REFERENCE-GUIDED WHITENING AND CONTRAST FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in image generation have made synthetic imagery more diverse and realistic, increasing the risks of manipulation and misinformation, yet existing methods often overfit to generator-specific artifacts or shortcuts in their training data, performing well within the training generator but degrading sharply on unseen generators. In this paper, we propose CoRe (Content-and-Real Referencing), a reference-guided approach with two components. Real-Referenced Whitening (RRW) whitens features with a covariance estimated from real images only, discounting the high-variance directions of natural variation and emphasizing the low-variance directions where transferable evidence lies. Content-Referenced Fake Contrastive (CRFC) then operates on the whitened features, contrasting each fake against content-similar reals and content-different fakes, so content cannot act as a shortcut. We instantiate CoRe on the Perception Encoder (PE), a recent vision foundation model whose features exhibit particularly favorable anisotropy for this purpose. On the recent and challenging AIGIBench and Chameleon benchmarks, our method achieves state-of-the-art mean accuracy, exceeding the strongest baseline on AIGIBench by 5.7 points, Chameleon (SD v1.4) by 7.9 points.

1 INTRODUCTION

Generative adversarial networks and diffusion models have rapidly expanded the quality, diversity, and controllability of image generation (Goodfellow et al., 2014; Rombach et al., 2022), but the same capabilities enable identity forgeries, non-consensual content, misleading online narratives, and false visual evidence (Mirsky & Lee, 2021). As visual quality improves, human inspection alone is no longer reliable (Lu et al., 2023), and new generators appear faster than detectors can be retrained for them. These developments make robust *cross-generator generalization*, i.e., the ability to detect images from generators that were never observed during training, a central requirement for deployable AI-generated image detectors.

To address these issues, we propose **CoRe** (Content-and-Real Referencing), a reference-guided framework with two components. **Real-Referenced Whitening** (RRW) estimates a covariance from source real images alone and uses it to define a shared transformation for both classes: generated images supply the supervised displacement but not the reference metric, so class differences aligned with high-variance real directions are discounted, while those along stable, low-variance directions receive greater emphasis (Figure 1). During adaptation, binary cross-entropy updates the backbone in the whitened space while the real statistics are re-estimated, so whitening serves as the metric under which the representation is optimized rather than a fixed post-processing step. **Content-Referenced Fake Contrastive** (CRFC) targets the remaining shortcut: even in the whitened space, binary labels can separate the classes through content that correlates with the source. On the same whitened features, CRFC contrasts each fake against content-similar real images (negatives) and content-different fakes (positives), so the decision rests on synthesis cues shared across content rather than on content itself.

We instantiate **CoRe** on the Perception Encoder (PE), which provides strong general-purpose features without task-specific forensic pretraining (Bolya et al., 2025). Our analysis shows a strongly anisotropic source-real representation whose long low-variance subspace retains discriminative evidence that carries over to unseen generators. For PE and DINOv3 most of the real–fake displacement, measured in the whitened metric, falls in that subspace, whereas for CLIP it is split roughly

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

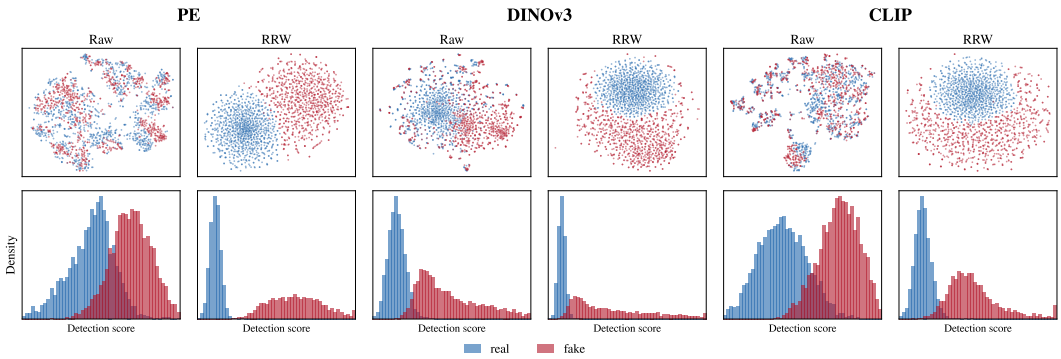


Figure 1: **Real-referenced whitening across three vision foundation models.** For PE, DINOv3, and CLIP, we show cls-token features before (Raw) and after real-referenced whitening (RRW). Top: t-SNE, bottom: detection-score distributions for real (blue) and generated (red) test images, in units of the real score's standard deviation. Whitening by the real-image covariance separates the two classes most for PE, whose geometry is the most anisotropic of the three.

evenly and the low-variance part sits near the level expected from estimation noise. This is what makes the whitened metric transferable rather than merely a better fit to the source: whitening frozen features multiplies the separability several-fold for PE and DINOv3, and the benefit follows the same ordering across the three backbones. Backbone adaptation under the whitened metric and CRFC each add further gains on PE, the latter on the more challenging benchmarks. **CoRe** achieves state-of-the-art mean accuracy on AIGIBench and Chameleon under the SD v1.4 protocol, while remaining competitive across the diverse generators in GenImage (Zhu et al., 2023; Li et al., 2025b; Yan et al., 2025a).

Our contributions are summarized as follows:

- We propose **RRW**, which fits the detection metric to real images alone and optimizes the backbone under that metric, and show that its benefit across backbones is predicted by where the real–fake displacement falls in the real-image spectrum.
- We introduce **CRFC**, a content-conditioned contrastive objective on the whitened features that draws negatives from content-similar real images and positives from content-different fakes, so that the real-fake decision rests on synthesis cues shared across content rather than on content itself.
- We evaluate **CoRe** on GenImage, AIGIBench, and Chameleon, obtaining state-of-the-art mean performance on AIGIBench and Chameleon and providing controlled analyses of backbone choice, covariance estimation, and robustness under distribution shift.

2 Related Work

Vision foundation models in generalizable AI-generated image detection. Broadly pretrained backbones anchor most recent detectors: frozen CLIP features with a lightweight readout (Radford et al., 2021; Ojha et al., 2023), category-common prompts (Tan et al., 2025), orthogonal-subspace adaptation (Yan et al., 2025b), simple probes on DINOv3 (Siméoni et al., 2025; Zhou et al., 2025), anchor-regularized DINOv3 adaptation (Choi et al., 2026), and calibrated textual prototypes on PE (Bolya et al., 2025; Cai et al., 2026). Each commits to a single backbone and treats its feature space as given; RRW instead asks when a backbone's geometry supports real-referenced whitening, and across CLIP, DINOv3, and PE its benefit tracks how the real–fake displacement projects onto the real-image spectrum, not the pretraining modality. Anchor-regularized adaptation also respects authentic-image geometry, but preserves it where RRW measures in it (Choi et al., 2026).

Covariance-aware discrimination. Fisher discrimination measures a class displacement relative to within-class scatter, yielding an inverse-covariance mean-difference direction (Fisher, 1936); shrinkage estimators stabilize the inverse in high dimensions (Ledoit & Wolf, 2004), Deep LDA and

differentiable closed-form solvers couple such classifiers to representation learning (Dorfer et al., 2016; Bertinetto et al., 2019), and Mahalanobis distances to a reference distribution underpin standard out-of-distribution detectors (Lee et al., 2018). RRW departs from these in its asymmetry: only source real images define the reference covariance, encoding the variation an authentic image is allowed to exhibit, while generated images supply the supervised displacement alone; this keeps generator-specific covariance out of the metric, and unlike post-hoc Mahalanobis scoring, the covariance shapes the space in which the detector is trained.

