RELATIVE PATCH RESPONSE LEARNING FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models can now synthesize highly realistic images, simultaneously increasing the risks of misinformation and visual forgery. Therefore, detecting AI-generated images becomes more essential, and a reliable detector must generalize to unseen generators and stay robust to unseen perturbations in the wild. Existing detectors are typically trained on either independently collected real and generated images or aligned real-generated pairs designed to mitigate content bias. Building on aligned pairs, recent methods form a mixed view by replacing some patches of the real image with their generated counterparts. However, we find that self-attention lets real and generated patches interact, so the feature of each patch no longer reflects its own source alone. This ***contextual shift*** makes a per-patch source label an imprecise target. To this end, we propose Relative Patch Response Learning (PRL). Instead of labeling each patch, PRL compares the same patch across two mixed views of an aligned pair and learns from its patch response, the change of its score between the views. (i) To give precise supervision under the contextual shift, a **relative response** objective measures the responses of source-changed patches against those of source-unchanged patches, which respond to the shift alone. (ii) To provide a reliable reference for the shift, a **reference coherence** objective keeps each group of source-unchanged patches moving as a whole. (iii) Since the two views contain different amounts of generated content, an **area ranking** objective asks the view with the larger generated area to have a higher mean patch score. Extensive experiments demonstrate the superior performance of PRL, which surpasses the best prior methods by 4.3% and 5.9% in average balanced accuracy across eight standard and three in-the-wild benchmarks, respectively.

1 INTRODUCTION

Generative models nowadays can produce highly realistic images (Karras et al., 2019; Dhariwal & Nichol, 2021; Rombach et al., 2022; Esser et al., 2024; Lu et al., 2023), expanding the possibilities for visual content creation (OpenAI, 2025; Black Forest Labs et al., 2025; Wu et al., 2025) while increasing the risks of misinformation and visual forgery (Feuerriegel et al., 2023; Chen et al., 2025b; Nightingale & Farid, 2022; Yang et al., 2024). These concerns motivate AI-generated image detection, which aims to distinguish AI-generated images from real ones (Wang et al., 2020; Frank et al., 2020; Chai et al., 2020; Ojha et al., 2023). The task has two central challenges. The first is ***generalization***: detectors are trained on a limited set of generators but must recognize images from unseen models (Corvi et al., 2023; Zhu et al., 2023). The second is ***robustness***: images in the wild are resized, compressed and re-shared online, and a detector must remain effective after such post-processing (Grommelt et al., 2024; Cozzolino et al., 2024; Yan et al., 2025a; Cavia et al., 2024).

To meet these challenges, detectors have relied on hand-crafted features that expose the traces of generation, such as frequency spectra and pixel-level residuals (Frank et al., 2020; Tan et al., 2024b; Li et al., 2025), or on the pretrained representations of vision foundation models (VFMs) (Ojha et al., 2023; Koutlis & Papadopoulos, 2024; Liu et al., 2024; Yan et al., 2025b). In either case, the training data are real and generated images collected separately (Figure 1a). Such unaligned data differ in content, format and resolution as well as in source, and these differences offer a detector shortcuts around the traces of generation (Guillaro et al., 2025; Rajan et al., 2025). Alignment-based methods mitigate content bias by pairing each real image with a reconstruction of matched content,

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.

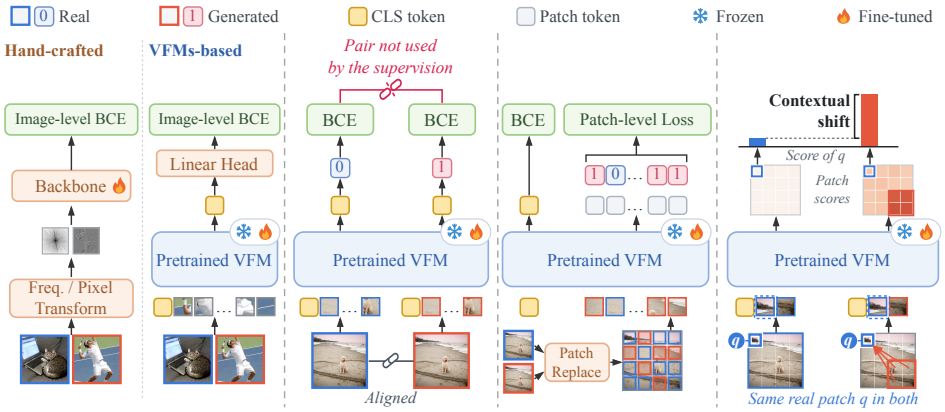


Figure 1: **Three training paradigms and the contextual shift.** (a) Detectors built on hand-crafted features or on vision foundation models (VFMs) learn from real and generated images collected separately. (b) Alignment-based methods pair each real image with a reconstruction of matched content, but treat the two as independent training samples with their own image-level labels; the pair itself is not used by the supervision. (c) Patch-level methods replace some patches of a real image by their reconstruction and label each patch by its own source, in addition to the image-level label. (d) Contextual shift: a real patch q has the same input in the real image and in a mixed view, yet its score rises once other patches are replaced, while the per-patch label of (c) still marks q as real.

and detectors trained on such pairs generalize better (Chen et al., 2024; Rajan et al., 2025; Guillaro et al., 2025; Chen et al., 2025a) (Figure 1b). Yet the alignment only shapes the data. The two images are still trained as independent samples with image-level labels, and their correspondence is never used as supervision. Recent methods go a step further on the aligned pair. PPL (Yang et al., 2026), for example, replaces some patches of a real image with their reconstruction, forming a mixed view in which the source of every patch is known, and supervises each patch by its own source (Figure 1c).

However, such a per-patch label is an imprecise target. In a vision transformer, self-attention lets every token gather evidence from the whole view. In a mixed view, the traces of generation in the replaced patches therefore spread into the real ones, and the real patches in turn dilute the replaced ones, so the feature of a patch moves with its context although the patch itself does not. We call this ***contextual shift*** (Figure 1d). The feature of a patch thus has two parts: the trace of its own source and the shift caused by the other patches, and a label that names only the source cannot tell them apart. It pushes the whole feature toward one class, so the supervision meant for the local trace is mixed with a context term, which Figure 2 makes visible in trained detectors. The label is exact about where the generated patches are, but it says nothing about what the feature of a patch should be once the rest of the view has changed.

To this end, we propose Relative Patch Response Learning (PRL), which supervises responses instead of labels. From each aligned pair, we build two mixed views that differ in which patches are replaced (Figure 3), so that every patch has aligned content in both views and only its source may differ. A linear head maps the feature of each patch to a score, higher for stronger evidence of generation, and the change of this score between the two views is its **patch response**. For a patch whose source is the same in both views, the input is identical, so the response is the contextual shift alone; for a patch whose source differs, the response adds the effect of its own change. The same-source patches thus serve as references for the source-changed ones. PRL has three objectives. (i) **Relative response** supervises each source-changed group relative to its reference, which subtracts the shift and leaves a precise target. (ii) **Reference coherence** keeps each reference group moving as a whole, which makes it a stable and reliable reference. (iii) **Area ranking** asks the view with the larger generated area to score higher. Experiments validate this design. Supervising the fine-grained correspondence within aligned pairs while accounting for the contextual shift improves the detector on both fronts. PRL averages 96.40% balanced accuracy on eight standard benchmarks of diverse generators and 93.74% on three in-the-wild benchmarks, surpassing the best previous method by 4.3% and 5.9%, respectively. Our contributions are as follows:

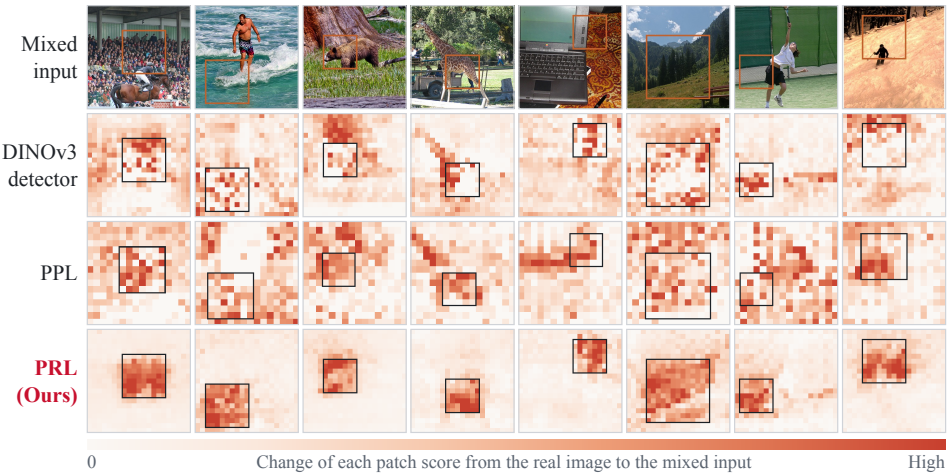


Figure 2: **Contextual shift in trained detectors.** Each mixed input takes the outlined square from the aligned reconstruction and the rest from the real image. **The maps show how much each patch score changes from the real image to the mixed input, darker for a larger change.** For DINOv3 and PPL, we train a patch head on them. Although only the square changed, the DINOv3 detector and PPL respond outside it almost as much as inside it. PRL keeps a shift outside the square and responds inside it.

- We point out that existing methods underuse the correspondence between aligned data. Image-level methods train the two images independently, and patch-level supervision on mixed views ignores the contextual shift, which makes its per-patch labels imprecise targets.
- We propose PRL, which supervises patch responses rather than labels. Relative response supervision reduces the shared contextual component against the same-source patches, reference coherence keeps the reference stable, and area-based ranking adds a view-level constraint.
- On eight standard benchmarks of diverse generators and three in-the-wild benchmarks, PRL surpasses the best previous methods by 4.3% and 5.9% in average balanced accuracy, respectively. It also transfers to the detection of locally manipulated images with no additional training data.

