Revisiting Deepfake Detection: BCNet for Robust Generalization Beyond Semantic Dependence

Jian Yang*, Shibo Yao*, Renshuai Tao✉, Chuangchuang Tan, and Yao Zhao

Institute of Information Science, Beijing Jiaotong University, China

Abstract. Recent deepfake detection methods leveraging vision foundation models (VFMs) like CLIP have made significant progress. However, the abundant semantic information in large training datasets has made VFMs highly dependent on semantics. As a result, VFM-based methods perform well on images from the same category as the training set but struggle with others. This bias limits generalization in real-world scenarios. To address this issue, we propose the Basis Correction Network (BCNet), which consists of two modules: the attention-guided semantic erasure (ASE), which adaptively identifies and erases semantic regions of the image by capturing the model's semantic attention, and the normalized-gradient perturbation enhancement (NPE) scales the gradients of fake samples to concentrated values and adds them as a small perturbation to the original sample, helping the model recognize more forgery patterns and improving its ability to distinguish fake from real samples. This design ensures the model focuses on the core distinction between real and fake categories rather than semantic information. Extensive experiments on 51 AI-generated datasets show that BCNet achieves 96.7% generalization accuracy on WildRF (collected from real social media) and outperforms other competitors by 8.7% on AIGI-Bench, offering a fresh perspective on semantic generalization in deepfake detection.The code is open-sourced and publicly available at `https://github.com/zwhyyyy/BCNet`.

Keywords: Deepfake Detection · AI-Generated Images Detection · AI Security

1 Introduction

The rapid development of generative AI [25, 32, 55, 56] has enabled the creation of highly realistic synthetic images, or deepfakes, which are often difficult to distinguish from real content [36]. While these technologies offer new opportunities for creativity, they also pose serious risks, including misinformation, identity fraud, and manipulation of public opinion [1,2]. As generative tools become more accessible, detecting AI-generated images has become a critical problem in the

*Equal Contribution.
✉Corresponding author. Email: `rstao@bjtu.edu.cn`

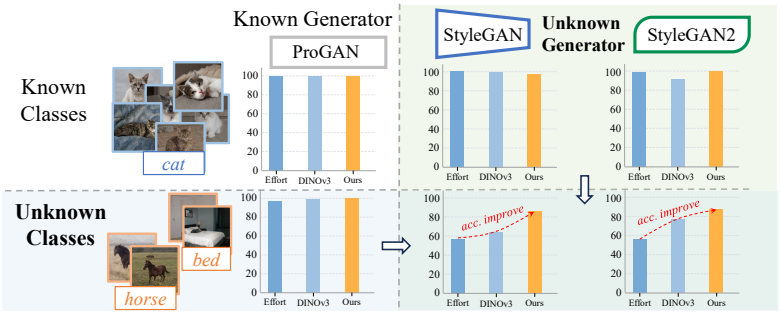

Fig. 1: Semantic Bias in VFM-based Detection. Detectors trained exclusively on **Cat** images are evaluated on both seen and unseen semantic classes. Baseline methods (Effort [78], LoRA-tuned [24] DINOv3 [64]) suffer a severe performance drop on novel semantics, revealing a heavy reliance on semantic cues rather than actual forgery patterns. In contrast, our **BCNet** corrects the decision basis to focus on intrinsic synthesis traces, achieving robust semantic generalization across diverse categories.

computer vision community, requiring highly robust and scalable methods that generalize across diverse and unforeseen scenarios [39, 77, 78].

Recent detection methods have increasingly leveraged vision foundation models (VFMs), such as CLIP [58], which are trained on massive, large-scale datasets. The richness of semantic information in these datasets allows VFMs to capture complex patterns and relationships within images, enabling strong performance on tasks like image classification and retrieval [9, 33, 52, 64, 80]. However, this abundance of semantic knowledge also introduces a subtle but critical limitation: VFMs become highly dependent on semantics when applied to deepfake detection [66]. While these methods can accurately identify manipulated content in images that resemble those in the training data, their strict reliance on these semantic cues becomes a clear vulnerability when encountering images from new categories not seen during the training phase.