Content-conditioned contrastive learning for detection. Contrastive objectives shape a metric by pulling matching examples together and pushing mismatched ones apart (van den Oord et al., 2018; Chen et al., 2020; Khosla et al., 2020). In generated image detection, content often correlates with the source labels, letting a detector key on scene statistics rather than synthesis traces; prior work reduces this content bias by suppressing content components in the representation (Gye et al., 2025), and face-forgery detectors use contrastive or consistency signals to stress manipulation traces (Sun et al., 2022). CRFC instead controls the bias in the objective: unlike supervised contrastive learning, which pairs examples by label alone (Khosla et al., 2020), it draws negatives from content-similar real images and positives from content-different fakes. Content is thereby matched across the real–fake boundary: reducing the loss requires evidence separating the classes at fixed content, leaving shared synthesis traces as the only alignment across fakes.

3 METHOD

3.1 MOTIVATION

Detectors built on strong pretrained backbones still degrade sharply on unseen generators, and we attribute this not to the backbone's capacity but to how its feature space is used under binary supervision, through two causes (Figure 1). *Geometry:* the real–fake differences that remain valid for unseen generators lie along low-variance directions of the strongly anisotropic real distribution, which a standard linear classifier underuses because high-variance directions dominate its decision and training signal. *Supervision:* binary labels do not indicate which differences carry over, so source-correlated content acts as a shortcut. CoRe addresses each with a matching component, shown as the two branches of Figure 2: a real-referenced metric that rescales the space by the second-order statistics of real images (RRW, §3.3), and a content-referenced objective that contrasts each generated image against content-matched references (CRFC, §3.4).

3.2 SETUP AND NOTATION

Given an image, label pair (x, y) with $y \in \{0, 1\}$, where $y = 1$ denotes a generated image and $y = 0$ a real one, the backbone provides an image encoder f_θ,

$$z = f_\theta(x) \in \mathbb{R}^D, \tag{1}$$

with θ trainable. Training data consist of real images $\mathcal{R}$ and generated images $\mathcal{F}$ from the generators. All statistics used by CoRe are computed from these features. Because the dimensions of z differ in scale by orders of magnitude, a covariance estimated directly from z would be dominated by a few large dimensions and poorly conditioned, so we first standardize each dimension,

$$q(x) = A\big(z(x) - m\big), \qquad A = \mathrm{diag}(\sigma_1 + \epsilon, \ldots, \sigma_D + \epsilon)^{-1}, \tag{2}$$

where m and σ are the mean and per-dimension standard deviation over the source real images $\mathcal{R}$, and $\epsilon > 0$ caps the rescaling of dimensions whose variance is close to zero. In these coordinates we form the mean feature of each class, their displacement, and the covariance of the real class,

$$\mu_R = \frac{1}{N_R} \sum_{i \in \mathcal{R}} q_i, \qquad \mu_F = \frac{1}{N_F} \sum_{i \in \mathcal{F}} q_i, \qquad \delta = \mu_F - \mu_R, \tag{3}$$

$$\widehat{\Sigma}_R = \frac{1}{N_R - 1} \sum_{i \in \mathcal{R}} (q_i - \mu_R)(q_i - \mu_R)^\top. \tag{4}$$

Both the per-dimension standardization (m, σ) and the real covariance $\widehat{\Sigma}_R$ are estimated from the source real images $\mathcal{R}$ alone; generated features enter only through the class displacement δ (Eq. 3). Since m is the real-class mean, the standardized real mean satisfies $\mu_R = 0$, so the reference metric C_R is defined entirely by the natural variation of real images.

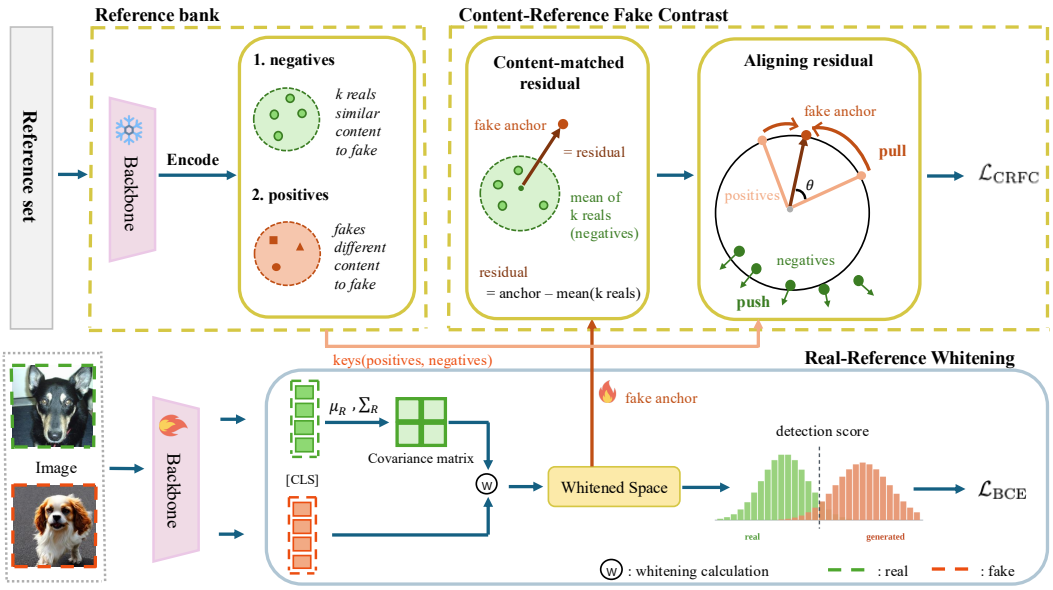


Figure 2: **Overview of CoRe.** Real features alone define a whitening transform Ⓦ applied to both classes, and the score in the whitened space is trained with $\mathcal{L}_{\mathrm{BCE}}$. Before training, the initial backbone encodes a set of reference images once and selects, for each generated image, its content-matched real neighbors and paired generated images of different content; the stored features are re-projected through the current transform at every refresh. During training, each generated image is centered on the mean of its content-matched reals, and the resulting residuals are aligned with $\mathcal{L}_{\mathrm{CRFC}}$. Only the whitening branch is used at test time.

3.3 Real-Referenced Whitening

A metric fitted on real images. The standard way to read a representation is to attach a linear head and train it with binary cross-entropy loss. Under an equal-covariance Gaussian model its optimum is Fisher's direction $\Sigma^{-1}(\mu_F - \mu_R)$, but this describes the minimizer of the expected loss, not what training on a finite set recovers. When the source classes are linearly separable, many directions attain zero training error, and the one gradient descent selects is determined by the sample rather than by what survives a change of generator. Content, resolution, or the fingerprint of the training generator separates the source as well as a synthesis trace does, and is easier to find, since such factors lie along directions in which the features vary most. The labels alone cannot say which of these directions to prefer; the distribution of real images can. Real-image variation does not change when the generator does, so we fix the direction in closed form and use that variation to choose the geometry in which it is measured. We therefore fix the direction in closed form and choose the geometry in which it is measured. Writing $C_R = \widehat{\Sigma}_R + \epsilon I = U\Lambda U^\top$ with a small ridge $\epsilon > 0$,

$$W = C_R^{-1/2} = U\Lambda^{-1/2}U^\top, \qquad h(x) = W\big(q(x) - \mu_R\big), \tag{5}$$

so the real class has identity covariance under W and a feature is expressed in units of authentic variation. We use the symmetric (ZCA) square root, which among all valid whitening maps is closest to the identity in Frobenius norm and so rotates the original dimensions as little as possible. The same W is applied to both classes; only the distribution on which it is *fitted* is real-only. Estimating the covariance on generated images, or on the pooled mixture, would admit the within-class structure of the training generator into the metric, and whitening divides by whatever serves as the reference: directions in which that generator varies would be discounted even though its traces may lie there.