2 Related Work

Forensic traces and pretrained representations. One family of detectors identifies generated images from the traces left by the generation process, including frequency-domain statistics (Frank et al., 2020; Tan et al., 2024a; Li et al., 2025; Karageorgiou et al., 2025; Li et al., 2026), pixel-level relations and residuals (Tan et al., 2024b; Fu et al., 2025) and reconstruction or regeneration error (Wang et al., 2023; Ricker et al., 2024; Luo et al., 2024; Yu et al., 2024; Chu et al., 2025; Qi et al., 2026), or from the trace of the camera pipeline that generated images lack (Wu et al., 2026). Another line builds on the pretrained representations of vision foundation models, with light classifiers on frozen features (Ojha et al., 2023; Koutlis & Papadopoulos, 2024; Yang et al., 2025) or with representation adaptation (Liu et al., 2024; Tan et al., 2025; Yan et al., 2025b); GAPL (Qin et al., 2026) organizes the feature space with generator-aware prototypes, and others scale the training set to thousands of generators (Park & Owens, 2025). Both lines are typically trained on real and generated images collected separately.

Aligned training data. Separately collected real and generated images differ in content, format and resolution as well as in source, and a detector can separate them by these differences instead of by the traces of generation (Guillaro et al., 2025; Rajan et al., 2025). Alignment removes content shortcut by giving each real image a generated counterpart of matched content. DRCT (Chen et al., 2024) and B-Free (Guillaro et al., 2025) construct generated samples of similar content through reconstruction or regeneration guided by real images; AlignedForensics (Rajan et al., 2025) obtains a tighter spatial correspondence through VAE reconstruction; DDA (Chen et al., 2025a) further removes the high-frequency difference and interpolates pixels within each pair. Detectors trained on such pairs generalize better across generators than those trained on separately collected images (Ra-

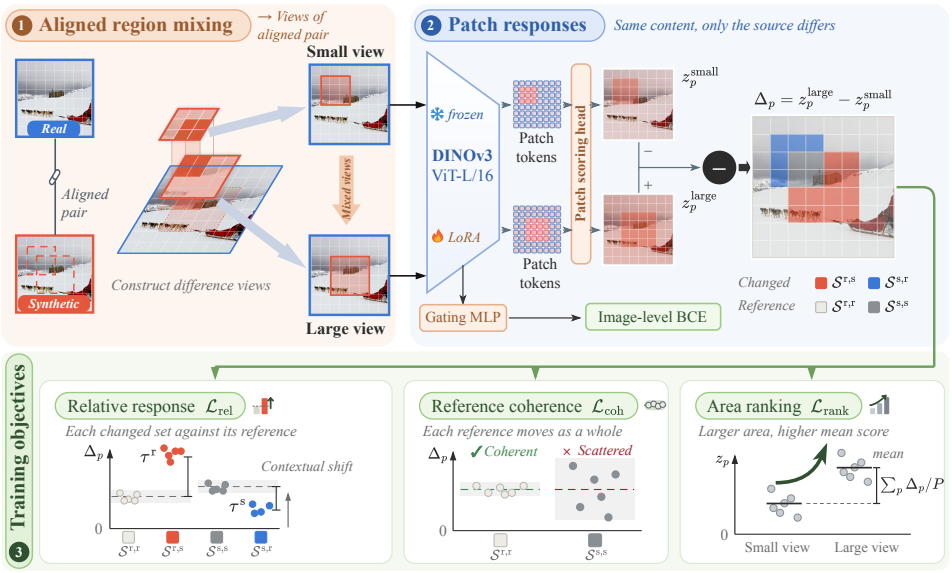


Figure 3: **Overview of PRL. ❶ Aligned region mixing**: two masks mix the real and the synthetic image into a small and a large view. **❷ Patch responses**: the patch scoring head scores every position in both views, and the patch response Δ_p (Eq. (3)) is read on the changed sets $\mathcal{S}^{\text{r,s}}$, $\mathcal{S}^{\text{s,r}}$ and on their references $\mathcal{S}^{\text{r,r}}$, $\mathcal{S}^{\text{s,s}}$; the gating MLP gives the image-level decision. **❸ Training objectives**: relative response, reference coherence and area ranking.

jan et al., 2025; Guillaro et al., 2025; Chen et al., 2025a). In all of these methods, however, the alignment only shapes the data. The two images of a pair are trained as independent samples with image-level labels, and their correspondence is never used as supervision.

Mixed views and patch-level supervision. Mixing real and generated content yields harder training samples. Without alignment, SimLBR (Dhakal et al., 2026) blends the feature of a real image with that of an unpaired generated image in the latent space of a foundation model and labels the blend as generated, which widens the training distribution but can only carry an image-level label. Within an aligned pair, the pixel-wise interpolation of DDA is again labeled at the image level, whereas PPL (Yang et al., 2026) replaces some real patches by their reconstruction and supervises each patch by its own source. Such a label, however, is decided by that patch alone and ignores the contextual shift. PRL instead compares the same patch across two mixed views and measures its response against the patches whose source is the same in both views.

3 Method

In this section, we detail PRL, whose framework is shown in Figure 3. Specifically, PRL supervises a ViT detector with the patch response of an aligned pair: how the score at each patch position changes between two mixed views of the pair. It has two parts: aligned region mixing, which constructs the mixed views, and response learning, which learns the traces of generation from the patch responses. Given a real image and its aligned synthetic image, we mix them patch by patch into two views and train the detector on how each patch score changes between the views.

3.1 Aligned region mixing

Given a real image $\boldsymbol{I}^{\text{r}}$ and its synthetic image $\boldsymbol{I}^{\text{s}}$, where the synthetic image is aligned with the real image, we use $\boldsymbol{\Omega}$ to split both images into P patches as $\mathcal{X}^{\text{r}} = \boldsymbol{\Omega}(\boldsymbol{I}^{\text{r}}) = \{\boldsymbol{x}_p^{\text{r}}\}_{p=1}^{P}$ and $\mathcal{X}^{\text{s}} = \boldsymbol{\Omega}(\boldsymbol{I}^{\text{s}}) = \{\boldsymbol{x}_p^{\text{s}}\}_{p=1}^{P}$, respectively, where $\boldsymbol{\Omega}$ is the function that splits an image into non-overlapping patches on the patch grid of the encoder and P is the number of patches. We then define a binary mixture mask $M = (m_1, \ldots, m_P) \in \{0,1\}^P$ to construct a new synthetic image, which we call a mixed view, from the patches

$$\hat{\boldsymbol{x}}_p = m_p\, \boldsymbol{x}_p^{\text{s}} + (1 - m_p)\, \boldsymbol{x}_p^{\text{r}}, \tag{1}$$

so that the mixed view consists of the patches $\hat{\mathcal{X}} = \{\hat{\boldsymbol{x}}_p\}_{p=1}^{P}$. The mask is generated at random: $m_p = 1$ takes the patch $\boldsymbol{x}_p^{\mathrm{s}}$ of $\boldsymbol{I}^{\mathrm{s}}$ at position p, and $m_p = 0$ takes $\boldsymbol{x}_p^{\mathrm{r}}$ of $\boldsymbol{I}^{\mathrm{r}}$. Thus, different masks M yield a large set of mixed views for training.

For each aligned pair, we sample two masks to construct two mixed views. The sum $\sum_p m_p$ counts the synthetic patches of a view, so a larger sum means more synthetic patches, and we require the two sums to differ; Appendix A details how the masks are sampled. Ordered by the sum, we denote the two masks by M^{small} and M^{large}, with entries m_p^{small} and m_p^{large}, and the mixed views they construct by $\hat{\mathcal{X}}^{\mathrm{small}} = \{\hat{\boldsymbol{x}}_p^{\mathrm{small}}\}_{p=1}^{P}$ and $\hat{\mathcal{X}}^{\mathrm{large}} = \{\hat{\boldsymbol{x}}_p^{\mathrm{large}}\}_{p=1}^{P}$. We then define four sets of positions by the patches the two views take:

$$\begin{aligned} \mathcal{S}^{\mathrm{r,r}} &= \{p : \hat{\boldsymbol{x}}_p^{\mathrm{small}} = \boldsymbol{x}_p^{\mathrm{r}},\ \hat{\boldsymbol{x}}_p^{\mathrm{large}} = \boldsymbol{x}_p^{\mathrm{r}}\}, & \mathcal{S}^{\mathrm{r,s}} &= \{p : \hat{\boldsymbol{x}}_p^{\mathrm{small}} = \boldsymbol{x}_p^{\mathrm{r}},\ \hat{\boldsymbol{x}}_p^{\mathrm{large}} = \boldsymbol{x}_p^{\mathrm{s}}\}, \\ \mathcal{S}^{\mathrm{s,s}} &= \{p : \hat{\boldsymbol{x}}_p^{\mathrm{small}} = \boldsymbol{x}_p^{\mathrm{s}},\ \hat{\boldsymbol{x}}_p^{\mathrm{large}} = \boldsymbol{x}_p^{\mathrm{s}}\}, & \mathcal{S}^{\mathrm{s,r}} &= \{p : \hat{\boldsymbol{x}}_p^{\mathrm{small}} = \boldsymbol{x}_p^{\mathrm{s}},\ \hat{\boldsymbol{x}}_p^{\mathrm{large}} = \boldsymbol{x}_p^{\mathrm{r}}\}. \end{aligned} \tag{2}$$

On $\mathcal{S}^{\mathrm{r,s}}$ and $\mathcal{S}^{\mathrm{s,r}}$ the patch changes source between the views; these are the changed sets in Figure 3, red and blue. On $\mathcal{S}^{\mathrm{r,r}}$ and $\mathcal{S}^{\mathrm{s,s}}$ the patch stays the same; these are the reference sets, light and dark grey. Because the pair is aligned, the patch at every position shows aligned content in both views, and only its source may differ.