This raises an important question regarding **whether these models are truly learning to distinguish real from fake, or merely recognizing familiar patterns tied to specific semantic categories [50,77].** To investigate this, we conducted a pre-experiment (Figure 1) in which an advanced model was trained solely on the "cat" category [12, 73]. The results show that while the model performs well on "cat" images regardless of the generative method, its performance drops significantly on other categories. Such semantic bias severely limits the generalization of existing detectors in real-world scenarios, where AI-generated images can appear across a wide range of unseen and diverse semantic contexts [21]. Consequently, current models that rely heavily on these specific semantic cues risk misclassifying manipulations in completely novel categories, thereby reducing their practical utility and overall robustness [7, 17].

To address this challenge, motivated by the goal of reducing reliance on semantics and refocusing on the core real/fake distinction, we propose the Basis

Correction Network (BCNet), a new framework for semantic-agnostic deepfake detection. BCNet comprises two complementary modules: attention-guided semantic erasure (ASE), which identifies semantic regions by analyzing the model's attention maps and selectively erases them, thereby forcing the model to rely on manipulation cues rather than semantics [48], and normalized-gradient perturbation enhancement (NPE), which scales the gradients of fake samples and adds them as small perturbations to the original images, helping the model recognize more diverse and subtle forgery patterns and ultimately improving its overall ability to accurately distinguish fake from real content [19, 23, 41].

We conduct extensive evaluations of our framework on 51 publicly available AI-generated datasets [5, 38, 72, 77, 83]. The experimental results show that our method consistently outperforms existing approaches, achieving 96.7% generalization accuracy on WildRF [5], a dataset collected from real social media, and exceeding prior methods by 8.7% on AIGI-Bench [38], demonstrating its superior robustness and strong semantic-agnostic capabilities across various complex scenarios. The main contributions are summarized as follows:

- We show that VFM-based detectors rely heavily on semantic cues, which limits its generalization to unseen categories. This finding offers a new perspective and motivates the development of semantic-agnostic detection methods.
- We propose BCNet, which combines ASE to remove category-specific semantic regions and NPE to recognize more forgery patterns, guiding the model to focus on the core real/fake distinction rather than semantic artifacts.
- Extensive experiments on 51 AI-generated datasets show that BCNet achieves 96.7% accuracy on WildRF [5] and outperforms existing methods by 8.7% on AIGI-Bench [38], demonstrating its robustness, broad applicability, and semantic-agnostic capabilities in real-world deepfake scenarios.

2 Related Works

2.1 CNN-based Synthetic Image Detection

Early synthetic image detection predominantly leveraged CNNs to isolate spatial or frequency artifacts. CNNSpot [71] achieved baseline generalization via meticulous augmentation. Subsequently, methods targeted frequency-aware cues, with F^3-Net [57] using adaptive decomposition and FreqNet [67] employing specialized convolutional layers. Structural inconsistencies were concurrently explored: NPR [65] targets up-sampling flaws, and FerretNet [39] exposes texture discontinuities via zero-masked reconstruction. To combat severe overfitting, SAFE [35] integrates random masking. However, these CNN methods struggle against modern high-fidelity models, which effectively suppress low-level traces and seamlessly reconstruct high-frequency details in generated images.

2.2 Adapting Vision Foundation Models for Detection

The emergence of Vision Foundation Models (VFMs) has shifted the paradigm toward leveraging universal visual representations. UnivFD [50] showed that

frozen CLIP [58] backbones achieve superior generalization, inspiring architectures like FatFormer [40], which blends spatial-frequency clues via parameter-efficient adapters. To further refine features, recent works have explored complex feature extraction methods: VIB-Net [82] utilizes information bottlenecks, Effort [78] applies SVD decoupling, AIDE [77] leverages mixture-of-experts, and DDA [8] aligns domains via VAE [30] reconstruction. Despite these advancements, existing methods remain susceptible to inherent semantic bias. As revealed by C2P-CLIP [66], models like CLIP naturally tend to rely on semantic content rather than authenticity traces. Unlike these approaches, our **BCNet** framework explicitly corrects the decision basis to move beyond semantic dependence. By integrating attention-guided semantic erasure and normalized-gradient perturbation enhancement, BCNet actively exposes underlying forgery patterns, ensuring the model focuses on the core distinction between real and fake categories rather than dataset-specific semantics, thereby achieving highly reliable and robust cross-domain generalization in practical applications.