Detection score. The score projects the transformed feature onto the transformed displacement,

$$s(x) = h(x)^\top W\delta = \big(q(x) - \mu_R\big)^\top C_R^{-1}\,\delta, \tag{6}$$

which weights the i-th eigendirection by λ_i^{-1}: directions along which real images vary widely are discounted, and those in which they are tightly concentrated are amplified. This is Fisher's direction

under a real-only within-class covariance, equivalently the matched filter for a known signal in colored Gaussian noise (Kay, 1998),

$$\max_{w \neq 0} \frac{(w^\top \delta)^2}{w^\top C_R w} = \delta^\top C_R^{-1} \delta, \qquad w \propto C_R^{-1} \delta. \tag{7}$$

The score fixes the decision *direction* but not the rule: the detector is the affine map $\ell(x) = \alpha\, s(x) + b$, with $\alpha > 0$ and b learned jointly with the backbone under $\mathcal{L}_{\mathrm{BCE}}$, and an image is called generated when $\ell(x) > 0$. Subtracting μ_R makes $\mathbb{E}_R[s] = 0$, so s is centered on the real class and b sets the operating threshold relative to it, in place of the fixed $\frac{1}{2}\delta^\top C_R^{-1}\delta$ that a hard Fisher rule would use. Two caveats bound Eq. 7. It assumes the true covariance, whereas with a finite reference set the factor λ_i^{-1} that amplifies weak evidence also amplifies the estimation error of small eigenvalues, which the ridge limits; and it concerns the source statistics, and does not assert that an unseen generator displaces the mean in the same direction.

3.4 Content-Referenced Fake Contrastive

Reference bank. The metric of §3.3 changes the geometry under which evidence is read, but not what the labels say: binary supervision still admits solutions that separate the classes through content that correlates with the source, since a generator's outputs differ from the real corpus in semantics as well as in synthesis traces. CRFC suppresses this shortcut through the way its pairs are formed, using a reference bank built once before training with the initial backbone. For each generated image i we retrieve the k real reference images whose features are nearest in cosine similarity, written $\mathcal{N}(i)$; as the cls-token features of the backbone are dominated by semantics, these are real images of similar content. We also pair i with generated reference images that are dissimilar to it in content, keeping those whose deviation from their own content-matched reals points in a direction similar to i's, so that a pair shares synthesis traces but not content; this set is written $\mathcal{P}_i$. The bank stores these indices together with the features of the reference images as encoded by the initial backbone.

Content-matched residuals and objective. During training, each generated image in the batch is encoded by the trainable backbone and serves as an anchor, with residual and direction

$$e_i = h_i - \frac{1}{k} \sum_{j \in \mathcal{N}(i)} \bar{h}_j, \qquad v_i = \frac{e_i}{\|e_i\|_2}, \tag{8}$$

where h_i is the anchor's transformed feature and $\bar{h}_j$ the stored feature of a content-matched real image. Subtracting the mean of content-matched real images cancels the component the two share, what the image depicts, and leaves the component they do not, how it was produced. This is the paired-comparison construction of classical statistics, with content playing the role of the matched pair. The normalization then keeps only the direction of the residual, the part we expect to be shared, and discards its magnitude, which varies with how strongly a particular image is marked. Each paired generated image $p \in \mathcal{P}_i$ is centered on its own content-matched reals in the same way, giving v_p; each real image $n \in \mathcal{N}(i)$ is centered on the mean of the remaining members of the neighborhood, giving v_n. With temperature τ, the term for anchor i is

$$\mathcal{L}_{\mathrm{CRFC}}^{(i)} = -\frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \frac{\exp\big(v_i^\top v_p/\tau\big)}{\sum_{p' \in \mathcal{P}_i} \exp\big(v_i^\top v_{p'}/\tau\big) + \sum_{n \in \mathcal{N}(i)} \exp\big(v_i^\top v_n/\tau\big)}, \tag{9}$$

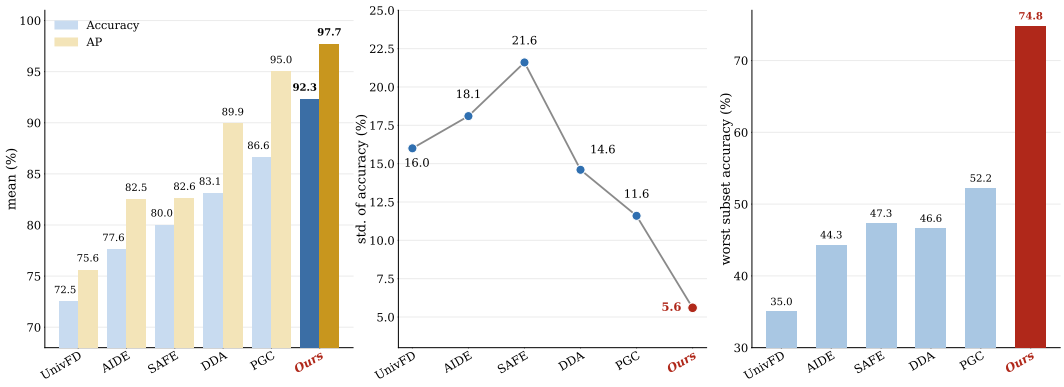
averaged over the generated images in the batch. Because a paired image depicts different content, the two residuals cannot be brought together through content, so increasing their similarity requires aligning a component that content does not explain; because a real neighbor depicts similar content, separating from it cannot be achieved by encoding what the image shows. The keys are detached, so the gradient reaches the backbone only through the anchor.

3.5 Training objective

The two components are combined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda\, \mathcal{L}_{\mathrm{CRFC}}. \tag{10}$$

where λ balances the two terms. The contrastive term acts only during training: it shapes the representation through the anchor. At test time an image is encoded, standardized, transformed by W, and scored by Eq. 6; the decision is $\ell(x) > 0$ with (α, β) fixed from the source.



(a) mean performance (b) variability across subsets (c) worst case

Figure 3: **Mean performance, variability, and worst case on AIGIBench.** (a) Mean accuracy and average precision over all AIGIBench 25 subsets. (b) Standard deviation of accuracy across the subsets. (c) Accuracy on the single hardest subset for each detector.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on three benchmarks that test generalization to unseen generators. GenImage (Zhu et al., 2023) is a million-scale benchmark spanning eight generators covering both GANs and diffusion models. AIGIBench (Li et al., 2025b) gathers a diverse set of generators under realistic distribution shifts; following its protocol, we train under *Setting-II*, in which the training set combines ProGAN and SD v1.4, and evaluate on all 25 of its test subsets. Chameleon (Yan et al., 2025a) is a test-only benchmark of highly photorealistic images that passed a human perception test: only generated images that human annotators judged to be real were retained, so every image is one that people fail to recognize as synthetic. It provides only a test set, so following its protocol we train two detectors, one on SD v1.4 and one on All GenImage, and report the results of each on the same Chameleon test images. AIGIBench and Chameleon are both recent benchmarks built to reflect the difficulty of current generators, on which existing detectors still perform poorly. We compare against detectors built on low-level artifacts (Wang et al., 2020; Frank et al., 2020; Tan et al., 2023; 2024a;b; Zhong et al., 2023), on reconstruction or diffusion inversion (Wang et al., 2023; Chen et al., 2024), on vision–language features (Ojha et al., 2023; Tan et al., 2025; Liu et al., 2024; Yan et al., 2025b), and on recent training or representation strategies (Yan et al., 2025a; Li et al., 2025a; Guillaro et al., 2025; Chen et al., 2025; Yang et al., 2026; Mu et al., 2026; Dhakal et al., 2026; Zhou et al., 2026), taking their reported numbers under each protocol.