3.2 Response learning

Patch response. Both mixed views are encoded by $\mathcal{F}_\phi$, a frozen DINOv3 ViT adapted with LoRA parameters ϕ (Hu et al., 2022). The encoder takes all P patches of a view and returns one final-layer token per position, $\boldsymbol{h}_p^{\mathrm{small}} = \mathcal{F}_\phi(\hat{\mathcal{X}}^{\mathrm{small}})_p$ and $\boldsymbol{h}_p^{\mathrm{large}} = \mathcal{F}_\phi(\hat{\mathcal{X}}^{\mathrm{large}})_p$. A linear *patch scoring head* $\boldsymbol{w}$ maps each token to a scalar *patch score*, $z_p^{\mathrm{small}} = \boldsymbol{w}^\top \boldsymbol{h}_p^{\mathrm{small}}$ and $z_p^{\mathrm{large}} = \boldsymbol{w}^\top \boldsymbol{h}_p^{\mathrm{large}}$, higher for stronger evidence of synthesis. The *patch response* at position p is the change of its score from the small to the large view,

$$\Delta_p = z_p^{\mathrm{large}} - z_p^{\mathrm{small}} = \boldsymbol{w}^\top \big(\boldsymbol{h}_p^{\mathrm{large}} - \boldsymbol{h}_p^{\mathrm{small}}\big). \tag{3}$$

Since the content at p is aligned across the two views, Δ_p is not driven by content. For $p \in \mathcal{S}^{\mathrm{r,r}} \cup \mathcal{S}^{\mathrm{s,s}}$, the patch at position p is identical in the two views, so Δ_p comes only from the patches that changed elsewhere in the image. This is the contextual shift (Figure 2), and the reference sets measure it directly. For $p \in \mathcal{S}^{\mathrm{r,s}} \cup \mathcal{S}^{\mathrm{s,r}}$, Δ_p responds both to the change of the patch at p and to the changed context, and a single position cannot tell the two apart. We therefore measure each changed set against the reference set with the same source in the small view: $\mathcal{S}^{\mathrm{r,s}}$ against $\mathcal{S}^{\mathrm{r,r}}$, and $\mathcal{S}^{\mathrm{s,r}}$ against $\mathcal{S}^{\mathrm{s,s}}$. The two sets of each couple start from the same source and are exposed to the same change in the rest of the image; they differ mostly in whether their own patch changes, so the mean response of the reference serves as the estimate of the contextual shift of its changed set.

Relative response ($\mathcal{L}_{\mathrm{rel}}$). A change from real to synthetic should raise the score of a position above its reference, and a change from synthetic to real should lower it below its reference. In both cases, of the two sets that share a source in the small view, the one synthetic in the large view should respond more than the one real in the large view. For the positions real and synthetic in the small view, we measure this by the *relative responses*

$$\tau^{\mathrm{r}} = \frac{1}{|\mathcal{S}^{\mathrm{r,s}}|} \sum_{p \in \mathcal{S}^{\mathrm{r,s}}} \Delta_p - \mathrm{sg}\Big(\frac{1}{|\mathcal{S}^{\mathrm{r,r}}|} \sum_{p \in \mathcal{S}^{\mathrm{r,r}}} \Delta_p\Big), \quad \tau^{\mathrm{s}} = \mathrm{sg}\Big(\frac{1}{|\mathcal{S}^{\mathrm{s,s}}|} \sum_{p \in \mathcal{S}^{\mathrm{s,s}}} \Delta_p\Big) - \frac{1}{|\mathcal{S}^{\mathrm{s,r}}|} \sum_{p \in \mathcal{S}^{\mathrm{s,r}}} \Delta_p, \tag{4}$$

where $|\cdot|$ is the number of elements in a set and $\mathrm{sg}(\cdot)$ is the stop-gradient function: it returns its input unchanged in the forward pass and passes no gradient to it in the backward pass. Applied to the means over the references $\mathcal{S}^{\mathrm{r,r}}$ and $\mathcal{S}^{\mathrm{s,s}}$, it keeps this loss from moving the references. With the softplus activation $\mathrm{softplus}(\cdot) = \log\big(1 + \exp(\cdot)\big)$, the loss asks both relative responses to be positive,

$$\mathcal{L}_{\mathrm{rel}} = \tfrac{1}{2}\big[\,\mathrm{softplus}(-\tau^{\mathrm{r}}) + \mathrm{softplus}(-\tau^{\mathrm{s}})\big]. \tag{5}$$

If a set in τ^{r} or τ^{s} is empty, that term is dropped and the other is used alone. Subtracting the reference reduces the contextual shift.

Reference coherence ($\mathcal{L}_{\text{coh}}$). The mean of a reference is a good estimate of the contextual shift only if its positions move together. We therefore penalize the deviation of every position in $\mathcal{S}^{\text{r,r}}$ and $\mathcal{S}^{\text{s,s}}$ from the mean response of its set, averaged over both sets,

$$\mathcal{L}_{\text{coh}} = \frac{1}{|\mathcal{S}^{\text{r,r}}| + |\mathcal{S}^{\text{s,s}}|} \left[\sum_{p \in \mathcal{S}^{\text{r,r}}} \left(\Delta_p - \frac{1}{|\mathcal{S}^{\text{r,r}}|} \sum_{q \in \mathcal{S}^{\text{r,r}}} \Delta_q \right)^2 + \sum_{p \in \mathcal{S}^{\text{s,s}}} \left(\Delta_p - \frac{1}{|\mathcal{S}^{\text{s,s}}|} \sum_{q \in \mathcal{S}^{\text{s,s}}} \Delta_q \right)^2 \right]. \quad (6)$$

Removing the mean leaves the shift itself free: a reference may shift with the context, but it has to shift as a whole.

Area ranking ($\mathcal{L}_{\text{rank}}$). The two views also differ in how much of the view is synthetic: the large view has $|\mathcal{S}^{\text{r,s}}| - |\mathcal{S}^{\text{s,r}}|$ more synthetic patches than the small view, a number that is always positive because the two masks have different sums. The loss

$$\mathcal{L}_{\text{rank}} = \frac{|\mathcal{S}^{\text{r,s}}| - |\mathcal{S}^{\text{s,r}}|}{P} \operatorname{softplus}\left(- \frac{\sum_{p=1}^{P} \Delta_p}{|\mathcal{S}^{\text{r,s}}| - |\mathcal{S}^{\text{s,r}}|} \right) \quad (7)$$

asks the large view to have a higher mean patch score, that is, a positive $\sum_p \Delta_p$. Inside the softplus, the score gain is divided by the number of added synthetic patches, so the loss looks at the gain per added patch. The factor in front keeps the gradient bounded for any area gap.

Training and inference. The image-level decision comes from the *detection head* g, a gating MLP that weights the tokens of an image, pools them and maps the result to a logit. For an image with patches $\mathcal{X}$, the detector outputs the probability $\hat{y}$ that the image is synthetic, and is trained with the binary cross-entropy $\mathcal{L}_{\text{img}}$,

$$\hat{y} = \sigma\big(g(\mathcal{F}_\phi(\mathcal{X}))\big), \qquad \mathcal{L}_{\text{img}} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}), \quad (8)$$

where σ is the sigmoid function and the label y is 0 for the real image and 1 for every image that contains synthetic patches; $\mathcal{L}_{\text{img}}$ is averaged over a batch with real and synthetic images weighted equally. The training loss is

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \lambda_{\text{rel}} \mathcal{L}_{\text{rel}} + \lambda_{\text{coh}} \mathcal{L}_{\text{coh}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}, \quad (9)$$

with which we train $\mathcal{F}_\phi$, $\boldsymbol{w}$ and g. The response terms reach the encoder through the patch scores and thereby shape the tokens that g classifies. **At inference**, the detector takes a single image $\boldsymbol{I}$ and outputs $\hat{y}$ of Eq. (8) with $\mathcal{X} = \boldsymbol{\Omega}(\boldsymbol{I})$.

4 Experiments

4.1 Experimental Setup

Implementation Details We build on DINOv3 ViT-L/16 (Siméoni et al., 2025), kept frozen and adapted with LoRA (Hu et al., 2022). The training pairs are those of DDA (Chen et al., 2025a): real images from MSCOCO (Lin et al., 2014) and, for each, its reconstruction through the VAE of Stable Diffusion 2.1 (Rombach et al., 2022). Images are 336×336 at training and at inference. We train with Adam (Kingma & Ba, 2015) at a learning rate of 10^{-4} and a batch of 16 pairs per GPU on two A100 GPUs, with $(\lambda_{\text{rel}}, \lambda_{\text{coh}}, \lambda_{\text{rank}}) = (0.3,\ 0.3,\ 0.1)$. Further details are given in Appendix A.

Evaluation Benchmarks We evaluate on eleven benchmarks, which we divide into two groups. Eight standard benchmarks are assembled from known generators: GenImage (Zhu et al., 2023), AIGCDetectBenchmark (AIGCDetect) (Zhong et al., 2023), DRCT-2M (Chen et al., 2024), DDA-COCO and EvalGEN (Chen et al., 2025a), Synthbuster (Bammey, 2024), ForenSynths (Wang et al., 2020) and UnivFD (Ojha et al., 2023); together they cover diverse generators, including GANs, diffusion models and auto-regressive models. Three in-the-wild benchmarks collect images from the web, with unknown generators and post-processing: Chameleon (Yan et al., 2025a), Synth-WildX (Cozzolino et al., 2024) and WildRF (Cavia et al., 2024). A third group, of locally manipulated images, is introduced in Section 4.3. Appendix B describes the benchmarks in more detail.

Table 1: **Cross-benchmark comparison.** Balanced accuracy (%) on eight standard and three in-the-wild benchmarks; EvalGEN has no real counterpart and is reported as accuracy on its synthetic images. Every compared method uses its official checkpoint and inference code. Bold: best per column; underlined: second best.

