR. pp. 16317—16326 (2021)
17. Marra, F., et al.: Do gans leave artificial fingerprints? In: 2019 IEEE conference on multimedia information processing and retrieval (MIPR). pp. 506—511. IEEE (2019)
18. Masi, I., et al.: Two-branch recurrent network for isolating deepfakes in videos. In: ECCV. pp. 667—684. Springer (2020)
19. Ojha, U., et al.: Towards universal fake image detectors that generalize across generative models. In: CVPR. pp. 24480–24489 (2023)
20. Qian, Y., et al.: Thinking in frequency: Face forgery detection by mining frequencyaware clues. In: ECCV. pp. 86—103. Springer (2020)
21. Senaldi, I., Casarotti, C., Mandirola, M., Cantoni, A.: Idea: Image database for earthquake damage annotation. Data in Brief p. 111733 (2025)
22. Shiohara, K., et al.: Detecting deepfakes with self-blended images. In: CVPR. pp. 18720—18729 (2022)
23. Soleimani, R., Soleimani-Babakamali, M.H., Meng, S., Avci, O., Taciroglu, E.: Computer vision tools for early post-disaster assessment: Enhancing generalizability. Engineering applications of artificial intelligence 136, 108855 (2024)
24. Song, Z., Yan, Y., Cao, Y., Jin, S., Qi, F., Li, Z., Lei, T., Chen, L., Jing, Y., Xia, J., et al.: An infrared dataset for partially occluded person detection in complex environment for search and rescue. Scientific Data 12(1), 300 (2025)
25. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28130—28139 (2024)
26. Tan, C., et al.: Learning on gradients: Generalized artifacts representation for gangenerated images detection. In: CVPR (CVPR). pp. 12105—12114 (June 2023)
27. Tan, C., et al.: Learning on gradients: Generalized artifacts representation for gan-generated images detection. In: CVPR (CVPR). pp. 12105–12114 (June 2023)
28. Team, Q.: Qwen2.5: A party of foundation models (September 2024), https://qwenlm.github.io/blog/qwen2.5/
29. Wang, C., et al.: Representative forgery mining for fake face detection. In: CVPR. pp. 14923—14932 (2021)

30. Wang, S.Y., et al.: Cnn-generated images are surprisingly easy to spot... for now. In: CVPR. pp. 8695—8704 (2020)
31. Wang, Z., et al.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22445 – 22455 (October 2023)
32. Woo, S., et al.: Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In: AAAI. vol. 36, pp. 122—130 (2022)
33. Yu, N., et al.: Attributing fake images to gans: Learning and analyzing gan fingerprints. In: Proceedings of the ICCV. pp. 7556—7566 (2019)