Implementation details. We instantiate CoRe on two backbones: the Perception Encoder (ViT-L/14 at 336×336) (Bolya et al., 2025) and DINOv3 (ViT-L/16 at 256×256) (Siméoni et al., 2025), using the `[CLS]` representation of the final layer. Each image is preprocessed by resizing its shorter side to the backbone resolution and taking a center crop; we apply no data augmentation. The model is trained with the AdamW optimizer (Loshchilov & Hutter, 2019), a batch size of 64, and a one-cycle learning-rate schedule, for a single epoch. We fine-tune all transformer blocks with layer-wise learning-rate decay: the backbone peak learning rate is 5×10^{-7} with a per-layer decay factor of 0.65 for PE and 0.9 for DINOv3. For CRFC we set the contrastive weight to λ=0.1 and temperature to τ=0.1.

4.2 Main Results

Evaluation on AIGIBench. We evaluate on all 25 test subsets of AIGIBench under Setting-II, which span GAN, face-swap, diffusion, personalized, and in-the-wild sources. CoRe reaches a mean accuracy of 92.3% and a mean average precision of 97.7%, 5.7 and 2.7 points above the strongest prior method (Figure 3a), with the largest gains on the in-the-wild subsets where prior detectors degrade most; per-subset numbers are reported in Appendix A.2. The average, however, is the least

Table 1: Generalization on Chameleon. For each cell the upper value is accuracy and the lower pair is fake accuracy / real accuracy (%). Baseline numbers are taken from Mu et al. (2026) and Dhakal et al. (2026). Best in **bold**, second best underlined. A dash indicates that the corresponding training source is not reported by that method.

Training Dataset	UnivFD	DIRE	PatchCraft	NPR	AIDE	HiDA-Net	SimLBR	Ours (DINOv3)	Ours (PE)
SD v1.4	55.6 17.7/93.5	59.7 11.9/95.7	56.3 3.1/96.4	58.1 2.4/100.0	62.6 20.3/94.4	73.4 65.5/79.4	85.6 93.0/80.1	<u>89.5</u> 90.0/89.2	**93.5** 90.0/96.1
All GenImage	60.4 85.5/41.6	57.8 2.1/99.7	55.7 1.4/96.5	57.8 1.7/100.0	65.8 26.8/95.1	79.1 76.8/82.2	– –	<u>80.5</u> 96.3/68.6	**92.4** 89.5/94.6

informative of the three views. CoRe attains the lowest spread across subsets (Figure 3b), which means it detects every family reliably rather than excelling on some while failing on others. This uniformity is clearest in the worst case: CoRe's hardest subset still reaches 74.8% (Figure 3c), 22.6 points above the best worst-case accuracy of any prior method: a floor at which the detector remains usable, where every prior method drops below 53%, near or below chance, on at least one family. This behavior follows from the construction rather than from tuning: because the reference is the real class, which every subset shares, no generator family is privileged, and the detector raises the floor across generators rather than the average alone: the number that matters in deployment.

Evaluation on Chameleon. We also evaluate on Chameleon, the most difficult of the three benchmarks, where most recent detectors remain close to chance (Table 1). Chameleon exposes a characteristic failure: because every image passed a human "Turing test," many detectors collapse to calling nearly everything real, reaching high real accuracy with fake accuracy near zero. Under the SD v1.4 protocol CoRe sets a new state of the art with both backbones, reaching 93.5% with PE and 89.5% with DINOv3 and exceeding the strongest prior detector by 7.9 and 3.9 points, and it does so with balanced errors: 90.0% on fake and 96.1% on real images, where prior methods trade one class against the other. Under the All GenImage protocol CoRe again sets the state of the art, reaching 92.4% with PE, 13.3 points above the strongest prior detector. (We subsample 162K generated images uniformly across the eight generators, together with 162K real images, so that it matches the SD v1.4 protocol in training set size and differs from it only in the diversity of the training generators; the baselines in that row use the full union of the eight splits.)

Evaluation on GenImage. On the large and diverse GenImage, CoRe reaches 94.6% mean accuracy with PE and 93.6% with DINOv3, and is the best method on Midjourney (95.8%)and strong on the GAN source BigGAN (95.3%), on which several diffusion-oriented detectors drop sharply (Table 2). The two backbones fail in the same place: ADM is the weakest subset for each, at 74.9% for PE, which points to a property of that generator rather than of a particular representation. Its samples are produced by a pixel-space diffusion model with classifier guidance, and its mean displacement is the least aligned with the source direction among the eight generators—precisely the transfer condition that §3.3 identifies but does not control, so the failure occurs where the analysis predicts it should. Together with the AIGIBench and Chameleon results, this indicates that CoRe generalizes across both GAN and diffusion generators without specializing to any one benchmark.

4.3 Ablation Studies

We next separate the two components of CoRe. RRW alone already performs well with PE and DINOv3 across the three benchmarks. Adding CRFC improves the mean by 4.1points for PE and 5.6 for DINOv3, and the gain is distributed unevenly in an informative way, as shown in Table 3. It is largest on Chameleon($+8.6$ for PE, $+11.6$ for DINOv3), smaller on AIGIBench ($+3.6$ and $+1.8$), and absent on GenImage for PE, where the real and generated images share the same ImageNet content distribution and accuracy is already near saturation. The two backbones respond differently in magnitude but identically in order. DINOv3 starts lower under RRW alone and gains more from CRFC, consistent with its flatter real-image spectrum, which admits a weaker whitening metric and therefore leaves more content dependence for the contrastive term to remove. PE remains the stronger instantiation under both configurations (Appendix A.5 reports the sensitivity to λ).

Table 2: Cross-generator accuracy (%) on GenImage. All detectors are trained on SD v1.4. Best is in bold and second best is underlined.

Method	**Ref.**	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	**mAcc**
CNNDet	CVPR 2020	50.1	50.3	50.3	53.0	51.7	51.4	50.0	69.8	53.3
FreDect	ICML 2020	32.1	28.8	28.9	62.9	42.8	35.9	72.1	26.1	41.2
LGrad	CVPR 2023	73.7	76.3	77.4	51.8	49.8	73.1	52.1	40.5	61.8
UnivFD	CVPR 2023	93.9	96.4	96.2	71.9	85.4	94.3	81.6	90.5	88.8
PatchCraft	arXiv 2023	89.7	95.0	94.6	81.6	83.5	90.9	88.2	91.5	86.3
FreqNet	AAAI 2024	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.8
NPR	CVPR 2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FatFormer	CVPR 2024	92.7	**100.0**	**99.9**	75.9	88.0	**99.9**	**98.8**	55.8	88.9
AIDE	ICLR 2025	79.4	99.7	<u>99.8</u>	78.5	91.8	98.7	80.3	66.9	86.9
Effort	ICML 2025	82.4	<u>99.8</u>	<u>99.8</u>	78.7	93.3	97.4	91.7	77.6	90.1
SAFE	KDD 2025	<u>95.3</u>	99.4	99.3	82.1	96.3	98.2	96.3	97.8	95.6
C2P-CLIP	AAAI 2025	88.2	90.9	97.9	<u>96.4</u>	<u>99.0</u>	98.8	96.5	**98.7**	<u>95.8</u>
HiDA-Net	ICLR 2026	94.2	99.1	99.2	92.7	**99.1**	98.0	<u>96.9</u>	97.8	**97.1**
LTD	CVPR 2026	63.0	97.2	97.2	90.4	97.6	96.3	92.5	<u>98.6</u>	91.6
Simlbr	CVPR 2026	91.7	98.3	98.1	**97.0**	92.4	97.4	93.5	88.0	94.6
Ours(DINOv3)	–	91.1	99.0	98.7	77.2	95.8	98.9	96.3	91.6	93.6
Ours(PE)	–	**95.8**	99.2	99.2	77.1	96.0	<u>99.1</u>	95.1	95.3	94.6

Table 3: **Ablation of the two CoRe components.** For each backbone, RRW denotes training with the real-referenced whitening metric alone (§3.3), and +CRFC adds the content-referenced contrastive objective (§3.4) on top; the latter row is the full CoRe. Accuracy (%) is the mean over the test subsets of each benchmark, with the trained-on generators included for GenImage; parentheses give the change in the four-column mean from adding CRFC.