Method	Standard Benchmarks								In-the-Wild Benchmarks			
	GenImage	AIGCDetect	DRCT-2M	DDA-COCO	EvalGEN	Synthbuster	ForenSynths	UnivFD	Chameleon	SynthWildX	WildRF	**Avg.**
UnivFD CVPR'23	64.05	72.50	59.64	52.32	15.39	55.31	77.69	78.95	50.71	52.27	55.08	57.63
FatFormer CVPR'24	61.03	74.78	50.82	51.63	27.17	53.62	86.13	82.07	51.11	53.31	59.62	59.21
NPR CVPR'24	46.74	47.78	49.62	49.07	3.43	41.83	37.88	41.92	52.26	49.80	59.34	43.61
DRCT ICML'24	80.42	67.67	90.41	66.89	50.93	73.04	54.86	64.27	68.93	74.47	74.22	69.65
C2P-CLIP AAAI'25	62.09	70.16	54.29	50.90	13.82	68.08	87.00	84.88	50.60	53.07	59.14	59.46
AIDE ICLR'25	60.35	53.98	58.94	50.09	13.09	56.87	42.36	47.05	57.36	55.25	62.08	50.67
AlignedForensics ICLR'25	77.44	65.63	95.13	86.32	61.42	77.29	51.47	60.47	68.04	79.21	82.20	73.15
Effort ICML'25	79.55	77.00	67.42	51.88	94.41	75.68	64.17	70.19	58.76	58.53	72.95	70.05
SAFE KDD'25	51.14	50.02	50.42	50.28	1.10	52.75	40.67	43.73	52.51	49.93	58.97	45.59
DDA NeurIPS'25	91.74	87.83	98.47	92.89	96.86	85.52	81.41	82.58	82.41	90.87	90.20	89.16
GAPL CVPR'26	95.56	96.08	95.23	79.27	96.10	90.63	90.94	92.89	69.30	86.23	86.01	88.93
PRL (Ours)	**96.83**	**97.75**	**99.27**	**94.39**	**99.14**	**97.00**	**92.57**	**94.22**	**90.01**	**93.60**	**97.59**	**95.67**

Evaluation Metrics and Comparative Methods We report balanced accuracy, the mean of the accuracies on real and on synthetic images. Following DDA, on benchmarks whose synthetic images are stored losslessly while real ones are JPEG (GenImage, AIGCDetect, ForenSynths and UnivFD), synthetic images are JPEG-compressed at quality 96 before evaluation, so that no detector can rely on the format bias. We compare with eleven detectors: UnivFD (Ojha et al., 2023), FatFormer (Liu et al., 2024), NPR (Tan et al., 2024b), DRCT (Chen et al., 2024), C2P-CLIP (Tan et al., 2025), AIDE (Yan et al., 2025a), AlignedForensics (Rajan et al., 2025), Effort (Yan et al., 2025b), SAFE (Li et al., 2025), DDA (Chen et al., 2025a) and GAPL (Qin et al., 2026). We evaluate each method using its official checkpoint and inference code; Appendix C describes the methods in more detail.

4.2 Cross-Benchmark Comparison

Overall comparison on multiple benchmarks. Table 1 compares PRL with the eleven detectors on the eleven benchmarks. PRL reaches an average balanced accuracy of 95.67%, surpassing DDA at 89.16% by 6.5% and GAPL at 88.93% by 6.7%, and is the best method on every benchmark. This consistent margin supports the design of PRL: supervising patch responses against internal references, instead of labeling patches, generalizes across every benchmark. PRL also performs best on Chameleon, the most challenging in-the-wild benchmark, with 90.01% against 82.41% for the strongest prior method. The two strongest prior methods fail in different places. GAPL, whose training set spans thousands of generators, stays close to PRL on the standard benchmarks but drops on all three in-the-wild benchmarks, to 69.30% on Chameleon, and to 79.27% on DDA-COCO; DDA, the best previous method, is strong on most benchmarks but falls behind on the in-the-wild benchmarks and on the GAN-based ones, with 81.41% on ForenSynths. Tables 2 and 3 and Appendix D report the detailed results for every subset.

Comparison on standard benchmarks. Table 2 gives the per-subset results on UnivFD; those of GenImage, AIGCDetect, Synthbuster, DRCT-2M, DDA-COCO and EvalGEN, and ForenSynths are in Appendix D. On ForenSynths, whose generators are GANs and other CNNs, the alignment-based methods DDA and AlignedForensics, trained on VAE reconstructions, fall to 81.41% and 51.47%, whereas PRL, trained on the same kind of pairs, reaches 92.57%; on DRCT-2M, whose sixteen subsets come from Stable Diffusion models and their predecessor LDM, PRL reaches 99.27% against 98.47% for DDA. GAPL is weakest on DDA-COCO, 79.27% against 94.39% for PRL. PRL is above 90% on every subset except five. On the DeepFake face swaps of ForenSynths it reaches 55.26%, where the CLIP-based detectors are stronger. On SITD and SAN, which hold low-light photographs and super-resolved images rather than generated ones, it reaches 84.72% and 88.81%. On the SD3.5 and Flux reconstructions of DDA-COCO it reaches 85.17% and 85.19%.

Comparison on in-the-wild benchmarks. Table 3 covers images collected from the web, with unknown generators and post-processing. PRL is above 90% on all three benchmarks and on every one of their subsets; the gap to the best prior method is largest on Chameleon, where every method drops the most.

Table 2: Comparison of balanced accuracy (%) on the UnivFD benchmark.

Method	BigGAN	CRN	CycleGAN	DeepFake	GauGAN	IMLE	ProGAN	SAN	SITD	StarGAN	StyleGAN	DALL-E	Glide-50-27	Glide-100-10	Glide-100-27	Guided	LDM-100	LDM-200	LDM-200-cfg	Avg.
UnivFD CVPR'23	87.50	55.67	96.92	69.39	98.79	68.11	99.38	58.22	62.22	95.12	79.99	79.55	77.80	77.40	77.80	67.15	91.20	90.75	67.05	78.95
FatFormer CVPR'24	97.90	69.32	99.32	86.03	99.17	69.45	99.21	62.10	80.83	**99.75**	87.51	88.40	70.10	69.60	70.60	61.45	88.40	89.45	70.75	82.07
NPR CVPR'24	43.25	0.00	59.83	33.30	35.62	0.00	50.45	48.40	22.22	49.47	50.18	52.30	50.50	50.85	51.00	45.75	51.15	51.10	51.05	41.92
DRCT ICML'24	52.45	49.80	48.83	56.38	49.45	49.26	53.02	86.76	54.44	54.38	54.45	79.05	60.05	66.45	63.15	59.00	94.80	94.80	94.65	64.27
C2P-CLIP AAAI'25	92.07	92.75	**99.39**	**91.85**	98.20	**97.68**	97.10	74.20	**87.22**	99.70	76.97	84.90	78.90	78.60	76.35	57.05	82.70	83.65	63.40	84.88
AIDE ICLR'25	43.68	19.96	51.73	45.54	39.73	19.38	49.31	51.37	30.56	51.43	49.57	51.80	51.25	51.55	51.50	50.25	60.80	62.25	62.35	47.05
AlignedForensics ICLR'25	50.52	49.95	49.69	52.69	50.24	49.91	50.35	61.87	50.56	51.50	51.46	74.25	52.20	52.30	52.10	50.20	**99.70**	**99.70**	**99.70**	60.47
Effort ICML'25	73.60	39.20	81.49	64.59	62.52	29.54	70.08	58.90	83.89	96.02	56.42	78.90	80.70	82.10	83.45	59.65	80.15	79.75	72.65	70.19
SAFE KDD'25	48.95	0.11	52.58	49.09	42.60	0.10	49.96	50.46	35.83	49.95	49.99	49.55	50.90	51.45	51.15	50.40	49.15	49.20	49.45	43.73
DDA NeurIPS'25	90.97	86.98	72.47	76.53	92.74	89.71	92.80	94.75	58.61	72.74	87.80	75.05	76.75	80.05	78.10	90.25	84.15	84.20	84.45	82.58
GAPL CVPR'26	98.62	67.85	97.54	84.88	**99.63**	67.12	**99.83**	**96.58**	82.50	97.05	97.89	97.15	**97.45**	**97.15**	**97.00**	94.15	97.50	97.55	97.55	92.89
PRL (Ours)	**99.28**	**96.07**	97.37	55.26	99.55	96.08	98.99	88.81	84.72	99.25	**98.21**	**98.25**	96.65	95.95	96.25	**94.95**	98.25	98.25	98.10	**94.22**

Table 3: Comparison of balanced accuracy (%) on the in-the-wild benchmarks Chameleon, SynthWildX and WildRF.

Method	Chameleon	SynthWildX				WildRF			
	Chameleon	DALL-E 3	Firefly	Midjourney	**Avg.**	Facebook	Reddit	Twitter	**Avg.**
UnivFD CVPR'23	50.71	45.42	65.24	46.14	52.27	49.06	60.20	55.98	55.08
FatFormer CVPR'24	51.11	48.01	61.62	50.30	53.31	54.37	69.67	54.82	59.62
NPR CVPR'24	52.26	49.53	50.55	49.31	49.80	50.31	74.93	52.78	59.34
DRCT ICML'24	68.93	84.72	59.24	79.45	74.47	79.06	66.27	77.33	74.22
C2P-CLIP AAAI'25	50.60	49.97	59.73	49.53	53.07	51.88	71.60	53.94	59.14
AIDE ICLR'25	57.36	63.47	50.60	51.67	55.25	56.88	72.07	57.31	62.08
AlignedForensics ICLR'25	68.04	86.85	56.95	93.83	79.21	93.44	70.87	82.31	82.20
Effort ICML'25	58.76	59.32	61.27	55.02	58.53	74.06	79.33	65.47	72.95
SAFE KDD'25	52.51	49.65	50.48	49.65	49.93	50.94	74.07	51.90	58.97
DDA NeurIPS'25	82.41	92.22	87.30	93.09	90.87	93.12	86.40	91.07	90.20
GAPL CVPR'26	69.30	87.06	87.22	84.39	86.23	88.44	82.20	87.40	86.01
PRL (Ours)	**90.01**	**94.92**	**91.31**	**94.58**	**93.60**	**97.19**	**97.20**	**98.39**	**97.59**

Table 4: Comparison of balanced accuracy (%) on AutoSplice at JPEG quality 100, 90 and 75.

