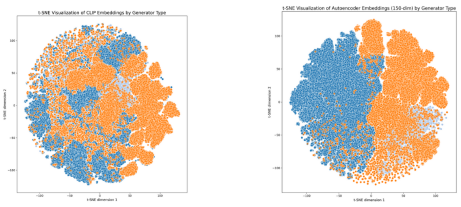
Reduced-Dimensional Anomaly-Aware Generalization for Deepfake Image Detection

The rapid advancement of generative models producing photorealistic imagery highlights the need for detectors that generalize to unseen synthesis techniques. Existing approaches including CNN classifiers, CLIP-derived embeddings, and local artifact analysis are typically trained as binary real/fake classifiers. Consequently, they tend to overfit to generator-specific artifacts present in the training data. When exposed to fakes from unseen generators lacking these artifacts, such detectors often misclassify them as real, resulting in poor robustness under distributional shift.



(a) Before (b) After

Figure 1: t-SNE visualization of real (blue) and fake (orange) embeddings before (left) and after (right) dimensionality reduction.

We propose a modular framework combining feature compression, student–teacher adversarial learning, and an anomaly-aware objective. We pose the problem as how far are fake images from the real ones. The pipeline extracts 768-dimensional embeddings from images using a pre-trained CLIPViT-L14 model, which are then compressed to 150 dimensions via a lightweight autoencoder that preserves discriminative structure while suppressing generator-specific noise. A transformer-based teacher-student network is trained on these compressed features: the student mimics the teacher on real images and diverges with a margin on fake images. The anomaly-aware component enforces tight clustering of real images in the latent space, enhancing generalization beyond training-specific artifacts. Additionally, we incorporate a generalized feature augmentation strategy using a feature augmenter G, which operates on extracted features to generate augmented fake features, promoting better generalization to unseen generators.

The training employs three key loss functions in an adversarial manner: (1) Discrepancy Loss for Real Images enforces student-teacher similarity on real samples, (2) Discrepancy Loss for Fake Images promotes divergence between networks on fake samples with a margin constraint, and (3) Generalized Feature Augmentation Loss uses the feature augmenter G to improve cross-generator generalization by making networks robust to training-specific artifacts.

Results

Performance on Known Generators

Method	ProGAN	CycleGAN	BigGAN	StyleGAN	GauGAN	StarGAN	Deepfakes
Before DR	95.23	94.19	91.03	95.03	90.27	90.27	82.18
After DR	99.84	98.49	98.57	99.30	98.13	99.05	89.12

Performance on Unseen Generators

Method	SITD	SAN	CRN	IMLE	Glide-27	Glide-50	Glide-100
Before DR	65.83	68.34	53.11	54.22	90.43	90.98	91.27
After DR	71.67	70.78	62.58	73.53	99.67	99.51	99.18

The framework achieves strong results across both seen and unseen generators.

Contributions

1. Problem Identification: We identify the tendency of binary classifiers to overfit to generator-specific artifacts, limiting cross-generator robustness.

2. Dimensionality Reduction Approach: We propose dimensionality reduction as a principled means of enforcing compact, anomaly-aware latent spaces where real images form a tight cluster.

References

Shao, R., Pang, T., Cao, J., Lin, M., Du, C., & Yan, S. (2023). GenDet: Towards Good Generalizations for AI-Generated Image Detection. arXiv preprint arXiv:2312.08880.