Backbone	Method	GenImage	Chameleon	AIGIBench	Mean
DINOv3	RRW	90.0	77.9	82.8	83.6
	<u>+CRFC</u>	**93.6**	**89.5**	**84.6**	**89.2** (+5.6)
PE	RRW	94.5	84.9	88.7	89.4
	<u>+CRFC</u>	**94.6**	**93.5**	**92.3**	**93.5** (+4.1)

Pair construction. Table 4 tests whether the content conditioning of §3.4 matters, on AIGIBench with PE. Replacing the content-different positives with random generated images costs 2.2 points: random pairs can be aligned partly through shared content, so the objective no longer isolates synthesis traces. Replacing the content-similar negatives with random reals costs 0.4: the anchor separates from a random real more easily through content, weakening the constraint, though the residual construction of Eq. 8 still removes most of it. The asymmetry is informative: the positives carry the conditioning, and the negatives refine it.

Table 4: Ablation on CRFC pair selection (AIGIBench, PE). Each row changes only how pairs are formed.

Pair construction	Acc.
Content-conditioned (ours)	**92.3**
Random generated positives	90.1 (−2.2)
Random real negatives	91.9 (−0.4)

Robustness. Images are rarely used as produced—they are recompressed, rescaled, and filtered en route through social platforms—so a detector must hold up under perturbations never seen in training. In Figure 4 we apply JPEG compression, blur, and resizing at test time only, on SD v1.4 and the eight-generator GenImage average. All methods lose accuracy as the perturbation strengthens, since these operations attenuate the high-frequency band where much generator evidence resides; but both instantiations of CoRe degrade far more slowly, stay highest at every blur level, and remain nearly flat under resizing, preserving the ranking throughout. We attribute this to what the score measures: it compares an image against the variability of authentic images rather than a fixed artifact signature, so a perturbation applied to both classes shifts the reference along with the image.

A transferable direction. Figure 5 shows this directly on two unseen face-swap generators. The raw source axis $\delta = \mu_F - \mu_R$ cleanly separates the training source but collapses to near-chance on

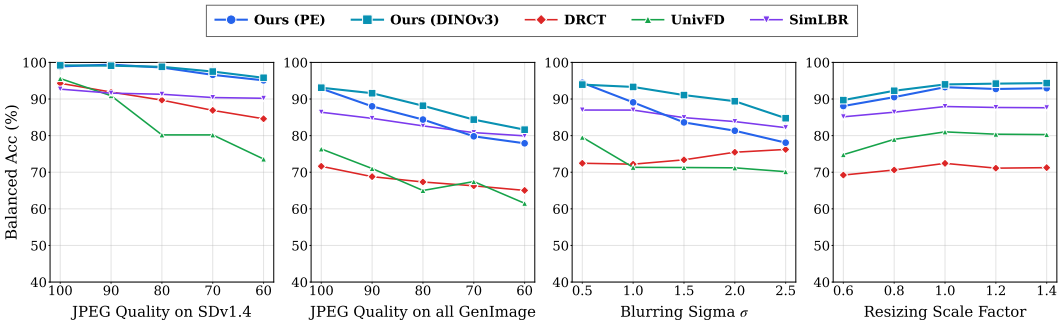


Figure 4: **Robustness to image corruptions on SD v1.4 and GenImage.** Performance is evaluated under JPEG compression (quality factor $Q \in \{100, 90, 80, 70, 60\}$), Gaussian blur (standard deviation $\sigma \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$), and resizing (scaling factor $S \in \{0.6, 0.8, 1.0, 1.2, 1.4\}$). Corruptions are applied to the test images only; no method is trained with corruption augmentation beyond its own protocol.

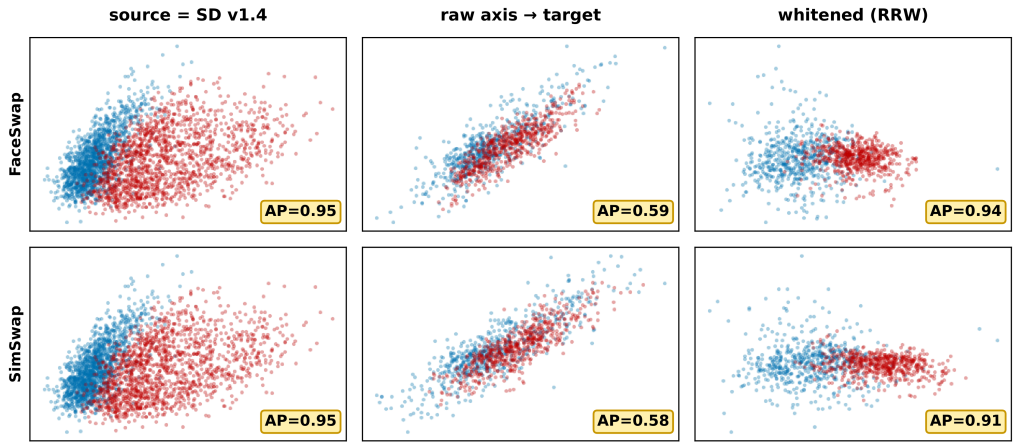


Figure 5: **Whitening makes the real–fake direction transfer.** DINOv3 trained on SD v1.4, tested on two unseen face-swap generators of AIGIBench. The raw source axis separates the source but collapses on an unseen generator; measuring the same displacement in the real-whitened metric restores it.

an unseen generator, being entangled with source content that does not carry over; measuring the *same* displacement in the real-whitened metric restores the separation. The forensic signal is present in the target features but hidden along low-variance real directions that whitening up-weights.

5 Conclusion

We proposed CoRe, a reference-guided framework that measures generated-image evidence relative to authentic images at two levels. Real-Referenced Whitening fits the metric to source real images alone and applies it to both classes, discounting differences that real images already exhibit and amplifying deviations along stable low-variance real directions, where we find transferable evidence to lie. Content-Referenced Fake Contrastive then centers each generated image on its content-matched real neighbors and aligns residuals across content, so the decision depends less on what an image depicts than on how it was produced. The benefit is governed by feature geometry rather than the backbone alone: whitening helps PE and DINOv3, whose real–fake displacement projects onto low-variance real directions, but not CLIP, which is consistent with where that displacement falls in its real-image spectrum. Instantiated on PE, CoRe attains state-of-the-art mean accuracy on AIGIBench and on Chameleon.

AI USE STATEMENT

We used a large language model to aid and polish writing, improving the grammar, clarity, and phrasing of author-written text, and to assist in the retrieval and discovery of related work. Additionally, we used generative AI tools for formatting LaTeX tables and generating boilerplate code for data visualization. We did not use generative AI tools for research ideation, method design, experimental design, implementation of the method, or the analysis of experimental results. All AI-assisted text was read and edited by the authors. The authors take full responsibility for the final content of this work, including all text and claims.