Method	JPEG 100	JPEG 90	JPEG 75	Avg.
UnivFD CVPR'23	64.27	58.12	57.47	59.95
FatFormer CVPR'24	73.78	57.84	53.88	61.83
NPR CVPR'24	51.37	50.30	50.18	50.62
DRCT ICML'24	70.04	62.90	60.84	64.59
C2P-CLIP AAAI'25	75.56	54.69	51.81	60.69
AIDE ICLR'25	55.22	51.52	51.27	52.67
AlignedForensics ICLR'25	50.34	50.12	50.80	50.42
Effort ICML'25	69.15	53.82	51.23	58.07
SAFE KDD'25	50.19	50.17	50.14	50.17
DDA NeurIPS'25	73.09	70.91	70.87	71.62
GAPL CVPR'26	82.62	78.91	78.64	80.06
PRL (Ours)	**86.45**	**84.74**	**80.36**	**83.85**

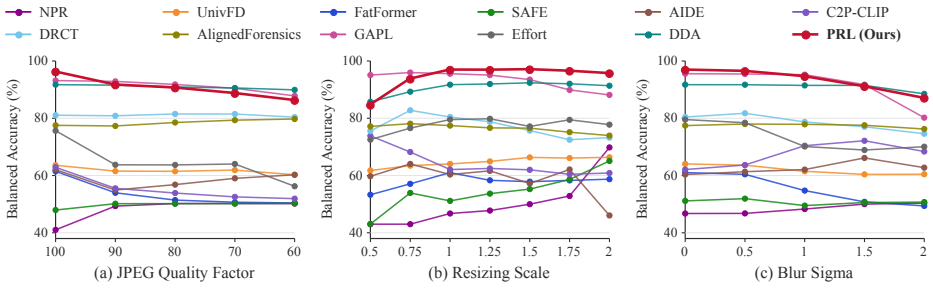


Figure 4: **Robustness to post-processing.** Balanced accuracy on debiased GenImage-JPEG96 dataset, when every test image is JPEG-compressed at decreasing quality (a), resized by a scale factor (b) or blurred with a Gaussian of increasing σ (c).

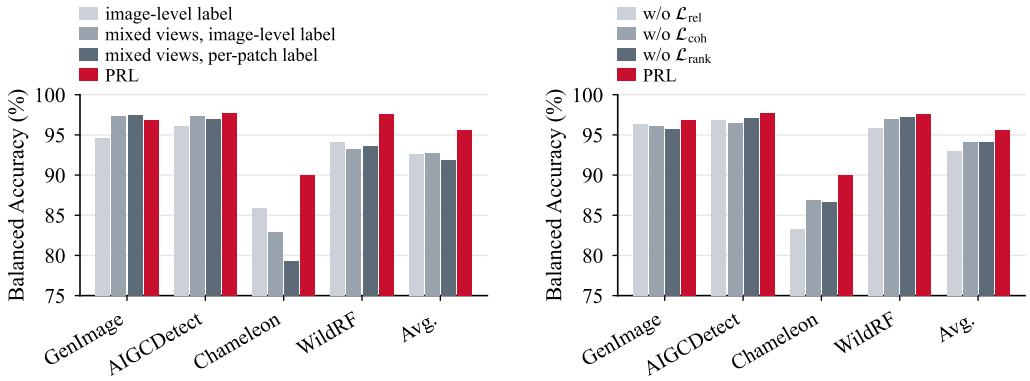
4.3 Comparison on Locally Manipulated Images

We additionally evaluate on two benchmarks of locally manipulated images, in which a region of a real image is replaced by generated content, AutoSplice (Jia et al., 2023) and the ManipulationBench subset of DailyBench (Jiang et al., 2026), both described in Appendix B.

Table 4 reports balanced accuracy on AutoSplice at JPEG quality 100, 90 and 75. PRL reaches 86.45% at quality 100 and remains the best at 90 and 75; GAPL is the strongest prior detector at 82.62%, DDA reaches 73.09%, and most of the remaining methods stay near chance. Results on the ManipulationBench subset of DailyBench are given in Appendix D, where PRL again has the best average, 68.32% against 66.88% for DDA.

4.4 Evaluation on Robustness

Figure 4 shows the results of three robustness evaluations on debiased GenImage-JPEG96 for all compared methods: JPEG compression at decreasing quality, resizing by a scale factor and Gaussian blur of increasing σ. PRL keeps the highest accuracy averaged over the fourteen degraded settings, 92.1% against 91.7% for GAPL and 90.6% for DDA. It is the best detector at every upscaling factor, staying above 95% up to a scale of 2 where GAPL falls to 88.2%, and it stays above 90% under blur up to $\sigma = 1.5$ and JPEG quality down to 80; at $\sigma = 2$ it remains within 1.6% of DDA while GAPL



(a) Form of supervision (b) Effect of each module

Figure 5: **Ablation.** Balanced accuracy (%) on GenImage, AIGCDetect, Chameleon and WildRF. (a) Three ways of using the aligned pairs: independent image-level label without and with the mixed views, the per-patch label on the mixed views, and the patch response. (b) PRL with one objective removed.

loses 15%. Under the two heaviest degradations, downscaling to half the size and JPEG quality below 80, it falls behind GAPL and DDA.

4.5 Ablation Studies

We conduct the ablation on four representative benchmarks: two standard benchmarks, GenImage and AIGCDetect, and two in-the-wild benchmarks, Chameleon and WildRF. Figure 5 presents the results and Table 12 in Appendix E lists the detailed numbers. Figure 5a compares the aligned pairs under independent image-level labels, the mixed views built from the pairs but still under the image-level label, a naive per-patch label on the mixed views, which **ignores the contextual shift**, and PRL. The mixed views by themselves change little, 92.68% against 92.65% on average, and lower Chameleon by 3.1%. The per-patch label gains slightly on GenImage but loses a further 3.6% on Chameleon. PRL raises the average to 95.55%, 2.9% above the image-level label and 3.7% above the per-patch label. The gain thus comes from how the mixed views are supervised, not from the views themselves, and a label per patch does worse in the wild than no patch label at all.

Figure 5b removes one objective at a time. Each removal costs between 1.4% and 2.5% on average and between 3.2% and 6.8% on Chameleon, $\mathcal{L}_{\text{rel}}$ the most. The sensitivity to the three loss weights is reported in Appendix E.

Besides, we adapt DDA, the best previous method, to our backbone: starting from its official code, we replace its DINOv2 backbone by DINOv3 with the same LoRA adaptation and keep the rest of its pipeline. Table 13 in Appendix E shows the results. The backbone lifts DDA from 89.16% to 93.20% on average, while PRL on the same backbone reaches 95.67%.

5 Conclusion

In this paper, we study patch-level supervision on aligned pairs for AI-generated image detection. Recent methods mix the two images of a pair and label each patch of the mixed view by its source. We find that patches of different sources in a mixed view interact through self-attention, so the feature of each patch also carries information passed from patches of the other source. A per-patch label ignores this contextual shift and is therefore an imprecise target. Relative Patch Response Learning (PRL) instead compares the same patch across two mixed views of an aligned pair and learns from the change of its score. Source-unchanged patches respond to the shift alone, so PRL measures the source-changed patches against them, keeps each group of them moving as a whole, and ranks the two views by generated area. PRL surpasses the best prior methods by 4.3% and 5.9% in average balanced accuracy on eight standard and three in-the-wild benchmarks, respectively.

AI USE STATEMENT

We used generative AI tools only to assist with writing and polishing the text of this paper and to search for related literature. We did not use them for any of the tasks for which the ICLR 2027 AI Policy for Authors requires disclosure, such as proposing hypotheses, designing experiments, implementing methods, generating or processing data, or interpreting results. The authors reviewed every AI-assisted passage and checked it against the method and the experimental results. We take full responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to reduce the harm of misinformation and visual forgery by improving the detection of AI-generated images. It uses only publicly available datasets under their respective licenses and involves no human subjects or personally identifiable information beyond what those datasets already contain. A detector can make mistakes, and its output should not be the sole basis for consequential decisions about an image.

REPRODUCIBILITY STATEMENT

Section 4.1 gives the backbone, the training pairs, the input resolution, the optimizer and the loss weights, and Appendix A the LoRA configuration, the two heads, the preprocessing and augmentation, and the sampling of the mixed views; Section 3 states every loss in closed form. Section 4.1 also specifies the evaluation protocol, including the JPEG compression applied to the synthetic images of GenImage, AIGCDetect, ForenSynths and UnivFD. Appendix B describes the benchmarks, and Appendix C the compared detectors, each run with its official checkpoint and inference code. Appendices D and E list the per-subset results, the ablation numbers and the sensitivity to the loss weights. Code, configuration files and the trained checkpoint will be released upon acceptance.

REFERENCES

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? Understanding properties that generalize. In *European Conference on Computer Vision (ECCV)*, 2020.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning (ICML)*, 2024.

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 (NeurIPS)*, 2025a.

Zhiyi Chen, Jinyi Ye, Beverlyn Tsai, Emilio Ferrara, and Luca Luceri. Synthetic politics: Prevalence, spreaders, and emotional reception of AI-generated political images on X. In *ACM Conference on Hypertext and Social Media (HT)*, 2025b.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Stefan Feuerriegel, Renée DiResta, Josh A. Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz, and Nicolas Pröllochs. Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, 7:1818–1821, 2023.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*, 2020.

Xinghe Fu, Zhiyuan Yan, Zheng Yang, Taiping Yao, Yandan Zhao, Shouhong Ding, and Xi Li. PiD: Generalized AI-generated images detection with pixelwise decomposition residuals. In *International Conference on Machine Learning (ICML)*, 2025.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or JPEG? Revealing common biases in generated image detection datasets. In *European Conference on Computer Vision Workshops (ECCVW)*, pp. 80–95, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. AutoSplice: A text-prompt manipulated image dataset for media forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

Xin Jiang, Hao Tang, Junyao Gao, Zijie Yang, Meiqi Cao, Fei Shen, Dongming Zhang, Zechao Li, and Yongdong Zhang. DailyBench: A unified benchmark for AI-generated and manipulated images from modern generative models. *arXiv preprint arXiv:2607.24016*, 2026.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18706–18717, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, 2024.

