Diversity Matters: Dataset Diversification and Dual-Branch Network for Generalized AI-Generated Image Detection

Supplementary Material

This supplemental material provides insights about qualitative analysis of failure cases for AIGI detection.

6. Qualitative Analysis of the Proposed Method

Figure 5 presents a qualitative analysis of representative failure cases of the proposed AIGI detection framework across multiple datasets. Each sample is annotated with a red bounding box indicating the predicted probability of being fake. A decision threshold of 0.5 is used, where predictions above this threshold are classified as fake and those below as real. The samples are grouped as follows: (a), (c), (e), and (g) correspond to real images, while (b), (d), (f), and (h) represent fake images.

In some examples, the model's predictions are influenced by the inherent visual characteristics of the input samples. For example, sample (a), a real image from the MGD-Deepfake [20, 49] dataset, is assigned a high probability of 0.93. This image exhibits noticeable blur and low-frequency noise patterns, which are often associated with synthetic content, leading to its misclassification. Similarly, sample (c), a real image from the GANDF-AttGAN [35] dataset, receives a probability of 0.74. Its texture degradation and subtle compression artifacts further contribute to its resemblance to generated imagery, causing the model to classify it as fake. These cases highlight how challenging real-world distortions can resemble generative artifacts.

Conversely, certain fake samples demonstrate highly realistic visual quality, making them difficult to distinguish from authentic images. For instance, sample (b), a fake image, is assigned a low probability of 0.26, and sample (d) receives an even lower score of 0.09. These images exhibit strong structural consistency, sharp textures, and realistic lighting, closely matching natural image statistics. As a result, they are misclassified as real, reflecting the increasing realism of modern generative models.

Additional observations can be made for samples (e) and (g), both real images from different datasets (GENI-DALLE2 [54] and MGD-Guided [20, 49], respectively). These samples are assigned high probabilities of 0.93 and 0.87, respectively. Both images contain variations such as uneven illumination, compression noise, or mild distribution shifts compared to the training data, which makes them visually closer to synthetic content. Such cross-domain variations can influence the perceived authenticity of the samples.

Finally, sample (f), a fake image from GENI-DALLE2 [54], receives a probability of 0.45, which lies near the decision boundary, while sample (h), a fake image from MNW-HyperSD [27], is assigned a probability of 0.28. These samples exhibit a high degree of realism with minimal visible artifacts, making them difficult even for human observers to distinguish. The model's near-threshold and low-confidence predictions reflect the subtle nature of these generative outputs.

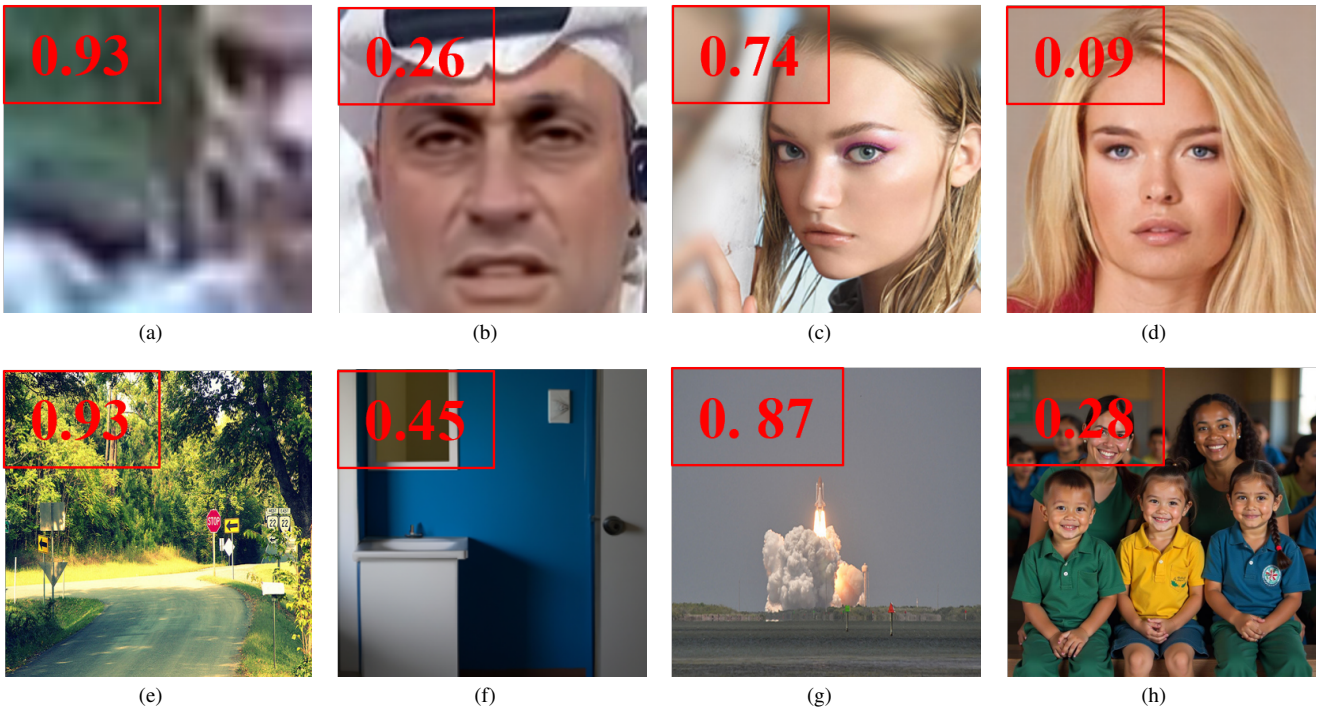

Figure 5. Qualitative analysis of AIGI detection across multiple datasets using the proposed method. (a) Real and (b) fake samples from MGD-Deepfake [20, 49], (c) real and (d) fake samples from GANDF-AttGAN [35], (e) real and (f) fake samples from GENI-DALLE 2 [54], and (g) real samples from MGD-Guided [20, 49] and (h) fake samples from MNW-HyperSD [27].