ETHICS STATEMENT

This work studies the detection of AI-generated images, a task motivated by the harms that synthetic imagery enables: impersonation, non-consensual content, fabricated evidence, and misinformation. All experiments use publicly released benchmarks (GenImage, AIGIBench, Chameleon, and the ProGAN training set of ForenSynths) under their original terms, and we neither collect new images nor generate new synthetic data. Several subsets of AIGIBench depict identifiable people, including face-swapped images of real individuals that were created by their original authors rather than by us; we use them only for evaluation, report aggregate metrics, and display no identifiable face in the paper. We did not collect data from human subjects and no annotation was performed for this work.

REPRODUCIBILITY STATEMENT

All information required to reproduce our results is contained in the paper. The method is fully specified in Sections 3.2–3.5: the reference statistics and the whitening transform in Eqs. 2–5, the detection score in Eq. 6, the reference bank and contrastive objective in Eqs. 8–9, and the training objective in Eq. 10. Training settings, backbones, input preprocessing, and the values of all hyperparameters are given in Appendix A.3. The benchmarks, their training sources, and the evaluation protocol we follow for each are described in Section 4.1; all datasets are publicly available and we use the official splits without modification, and we state which numbers are taken from prior work and which are produced by us in the corresponding table captions. Appendix A.1 details the feature-geometry analyses, Appendix A.4 the construction of the reference bank, including the retrieval procedure and the number of neighbors and paired images.

REFERENCES

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.08136.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. URL https://arxiv.org/abs/2504.13181.

Weihan Cai, Hao Tan, Zichang Tan, Jun Wan, and Xinping Gao. Unleashing the potential of vision-language models for generalizable AI-generated image detection. *arXiv preprint arXiv:2608.04935*, 2026. URL https://arxiv.org/abs/2608.04935.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/chen24ay.html.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.14359.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Hyeongjun Choi, Juhun Lee, Davide Cozzolino, Luisa Verdoliva, and Simon S. Woo. Anchor-regularized adaptation for generalizable AI-generated image detection with DINOv3. *arXiv preprint arXiv:2608.15196*, 2026. URL https://arxiv.org/abs/2608.15196.

Aayush Dhakal, Subash Khanal, Srikumar Sastry, Jacob Arndt, Philipe Ambrozio Dias, Dalton Lunga, and Nathan Jacobs. SimLBR: Learning to detect fake images by learning to detect real images. *arXiv preprint arXiv:2602.20412*, 2026. URL https://arxiv.org/abs/2602.20412.

Matthias Dorfer, Rainer Kelz, and Gerhard Widmer. Deep linear discriminant analysis. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.04707.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3247–3258, 2020. URL https://proceedings.mlr.press/v119/frank20a.html.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.17671.

Seoyeon Gye, Junwon Ko, Hyounguk Shon, Minchan Kwon, and Junmo Kim. Reducing the content bias for AI-generated image detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 399–408, February 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Gye_Reducing_the_Content_Bias_for_AI-Generated_Image_Detection_WACV_2025_paper.html.

Steven M. Kay. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall, 1998.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. 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*, 2025a. doi: 10.1145/3690624.3709392. URL https://arxiv.org/abs/2408.06741.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-5739. URL https://papers.nips.cc/paper_files/paper/2025/hash/fb693c67f61e5321746ffce8b6fdd2d0-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: Benchmarking human and model perception of AI-generated images. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.13023.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1):1–41, 2021. doi: 10.1145/3425780. URL https://doi.org/10.1145/3425780.

Lianrui Mu, Xingze Zou, Jianhong Bai, Jiaqi Hu, Wenjie Zheng, Jiangnan Ye, Jiedong Zhuang, Mudassar Ali, Jing Wang, and Haoji Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution AI-generated image detection. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.17346.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, 2023. doi: 10.1109/CVPR52729.2023.02345. URL https://arxiv.org/abs/2302.10174.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2316–2324, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/20130.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114,

2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tan_Learning_on_Gradients_Generalized_Artifacts_Representation_for_GAN-Generated_Images_Detection_CVPR_2023_paper.html.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a. URL https://ojs.aaai.org/index.php/AAAI/article/view/28310.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 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, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tan_Rethinking_the_Up-Sampling_Operations_in_CNN-based_Generative_Network_for_Generalizable_CVPR_2024_paper.html.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025. doi: 10.1609/aaai.v39i7.32772. URL https://ojs.aaai.org/index.php/AAAI/article/view/32772.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, 2020. doi: 10.1109/CVPR42600.2020.00872. URL https://arxiv.org/abs/1912.11035.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DIRE_for_Diffusion-Generated_Image_Detection_ICCV_2023_paper.html.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b0303773962ea1b5394c3a83cc7dd066-Abstract-Conference.html.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 2025b. URL https://proceedings.mlr.press/v267/yan25b.html.

Yawen Yang, Feng Li, Shuqi Kong, Yunfeng Diao, Xinjian Gao, Zenglin Shi, and Meng Wang. Layer consistency matters: Elegant latent transition discrepancy for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.10598.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient AI-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023. URL https://arxiv.org/abs/2311.12397.

Xiaoyu Zhou, Jianwei Fei, Peipeng Yu, Jingchang Xie, Chong Cheng, and Zhihua Xia. PGC: Peak-guided calibration for generalizable AI-generated image detection. *arXiv preprint arXiv:2605.21207*, 2026. URL https://arxiv.org/abs/2605.21207. Accepted at the 43rd International Conference on Machine Learning.

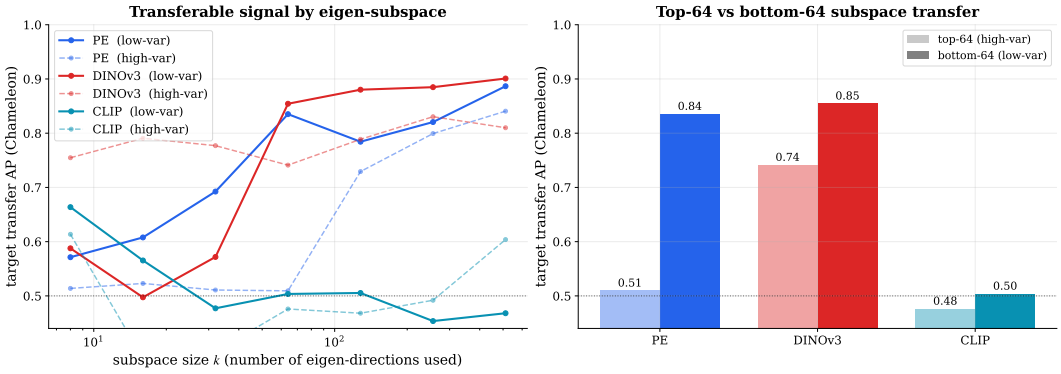
Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, Jinhua Zeng, and Bin Li. Brought a gun to a knife fight: Modern VFM baselines outgun specialized detectors on in-the-wild AI image detection. *arXiv preprint arXiv:2509.12995*, 2025. URL https://arxiv.org/abs/2509.12995.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782, 2023. doi: 10.52202/075280-3398. URL https://papers.nips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