Kai Li, Wenqi Ren, Wei Wang, and Xiaochun Cao. Detecting compressed AI-generated images via phase spectrum robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

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

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 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17006–17015, 2024.

Sophie J. Nightingale and Hany Farid. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

OpenAI. Addendum to GPT-4o system card: Native image generation. `https://openai.com/index/gpt-4o-image-generation-system-card-addendum/`, 2025.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257, 2025.

Xinyi Qi, Kai Ye, Chengchun Shi, Ying Yang, Jin Zhu, and Hongyi Zhou. A difference-in-difference approach to detecting AI-generated images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up AI-generated image detection with generator-aware prototypes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations (ICLR)*, 2025.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pp. 5052–5060, 2024a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

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 *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Haiwei Wu, Fengpeng Li, Zhilin Tu, Yuanman Li, Xiong Li, and Jiantao Zhou. Zero-shot detection of AI-generated image via RAW-RGB alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

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 (ICLR)*, 2025a.

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 *International Conference on Machine Learning (ICML)*, 2025b.

Kai-Cheng Yang, Danishjeet Singh, and Filippo Menczer. Characteristics and prevalence of fake social media profiles with AI-generated faces. *Journal of Online Trust and Safety*, 2(4), 2024.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D^3: Scaling up deepfake detection by learning from discrepancy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23850–23859, 2025.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, and Xi Li. All patches matter, more patches better: Enhance AI-generated image detection via panoptic patch learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Xiao Yu, Kejiang Chen, Kai Zeng, Han Fang, Zijin Yang, Xiuwei Shang, Yuang Qi, Weiming Zhang, and Nenghai Yu. SemGIR: Semantic-guided image regeneration based method for AI-generated image detection and attribution. In *ACM International Conference on Multimedia (ACM MM)*, pp. 8480–8488, 2024.

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.

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 (NeurIPS), Datasets and Benchmarks Track*, 2023.

A Implementation Details

Architecture. The frozen DINOv3 ViT-L/16 backbone is adapted with LoRA of rank 8 and scaling $\alpha = 1$, applied in every transformer block to the query, key, value and output projections of self-attention and to the two projections of the MLP. The patch scoring head is a single linear layer on the final-layer patch tokens. The detection head is a gating MLP that assigns a weight to every token, aggregates the tokens with the normalized weights and maps the result to the image logit.

Data augmentation. At both training and inference, we center-crop every image to 336×336 pixels, and we resize an image to the input size instead when both its width and height exceed $1{,}024$ pixels, so that the input covers the whole image. We apply the common augmentations in this field, such as random JPEG compression, Gaussian blur and colour jitter. Each augmentation is drawn once per pair and applied identically to both members, so that the alignment is preserved.

View sampling. For each training pair, with probability 0.5 we use the real image with two mixed views, and otherwise the real image with the synthetic image. Each mask replaces one axis-aligned rectangle of patches on the 21×21 patch grid of the encoder, so $P = 441$. The two masks of a pair share a base image, the real or the synthetic one with equal probability, and the rectangle takes its patches from the other image. For each mask, we draw the synthetic fraction uniformly from $[0.1, 0.9]$ and the aspect ratio of the rectangle log-uniformly from $[0.5, 2]$, resample until the two fractions differ, and place the rectangle uniformly at random among the positions where it fits.

B Benchmarks

GenImage (Zhu et al., 2023) pairs ImageNet photographs with images from eight generators: ADM, BigGAN, Glide, Midjourney, Stable Diffusion v1.4 and v1.5, VQDM and Wukong.

AIGCDetectBenchmark (Zhong et al., 2023) collects seventeen subsets from GANs (ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN and the WFIR face set) and diffusion models (ADM, Glide, Midjourney, Stable Diffusion v1.4, v1.5 and XL, VQDM, Wukong and DALL-E 2), with real images drawn from LSUN, ImageNet, MSCOCO, CelebA and FFHQ.

DRCT-2M (Chen et al., 2024) generates images from MSCOCO with sixteen diffusion models. Ten text-to-image models (LDM, SD1.4, SD1.5, SD2.1, SDXL, SDXL-Refiner, SD-Turbo, SDXL-Turbo, LCM-SD1.5 and LCM-SDXL) are prompted by the MSCOCO captions, three ControlNet models are conditioned on the captions and Canny edges, and three inpainting models reconstruct the real images with an empty prompt and an all-zero mask (SDv1-DR, SDv2-DR and SDXL-DR). Every subset is paired with the shared real set.

DDA-COCO and EvalGEN (Chen et al., 2025a) are released with DDA. DDA-COCO reconstructs the MSCOCO validation images with six VAEs, including those of SD1.x, SD2.1, SDXL, SD3.5 and Flux, with frequency-level alignment to the real images. EvalGEN holds $2{,}765$ images from five recent text-to-image generators (Flux, GoT, Infinity, OmniGen and NOVA) driven by GenEval prompts; it has no real counterpart, so we report accuracy on its synthetic images.

Synthbuster (Bammey, 2024) provides $1{,}000$ images from each of nine generators (DALL-E 2, DALL-E 3, Firefly, Glide, Midjourney v5, Stable Diffusion 1.3, 1.4, 2 and XL).

ForenSynths (Wang et al., 2020) is the test set of CNNSpot, with thirteen subsets from CNN-based generators: ProGAN, StyleGAN, StyleGAN2, BigGAN, CycleGAN, StarGAN, GauGAN, CRN, IMLE, SITD, SAN, DeepFake and WFIR.

UnivFD (Ojha et al., 2023) extends the eleven ForenSynths subsets with eight diffusion and autoregressive subsets: Guided, three LDM variants, three Glide variants and DALL-E, nineteen subsets in total.

Chameleon (Yan et al., 2025a) collects about $26{,}000$ real and generated images from the web. The generated images come from AI-art communities and passed a human perception test, every image is at least 448×448 pixels, and the generators are unknown.

SynthWildX (Cozzolino et al., 2024) holds 500 real and $1{,}500$ generated images collected from X, attributed by the tags and annotations on the platform to DALL-E 3, Midjourney and Firefly, which form its three subsets.

WildRF (Cavia et al., 2024) collects real and generated images from Reddit, Facebook and X, one subset per platform, with unknown generators and varying resolutions, formats, compressions and editing transformations.

AutoSplice (Jia et al., 2023) replaces one object region of a real photograph by a DALL-E 2 inpainting guided by a text prompt, and releases the forged images at JPEG qualities 100, 90 and 75.

DailyBench (Jiang et al., 2026) collects edited versions of real images from LAION-Aesthetics; its ManipulationBench subset, which we use, contains edits by four open-source editors (FLUX-Fill with object masks or random rectangular masks, FLUX.2-klein-9B, Qwen-Image-Edit-2511 and Step1X-Edit-v1p2) and two proprietary ones (Nano Banana 2 and GPT-Image 2), mostly object removals and replacements.

C Compared Methods

UnivFD (Ojha et al., 2023) keeps the CLIP ViT-L/14 image encoder frozen and trains only a linear classifier on its features, arguing that a feature space not tuned to the training generator generalizes better.

FatFormer (Liu et al., 2024) adds forgery-aware adapters with a frequency branch to the CLIP image encoder and aligns the adapted image features with text prompts.

NPR (Tan et al., 2024b) models the relations between neighbouring pixels that up-sampling layers of generators leave behind and feeds them to a ResNet classifier.

DRCT (Chen et al., 2024) reconstructs both real and generated images with a diffusion model to obtain hard samples of the same content, labels every reconstruction as generated, and trains a detector with a classification and a contrastive loss.

C2P-CLIP (Tan et al., 2025) injects category-common prompts into the CLIP image encoder through the text encoder, so that the image features carry the concepts of real and generated images.

AIDE (Yan et al., 2025a) combines the highest- and lowest-frequency patches of an image, selected by DCT scoring and processed by a CNN, with the semantic features of CLIP; the paper also introduces the Chameleon benchmark.

AlignedForensics (Rajan et al., 2025) pairs every real image with its VAE reconstruction so that the two differ only in the traces of generation, and trains a detector on these aligned images.

Effort (Yan et al., 2025b) decomposes the weights of the CLIP encoder into orthogonal subspaces by singular value decomposition, freezes the principal subspace to preserve the pretrained knowledge and trains the residual subspace for detection.

SAFE (Li et al., 2025) feeds the high-frequency components of a discrete wavelet transform to a lightweight ResNet and pairs it with image transformations that preserve the traces of generation: cropping instead of resizing, colour jitter, random rotation and random masking.

DDA (Chen et al., 2025a) aligns generated images with real ones in both the pixel and the frequency domain: VAE reconstruction, high-frequency debias from the real image and pixel mixup.

Table 5: Comparison of balanced accuracy (%) on GenImage.

Method	ADM	BigGAN	Glide	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	**Avg.**
UnivFD CVPR'23	62.51	84.37	61.28	55.06	55.57	55.67	76.88	61.10	64.05
FatFormer CVPR'24	59.82	81.77	60.28	51.99	52.81	52.76	69.97	58.86	61.03
NPR CVPR'24	46.73	44.25	46.06	43.86	46.73	47.06	50.00	49.22	46.74
DRCT ICML'24	60.74	58.38	67.86	88.68	99.37	99.21	70.03	99.10	80.42
C2P-CLIP AAAI'25	53.62	92.67	65.91	50.60	55.02	55.31	63.58	60.00	62.09
AIDE ICLR'25	50.22	50.42	52.26	57.46	75.96	76.19	50.78	69.47	60.35
AlignedForensics ICLR'25	50.49	52.73	52.80	96.33	**99.82**	**99.74**	67.83	**99.79**	77.44
Effort ICML'25	69.75	51.60	83.43	76.37	96.03	95.82	73.24	90.17	79.55
SAFE KDD'25	50.41	51.84	53.73	49.89	50.55	50.54	51.01	51.15	51.14
DDA NeurIPS'25	89.55	86.52	89.58	95.60	98.72	98.55	76.58	98.78	91.74
GAPL CVPR'26	**94.46**	**97.17**	**97.52**	84.41	97.81	97.59	97.69	97.81	95.56
PRL (Ours)	94.40	91.55	95.02	**96.73**	99.33	99.25	**99.05**	99.36	**96.83**

Table 6: Comparison of balanced accuracy (%) on AIGCDetectBenchmark.