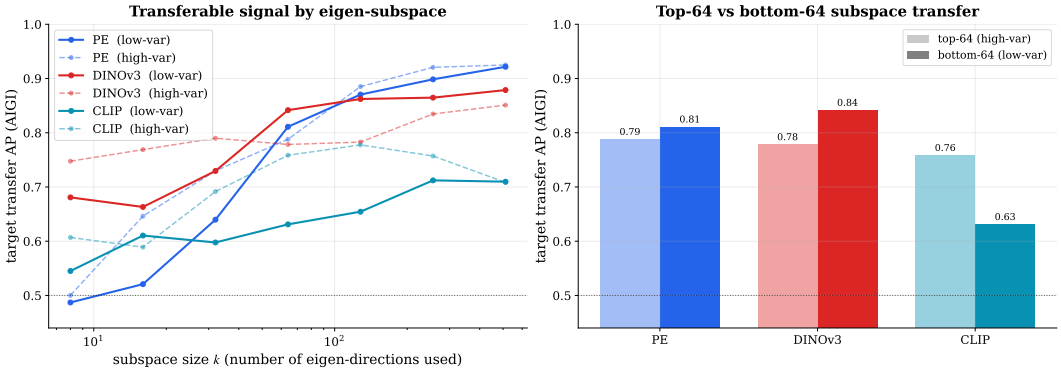
A Appendix

A.1 Analysis of vision foundation models

Low-variance directions carry the transferable signal. Whether a direction separates fake from real on the source says little about whether it *transfers*; what matters is where the cross-domain signal lives. We therefore build the RRW score from isolated eigen-subspaces and measure transfer to *unseen* target generators, on Chameleon and on AIGIBench (Figure 6). For PE and DINOv3 the low-variance tail of the real covariance carries this signal on both targets (bottom-$k \gg$top-k), so down-weighting the high-variance content directions, which is what whitening does, is the right operation. For CLIP the low-variance directions are largely uninformative, and whitening brings little benefit. This suggests a simple check for a new backbone, judged by transfer to unseen generators rather than by a source-only proxy: RRW helps mainly when bottom-k transfer exceeds top-k.



(a) Target = Chameleon.



(b) Target = AIGIBench.

Figure 6: **Where the transferable forgery signal lives, per backbone.** We build the RRW score from *only* the top-k (high-variance, dashed) or bottom-k (low-variance, solid) eigen-directions of the real covariance and measure transfer to *unseen* target generators, for two targets: **(a)** Chameleon and **(b)** AIGI. For PE and DINOv3 the low-variance directions carry the signal, which is exactly what whitening up-weights; for CLIP they do not.

A.2 Results on AIGIBench

We report accuracy (Acc) and average precision (AP) on all 25 AIGIBench test subsets under Setting-II, using ProGAN (four classes) and SD v1.4 as the training sources. Table 5 reports our results alongside baselines transcribed from PGC (Zhou et al., 2026). CoRe with PE obtains a mean Acc of 92.3% and a mean AP of 97.7%, exceeding PGC, the strongest prior method by mean Acc, by 5.7 and 2.7 percentage points, respectively. CoRe with DINOv3 obtains 84.6% Acc and 90.1% AP, the third-highest mean Acc among the methods displayed, after CoRe with PE and PGC.

Table 5: Generalization on AIGIBench (Acc / AP, %) under Setting-II (trained on ProGAN four-class and SD v1.4). Baseline numbers are taken from PGC (Zhou et al., 2026); B-Free and DDA use different training data and are included for reference. **Best** and second-best denote the top two distinct reported values among all displayed methods in each column; ties share the same mark.

	ProGAN		R3GAN		StyleGAN3		StyleGAN-XL		StyleSwin		WFIR		BlendFace		E4S		FaceSwap	
Method	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet	97.6	99.9	50.4	52.7	55.8	73.1	52.8	64.2	52.6	76.5	49.8	50.0	52.4	73.4	51.1	68.9	50.3	58.7
LGrad	96.6	99.8	54.4	58.7	70.5	80.5	65.7	74.6	81.3	90.0	51.7	49.4	41.8	34.9	41.5	32.8	45.3	37.5
UnivFD	98.4	99.9	83.5	91.2	79.6	84.5	84.6	93.3	86.4	95.2	70.0	82.0	35.0	35.3	57.0	57.1	53.1	52.4
FreqNet	99.3	**100.0**	62.3	56.8	83.0	92.4	79.8	84.1	80.8	91.8	58.5	48.9	23.3	34.1	25.8	34.7	40.4	43.4
NPR	99.4	**100.0**	50.8	61.1	78.4	91.7	60.3	75.3	85.7	94.9	51.6	65.5	44.5	34.7	45.0	34.4	48.1	43.6
B-Free	92.7	97.2	82.2	93.5	62.5	79.3	65.2	82.9	70.2	87.2	77.0	92.6	60.3	79.9	77.5	91.8	61.8	78.3
AIDE	97.2	99.6	92.9	97.1	88.1	91.4	88.7	93.2	83.7	89.3	71.4	90.8	51.5	54.2	44.3	44.3	52.1	56.3
SAFE	**100.0**	**100.0**	93.9	98.2	89.7	97.6	93.1	97.6	**97.8**	99.6	60.4	81.8	47.3	45.6	47.6	46.0	50.7	45.7
DDA	84.7	97.7	94.0	98.6	58.3	76.2	46.6	35.9	56.5	68.5	51.7	68.8	**77.9**	**90.0**	82.3	87.8	77.3	84.0
PGC	92.2	99.9	95.4	99.8	90.0	97.7	94.8	96.5	95.5	99.4	52.2	92.5	71.8	82.3	**88.5**	**95.9**	87.0	**95.6**
Ours (DINOv3)	90.9	99.3	89.8	96.6	**95.2**	98.5	91.6	97.8	89.8	96.6	**89.1**	95.5	55.8	60.4	55.0	52.7	52.2	55.7
Ours (PE)	96.2	**100.0**	**97.2**	**100.0**	94.8	**99.9**	**97.3**	**100.0**	97.1	**100.0**	86.1	**100.0**	74.8	77.8	87.2	93.5	**87.7**	95.2

	InSwap		SimSwap		FLUX1-dev		Midjourney-V6		GLIDE		DALL-E 3		Imagen 3		SD3		SDXL	
Method	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet	54.5	77.9	52.1	70.0	57.4	72.3	52.3	59.8	51.1	60.0	53.9	68.6	51.2	57.4	55.8	73.1	52.8	64.2
LGrad	44.6	35.0	44.2	37.6	80.4	88.1	60.4	64.5	82.4	91.0	57.2	62.7	62.2	69.7	63.4	72.1	73.6	82.8
UnivFD	43.7	40.2	43.7	40.4	80.0	79.5	65.3	61.5	76.7	80.3	75.1	76.3	78.9	79.3	84.5	87.2	84.7	88.0
FreqNet	37.5	42.1	36.5	41.9	78.5	87.3	53.9	55.9	75.8	77.4	66.2	61.0	73.6	80.7	77.3	82.6	82.7	95.2
NPR	47.8	40.7	47.4	42.7	95.2	99.0	68.8	76.9	82.5	94.3	57.1	70.0	85.9	94.4	91.9	97.2	86.6	94.4
B-Free	62.1	79.6	63.4	81.2	53.3	55.0	60.0	73.3	83.0	93.1	78.2	88.6	50.9	53.2	69.1	82.8	89.0	96.3
AIDE	50.9	54.6	54.9	62.7	88.0	93.4	76.4	83.0	93.4	97.7	55.1	63.1	89.8	95.2	94.3	98.3	93.5	95.7
SAFE	49.7	49.9	49.0	49.5	**98.1**	99.7	94.1	98.4	92.5	97.9	49.0	45.8	**96.7**	**98.8**	94.1	98.8	**98.3**	99.7
DDA	82.0	87.8	83.3	89.8	95.2	99.3	**96.4**	**99.5**	85.1	93.7	93.1	98.1	90.8	97.2	95.8	**100.0**	95.8	**100.0**
PGC	80.4	92.3	82.7	93.5	81.3	92.2	85.4	95.8	**96.1**	**99.7**	86.3	94.9	86.7	95.6	**96.0**	99.6	96.2	99.9
Ours (DINOv3)	55.8	59.1	57.3	59.7	94.8	99.5	94.2	99.4	91.3	97.4	94.4	**100.0**	92.3	97.3	93.5	**100.0**	93.5	**100.0**
Ours (PE)	**86.2**	**94.7**	**85.9**	**94.2**	97.3	**99.9**	86.3	**99.5**	95.1	99.0	**95.8**	**100.0**	86.1	93.5	**96.0**	**100.0**	96.0	**100.0**