Method	ADM	BigGAN	CycleGAN	DALLE2	GauGAN	Glide	Midjourney	ProGAN	SDXL	SDv1.4	SDv1.5	StarGAN	StyleGAN	StyleGAN2	VQDM	WFIR	Wukong	**Avg.**
UnivFD CVPR'23	62.51	87.50	96.92	49.95	98.79	61.28	55.06	99.38	58.15	55.57	55.67	95.12	79.99	69.41	76.88	69.20	61.10	72.50
FatFormer CVPR'24	59.82	97.90	99.32	49.00	99.17	60.28	51.99	99.21	63.75	52.81	52.76	**99.75**	87.51	81.63	69.97	87.50	58.86	74.78
NPR CVPR'24	46.73	43.25	59.83	46.65	35.62	46.06	43.86	50.45	47.48	46.73	47.06	49.47	50.18	50.02	50.00	49.65	49.22	47.78
DRCT ICML'24	60.74	52.45	48.83	68.25	49.45	67.86	88.68	53.02	80.65	99.37	99.21	54.38	54.45	53.74	70.03	50.15	99.10	67.67
C2P-CLIP AAAI'25	53.62	92.07	**99.39**	50.20	98.20	65.91	50.60	97.10	51.20	55.02	55.31	99.70	76.97	58.51	63.58	65.35	60.00	70.16
AIDE ICLR'25	50.22	43.68	51.73	50.95	39.73	52.26	57.46	49.31	50.55	75.96	76.19	51.43	49.57	49.20	50.78	49.20	69.47	53.98
AlignedForensics ICLR'25	50.49	50.52	49.69	50.30	50.24	52.80	96.33	50.35	94.45	**99.82**	**99.74**	51.50	51.46	50.31	67.83	50.00	**99.79**	65.63
Effort ICML'25	69.75	73.60	81.49	79.15	62.52	83.43	76.37	70.08	86.88	96.03	95.82	96.02	56.42	47.67	73.24	70.35	90.17	77.00
SAFE KDD'25	50.41	48.95	52.58	50.05	42.60	53.73	49.89	49.96	49.93	50.55	50.54	49.95	49.99	49.99	51.01	49.10	51.15	50.02
DDA NeurIPS'25	89.55	90.97	72.47	94.55	92.74	89.58	95.60	92.80	99.38	98.72	98.55	72.74	87.80	90.17	76.58	52.10	98.78	87.83
GAPL CVPR'26	**94.46**	98.62	97.54	84.05	**99.63**	**97.52**	84.41	**99.83**	98.78	97.81	97.59	97.05	97.89	**98.62**	97.69	**94.10**	97.81	96.08
PRL (Ours)	94.40	**99.28**	97.37	**96.55**	99.55	95.02	**96.73**	98.99	**99.55**	99.33	99.25	99.25	**98.21**	98.19	**99.05**	91.65	99.36	**97.75**

GAPL (Qin et al., 2026) builds generator-aware prototypes from the principal components of the embeddings of a few generators and fine-tunes the CLIP ViT-L/14 encoder with LoRA so that image embeddings are mapped onto these prototypes. It is trained on the Community Forensics set (Park & Owens, 2025) of several thousand generators.

D Detailed Benchmark Results

Tables 5, 6, 7, 8, 9 and 10 report the per-subset results on GenImage, AIGCDetect, Synthbuster, DRCT-2M, DDA-COCO and EvalGEN, and ForenSynths, which Table 1 and Section 4.2 summarize. PRL is above 98.8% on every DRCT-2M subset and above 99% on every EvalGEN generator; on DDA-COCO its two lowest subsets are the SD3.5 and Flux reconstructions. On ForenSynths, its margin over the alignment-based methods opens on every GAN generator, while the DeepFake, SITD and SAN subsets remain its hardest.

Table 11 reports balanced accuracy on the ManipulationBench subset of DailyBench (Section 4.3), with the rows of the other detectors, including the detector FPD proposed with the benchmark, as published by DailyBench. Most of the detectors it evaluates stay within a few points of chance on most editors. PRL has the best average, 68.32% against 66.88% for DDA and 63.41% for FPD, and the best result on the three FLUX-based editors, where every other detector except DDA is at chance. DDA is stronger on Qwen-Image-Edit and Step1X-Edit, and SAFE on GPT-Image 2.

Table 7: Comparison of balanced accuracy (%) on Synthbuster.

Method	DALL-E 2	DALL-E 3	Firefly	GLIDE	Midjourney	SD1.3	SD1.4	SD2	SDXL	**Avg.**
UnivFD CVPR'23	70.98	34.83	77.38	40.73	39.93	57.88	57.38	63.13	55.48	55.31
FatFormer CVPR'24	53.53	35.33	59.83	67.28	43.88	52.43	53.18	51.03	66.08	53.62
NPR CVPR'24	55.76	12.36	9.46	53.01	40.81	55.76	56.41	35.86	57.11	41.83
DRCT ICML'24	47.54	53.39	48.59	60.19	89.29	94.09	94.09	90.79	79.39	73.04
C2P-CLIP AAAI'25	54.89	43.64	81.54	81.59	52.74	79.69	79.94	65.64	72.99	68.08
AIDE ICLR'25	37.83	36.63	27.73	67.93	60.48	77.03	76.68	56.18	71.38	56.87
AlignedForensics ICLR'25	49.15	48.60	50.00	56.60	**98.55**	**98.60**	**98.60**	**98.30**	97.25	77.29
Effort ICML'25	67.88	79.28	56.08	78.98	77.23	81.63	81.68	77.88	80.48	75.68
SAFE KDD'25	64.27	16.22	16.57	58.47	63.02	65.72	65.42	59.32	65.77	52.75
DDA NeurIPS'25	81.79	85.44	87.39	71.94	88.94	88.34	88.14	88.79	88.94	85.52
GAPL CVPR'26	91.69	74.39	91.54	93.29	90.64	93.54	93.59	92.64	94.29	90.63
PRL (Ours)	**97.65**	**97.85**	**93.40**	**94.85**	97.85	97.85	97.85	97.85	**97.85**	**97.00**

Table 8: Comparison of balanced accuracy (%) on DRCT-2M.

Method	SDXL-Ref	SDXL-Ctrl	LCM-SDv1.5	LCM-SDXL	LDM	SD-Ctrl	SD-Turbo	SD2.1-Ctrl	SDXL-Turbo	SD2.1	SDv2-DR	SDv1-DR	SDv1.4	SDv1.5	SDXL-DR	SDXL	**Avg.**
UnivFD CVPR'23	54.30	72.80	53.48	65.20	81.27	66.26	55.25	63.97	52.42	56.08	52.56	56.47	55.16	54.78	52.18	62.04	59.64
FatFormer CVPR'24	48.60	61.06	48.61	49.97	52.93	49.97	48.59	50.20	48.54	48.56	53.71	54.98	48.55	48.55	51.78	48.55	50.82
NPR CVPR'24	47.72	57.20	49.16	47.72	48.85	48.11	48.28	48.57	48.07	48.50	50.78	52.60	48.84	48.77	53.03	47.76	49.62
DRCT ICML'24	85.86	80.19	99.65	78.73	99.90	99.89	91.94	95.04	70.54	96.42	92.63	99.88	99.90	99.88	72.27	83.90	90.41
C2P-CLIP AAAI'25	50.92	72.09	50.67	53.56	63.61	57.02	51.07	54.28	50.44	50.95	50.91	55.89	51.88	51.64	50.68	52.98	54.29
AIDE ICLR'25	66.05	53.42	69.95	54.13	60.74	64.80	53.03	52.95	53.07	58.09	50.02	54.52	74.82	74.28	50.01	53.11	58.94
AlignedForensics ICLR'25	80.49	84.08	**99.85**	89.17	**99.95**	**99.98**	**99.88**	99.30	88.15	**99.77**	**99.88**	**99.98**	**99.97**	**99.98**	92.68	88.90	95.13
Effort ICML'25	64.81	65.94	74.96	69.65	49.99	68.00	64.20	62.27	58.91	84.89	54.37	64.80	85.73	85.66	54.10	70.49	67.42
SAFE KDD'25	50.11	55.28	50.13	49.94	50.22	49.96	49.97	50.04	49.95	49.97	50.35	50.22	50.01	50.02	50.67	49.92	50.42
DDA NeurIPS'25	97.47	99.45	96.98	98.60	99.32	98.96	98.31	99.24	95.53	98.54	99.47	99.18	99.06	99.06	98.11	98.18	98.47
GAPL CVPR'26	96.57	95.55	98.44	94.99	99.21	98.27	95.48	96.33	93.61	97.07	90.29	93.34	97.94	97.88	84.34	94.33	95.23
PRL (Ours)	**99.21**	**99.48**	99.38	**99.40**	99.41	99.52	99.17	**99.44**	**98.82**	99.15	99.45	99.48	99.15	99.10	**98.98**	**99.16**	**99.27**

Table 9: Comparison of balanced accuracy (%) on DDA-COCO and of accuracy (%) on the synthetic images of EvalGEN.