	BLIP		Infinite-ID		InstantID		IP-Adapter		PhotoMaker		SocialRF		CommunityAI		Mean	
Method	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet	77.2	92.9	49.7	49.5	53.2	80.2	52.0	65.8	50.1	58.2	51.1	50.6	51.3	59.1	55.1	67.1
LGrad	93.0	97.4	50.9	54.6	72.6	81.5	70.3	78.3	59.9	67.2	53.5	54.9	55.5	69.4	62.9	66.6
UnivFD	88.6	95.8	84.5	89.6	85.4	93.5	82.6	87.3	69.3	72.3	54.4	55.2	67.0	73.2	72.5	75.6
FreqNet	93.8	**100.0**	79.0	74.5	79.8	86.3	78.8	79.9	77.0	74.9	54.2	58.1	55.9	69.7	66.1	70.1
NPR	99.2	**100.0**	63.9	80.4	63.8	79.2	82.4	91.7	48.1	43.6	59.1	68.4	54.0	62.9	67.9	73.5
B-Free	90.0	96.6	79.5	91.3	75.4	88.5	77.1	89.6	61.6	74.8	66.7	75.7	56.1	61.2	70.6	82.5
AIDE	96.4	95.5	92.2	94.7	91.8	96.3	90.0	95.4	91.7	95.6	57.8	65.0	54.1	61.0	77.6	82.5
SAFE	**99.7**	**100.0**	**96.9**	99.2	**98.2**	99.6	92.8	98.1	**97.0**	**99.3**	58.0	64.2	54.2	55.2	80.0	82.6
DDA	95.9	99.9	96.5	**100.0**	96.6	99.9	95.0	99.1	78.7	90.8	79.6	88.9	88.6	95.7	83.1	89.9
PGC	95.9	98.9	94.9	99.3	95.4	99.6	93.8	98.7	93.7	98.6	74.4	84.1	57.7	72.2	86.6	95.0
Ours (DINOv3)	94.9	99.6	93.8	99.7	93.7	99.5	92.2	98.0	85.8	93.3	84.5	**98.0**	93.1	**98.0**	84.6	90.1
Ours (PE)	97.1	**100.0**	94.9	**100.0**	97.1	**100.0**	**96.0**	**99.3**	95.8	99.2	**90.0**	**98.0**	**93.6**	**98.0**	**92.3**	**97.7**

CoRe with PE has a population standard deviation of 5.6 percentage points across subset accuracies, compared with 10.1 for CNNDet, the next lowest among the methods displayed. Its lowest subset accuracy is 74.8% on BlendFace, 22.6 points above PGC's 52.2%, the highest minimum among the prior methods displayed. On SocialRF and CommunityAI, CoRe with PE obtains 90.0% and 93.6% Acc; these exceed the strongest reported baselines for each subset by 10.4 and 5.0 points, respectively (DDA: 79.6% and 88.6%). On WFIR it obtains 86.1%, 9.1 points above the best reported baseline (B-Free: 77.0%). B-Free and DDA use different training data and are included for reference; the corresponding differences should therefore be interpreted as comparisons of reported results rather than controlled improvements under identical training conditions. Face-swap subsets remain challenging. Compared with the best prior result reported for each subset, CoRe with PE improves on InSwap, SimSwap, and FaceSwap by 4.2, 2.6, and 0.7 points, respectively, but falls behind on BlendFace and E4S by 3.1 and 1.3 points. A single global cls-token score may miss evidence confined to a small edited region; the present results do not establish this as the sole explanation for those gaps.



Figure 7: **Content-referenced fake contrast (CRFC)**, ProGAN with frozen PE features. **Left**: Each fake anchor (red) and its top-3 nearest reals (green). They share the anchor's content (chair → chairs, vehicle → vehicles, ...) and define the content-removed residual $v = h - \mathrm{mean}(\text{reals})$ (the InfoNCE negatives). **Right**: The same anchor and the two fakes it is pulled toward (purple). Content-different yet sharing a forgery residual (a chair with a motorcycle and a car).

A.3 Implementation details

The real statistics (μ_R, σ_R, C_R) and the decision direction are refreshed every $N{=}200$ steps from a running probe of 2,000 examples, so that the metric tracks the backbone as it adapts. The CRFC memory bank stores *raw* initial-model features rather than whitened ones, and is re-whitened into the *current* metric at every refresh (re-projection); this keeps the contrastive objective measured in the same coordinate system as the detector, rather than in a frozen initial coordinate frame. CRFC operates in the whitened space. For each sample we form a residual against the mean of its content-matched real references ($K{=}16$ nearest reals by feature cosine), ℓ_2-normalize it, and apply an InfoNCE loss with temperature $\tau{=}0.1$: positives are content-different fakes and negatives are the content-nearest reals. The contrastive weight is $\lambda{=}0.1$. The total objective is $\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda\,\mathcal{L}_{\mathrm{CRFC}}$. The linear detection head (the calibration parameters α, b) uses a separate learning rate of 10^{-2}, and is supervised with a binary cross-entropy loss.

A.4 Examples from the Reference Bank

Figure 7 shows entries of the reference bank used by CRFC. For each generated image (red), we compute the cosine similarity between its feature and those of all real reference images and keep the k nearest as its content-matched reals (green; we show three of the $k = 16$). Because the feature is dominated by semantics, these neighbors depict the same kind of object or scene: chairs for a generated chair, cars for a generated car, and indoor scenes for a generated interior. To select paired generated images (purple), we first compute a residual for every generated reference image

Table 6: Effect of the contrastive weight λ, evaluated on AIGIBench with PE. Accuracy (%) is the mean over the 25 subsets; λ=0 is RRW alone.

λ	0 (RRW only)	0.1 (ours)	0.3	0.5
Accuracy	88.7	**92.3**	91.6	90.8

by subtracting the mean of its own content-matched reals and normalizing the result. Among the generated reference images whose content similarity to the anchor is below the median, we then keep the two whose residuals have the highest cosine similarity to the anchor's residual. The resulting pairs therefore depict different content, such as vehicles, crowds, or animals paired with a chair, while sharing a similar residual direction, which is what the contrastive objective of Eq. 9 aligns.

A.5 Sensitivity to the contrastive weight

The contrastive term of Eq. 10 is weighted by λ, which balances it against the classification objective. We vary λ while keeping the backbone, the training data, and the schedule fixed, and report accuracy on AIGIBench in Table 6; the column λ=0 corresponds to RRW alone. Accuracy improves over RRW alone for every positive value we tested, peaks at λ=0.1, and declines slowly thereafter: a small weight leaves the contrastive signal too weak to reshape the residual directions, while a large one lets the contrastive term dominate the objective that the detection score is actually trained on. Among the positive values the spread is only 1.5 points, against a 3.6-point gain over λ=0, so the method does not depend on a finely tuned value.

A.6 Limitations.

Both components rely on the source real set: the whitening metric is only as representative as the real images it is fitted on, and content matching is approximated by nearest neighbors in a frozen feature space, so a residual may retain content when the real set covers a scene poorly, and it does not remove factors that are shared across content, such as compression or acquisition pipeline. A broader reference would likely improve both. The optimality of the closed-form direction is a statement about the source statistics; transfer additionally depends on how the displacement of an unseen generator aligns with the source one, which we measure but do not control, and which is weakest for pixel-space diffusion models such as ADM. Finally, the benefit of whitening varies with the backbone's feature geometry, and with three backbones we can report a consistent ordering but not a validated predictor for an unseen backbone.