Method	DDA-COCO							EvalGEN					
	VAE-ema	VAE-mse	SDXL	SD2.1	SD3.5	Flux	**Avg.**	Flux	GoT	Infinity	NOVA	OmniGen	**Avg.**
UnivFD CVPR'23	54.40	53.24	51.52	53.29	51.28	50.21	52.32	3.94	10.27	15.30	39.72	7.71	15.39
FatFormer CVPR'24	51.49	53.17	50.77	53.31	51.63	49.42	51.63	1.59	16.14	19.03	91.85	7.22	27.17
NPR CVPR'24	49.39	48.88	48.30	48.85	48.52	50.46	49.07	2.04	1.10	6.01	2.42	5.57	3.43
DRCT ICML'24	83.31	76.99	62.49	76.82	51.47	50.24	66.89	36.26	49.59	63.40	49.68	55.70	50.93
C2P-CLIP AAAI'25	51.42	51.27	50.22	51.23	51.28	49.98	50.90	1.78	6.71	11.39	46.57	2.66	13.82
AIDE ICLR'25	50.14	50.12	50.10	50.09	50.03	50.04	50.09	14.28	18.93	3.20	12.65	16.38	13.09
AlignedForensics ICLR'25	**99.80**	**99.69**	91.98	**99.67**	75.61	51.16	86.32	20.96	64.47	71.79	78.48	71.39	61.42
Effort ICML'25	52.63	52.34	52.14	52.40	51.67	50.09	51.88	85.56	99.17	92.98	95.67	98.66	94.41
SAFE KDD'25	50.35	50.39	50.21	50.42	50.07	50.25	50.28	1.00	0.48	1.88	0.59	1.56	1.10
DDA NeurIPS'25	99.15	99.35	96.99	99.36	83.56	78.92	92.89	88.68	99.14	97.73	99.53	99.20	96.86
GAPL CVPR'26	89.60	88.93	78.36	88.90	70.95	58.85	79.27	90.76	97.46	98.39	97.59	96.31	96.10
PRL (Ours)	99.28	99.27	**98.13**	99.27	**85.17**	**85.19**	**94.39**	**96.84**	**99.64**	**99.98**	**99.60**	**99.61**	**99.14**

Table 10: Comparison of balanced accuracy (%) on ForenSynths.

Method	BigGAN	CRN	CycleGAN	DeepFake	GauGAN	IMLE	ProGAN	SAN	SITD	StarGAN	StyleGAN	StyleGAN2	WFIR	**Avg.**
UnivFD CVPR'23	87.50	55.67	96.92	69.39	98.79	68.11	99.38	58.22	62.22	95.12	79.99	69.41	69.20	77.69
FatFormer CVPR'24	97.90	69.32	99.32	86.03	99.17	69.45	99.21	62.10	80.83	**99.75**	87.51	81.63	87.50	86.13
NPR CVPR'24	43.25	0.00	59.83	33.30	35.62	0.00	50.45	48.40	22.22	49.47	50.18	50.02	49.65	37.88
DRCT ICML'24	52.45	49.80	48.83	56.38	49.45	49.26	53.02	86.76	54.44	54.38	54.45	53.74	50.15	54.86
C2P-CLIP AAAI'25	92.07	92.75	**99.39**	**91.85**	98.20	**97.68**	97.10	74.20	**87.22**	99.70	76.97	58.51	65.35	87.00
AIDE ICLR'25	43.68	19.96	51.73	45.54	39.73	19.38	49.31	51.37	30.56	51.43	49.57	49.20	49.20	42.36
AlignedForensics ICLR'25	50.52	49.95	49.69	52.69	50.24	49.91	50.35	61.87	50.56	51.50	51.46	50.31	50.00	51.47
Effort ICML'25	73.60	39.20	81.49	64.59	62.52	29.54	70.08	58.90	83.89	96.02	56.42	47.67	70.35	64.17
SAFE KDD'25	48.95	0.11	52.58	49.09	42.60	0.10	49.96	50.46	35.83	49.95	49.99	49.99	49.10	40.67
DDA NeurIPS'25	90.97	86.98	72.47	76.53	92.74	89.71	92.80	94.75	58.61	72.74	87.80	90.17	52.10	81.41
GAPL CVPR'26	98.62	67.85	97.54	84.88	**99.63**	67.12	**99.83**	**96.58**	82.50	97.05	97.89	**98.62**	**94.10**	90.94
PRL (Ours)	**99.28**	**96.07**	97.37	55.26	99.55	96.08	98.99	88.81	84.72	99.25	**98.21**	98.19	91.65	**92.57**

Table 11: Comparison of balanced accuracy (%) on the ManipulationBench subset of DailyBench; rows other than ours are as reported by DailyBench, including its own detector FPD.

Method	FLUX-Fill (rand.)	FLUX-Fill (obj.)	FLUX.2 -klein	Qwen-Edit	Step1X-Edit	Nano Banana 2	GPT-Image 2	**Avg.**
UnivFD CVPR'23	50.16	50.83	50.72	50.39	50.72	50.26	50.23	50.47
FatFormer CVPR'24	49.86	49.95	50.11	50.05	50.26	50.00	56.96	51.02
NPR CVPR'24	45.09	46.02	49.81	55.17	48.77	55.53	80.33	54.38
DRCT ICML'24	51.00	50.86	50.56	52.17	50.13	57.21	63.99	53.70
AIDE ICLR'25	51.13	50.65	49.63	58.63	57.54	54.43	60.43	54.63
AlignedForensics ICLR'25	50.51	50.44	50.42	73.18	50.43	**64.14**	56.96	56.58
Effort ICML'25	45.60	45.70	50.67	56.98	53.52	60.85	68.99	54.61
SAFE KDD'25	52.22	51.94	49.99	48.96	48.26	49.46	**86.67**	55.35
DDA NeurIPS'25	54.54	54.82	62.29	**89.20**	**62.88**	62.48	81.96	66.88
FPD arXiv'26	53.68	54.09	58.00	80.47	60.70	53.22	83.71	63.41
PRL (Ours)	**62.58**	**65.89**	**64.86**	81.18	58.12	59.94	85.69	**68.32**

E More Ablation Studies

Full ablation results. Table 12 lists the numbers behind Figure 5.

Table 12: **Ablation.** Balanced accuracy (%) of the variants of Figure 5 on GenImage, AIGCDetect, Chameleon and WildRF. All rows use the same DINOv3 backbone with LoRA and the same aligned training pairs. Top: three ways of using the aligned pairs; bottom: PRL with one objective removed.

Method	GenImage	AIGCDetect	Chameleon	WildRF	**Avg.**
Image-level BCE on the aligned pairs	94.58	96.06	85.89	94.06	92.65
+ mixed views, image-level BCE	97.33	97.38	82.84	93.17	92.68
+ mixed views, per-patch label	**97.45**	96.92	79.29	93.62	91.82
PRL (Ours)	96.83	**97.75**	**90.01**	**97.59**	**95.55**
w/o $\mathcal{L}_{\text{rel}}$	96.32	96.77	83.20	95.82	93.03
w/o $\mathcal{L}_{\text{coh}}$	96.05	96.43	86.83	96.89	94.05
w/o $\mathcal{L}_{\text{rank}}$	95.72	97.05	86.61	97.20	94.14

DDA on the same backbone. Table 13 compares DDA with its released DINOv2 backbone, DDA adapted to our DINOv3 backbone with the same LoRA, and PRL on the eleven benchmarks (Section 4.5).

Table 13: **DDA and PRL on the same backbone.** Balanced accuracy (%) on the eleven benchmarks of Table 1 for the released DDA model (DINOv2), for DDA adapted from its official code to our DINOv3 backbone with LoRA, and for PRL, which uses the same backbone, adaptation and aligned pairs.

Method	Standard Benchmarks								In-the-Wild Benchmarks			
	GenImage	AIGCDetect	DRCT-2M	DDA-COCO	EvalGEN	Synthbuster	ForenSynths	UnivFD	Chameleon	SynthWildX	WildRF	**Avg.**
DDA, released model (DINOv2)	91.74	87.83	98.47	92.89	96.86	85.52	81.41	82.58	82.41	90.87	90.20	89.16
DDA on DINOv3	**97.93**	96.09	99.04	**95.98**	**99.40**	88.77	87.50	85.52	**91.82**	89.01	94.17	93.20
PRL (Ours)	96.83	**97.75**	**99.27**	94.39	99.14	**97.00**	**92.57**	**94.22**	90.01	**93.60**	**97.59**	**95.67**

Sensitivity to the loss weights. Figure 6 varies one loss weight at a time on the same four benchmarks, with the other two kept at their default values $(\lambda_{\text{rel}}, \lambda_{\text{coh}}, \lambda_{\text{rank}}) = (0.3, 0.3, 0.1)$. Setting a weight to zero reproduces the corresponding row of Table 12. PRL is robust to the choice of loss weights. The default is the best point on every axis, every nonzero weight in $[0.1, 0.5]$ stays within 1.5% of it, and every setting remains above the image-level label.

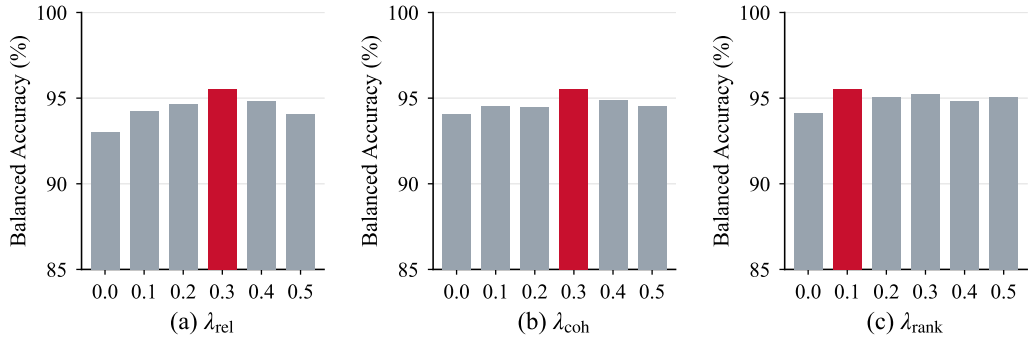


Figure 6: **Sensitivity to the loss weights.** Mean balanced accuracy (%) over GenImage, AIGCDetect, Chameleon and WildRF when one weight varies and the other two keep their default values; the red bar is the default setting.