3 Methodology

3.1 Overview

We propose **BCNet** (Basis Correction Network) to correct the semantic bias of a frozen **VFM** [64]. As revealed by our motivating experiment (Figure 1), VFMs often anchor judgments on specific semantic categories, leading to a biased decision basis. As illustrated in Figure 2, we use **Low-Rank Adaptation (LoRA)** [24] to efficiently adapt the pre-trained features. To shift the model's focus from misleading semantic information to the core differences between real and fake categories, BCNet uses two correction modules. First, **Attention-Guided Semantic Erasure (ASE)** adaptively identifies and masks prominent semantic regions. By removing these regions, ASE forces the model to correct its judgment basis and mine intrinsic synthesis traces. Second, **Normalized-Gradient Perturbation Enhancement (NPE)** scales the gradients of fake samples to concentrated values and adds them as small perturbations to the original images. This actively exposes underlying forgery patterns, improving the model's ability to distinguish fake from real samples. By forcing the network to focus on the fundamental categorical distinction rather than memorizing dataset-specific semantics, BCNet effectively widens the decision margin. This ensures strong generalization to unseen, high-fidelity generative domains.

3.2 Attention-Guided Semantic Erasure (ASE)

Vision foundation models (VFMs) tend to rely heavily on high-level semantic information [66]. This reliance creates a major generalization bottleneck where the detector uses semantic content as the basis for its predictions rather than intrinsic synthesis traces. To rectify this bias, we propose the ASE module to adaptively eliminate semantic dependencies and compel the model to establish a more fundamental forensic basis for robust deepfake detection.

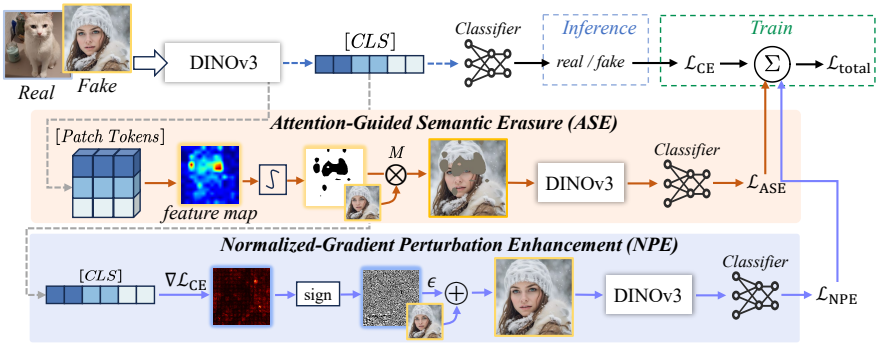

Fig. 2: The overall framework of the Basis Correction Network (BCNet). Built upon a frozen DINOv3 backbone with parameter-efficient LoRA, BCNet utilizes a complementary correction strategy to eliminate semantic dependence. The entire framework is optimized end-to-end, asymmetrically aggregating the baseline loss with the correction losses ($\mathcal{L}_{ASE}$ and $\mathcal{L}_{NPE}$) applied exclusively to fake samples.

During the training phase, we extract self-attention maps from the final Transformer [68] block. Let S denote the attention map upsampled to match the input image size. A binary correction mask M is defined as:

$$M_{i,j} = \mathbb{I}(S_{i,j} < \tau) \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ is the semantic threshold. By assigning 0 to regions with high semantic attention ($S_{i,j} \geq \tau$), the module physically erases the primary semantic basis, forcing the network to seek a new, robust judgment basis in the remaining background and boundary regions (assigned as 1) [10,84]. This mask is applied to the input fake image X_f to generate a semantically-corrected view $X'_f = X_f \odot M$. The corresponding ASE loss is formulated as:

$$\mathcal{L}_{ASE} = \mathcal{L}_{CE}(f_\theta(X'_f), y) \tag{2}$$

By erasing semantic cues from synthetic inputs, ASE forces the LoRA layers to prioritize core synthesis traces that stay consistent across different categories, effectively correcting the model's decision basis.

3.3 Normalized-Gradient Perturbation Enhancement (NPE)

To fully rectify the decision basis, we introduce the NPE module as a complementary correction strategy. While ASE removes primary semantic bias in the input domain, the learning process can still be distracted by residual, semantic-correlated shortcuts. NPE addresses this by scaling the gradients of fake samples to concentrated values and adding them as small perturbations to the original images. This actively helps the model recognize more underlying forgery patterns, significantly improving its ability to distinguish fake from real samples.

Ultimately, this design ensures the model completely breaks away from semantic dependence, remaining strictly focused on the core categorical distinction between real and fake rather than dataset-specific semantic information.

Specifically, we use an on-the-fly normalized-gradient perturbation [19] strategy to expose diverse synthesis traces and refine the binary decision boundary. For a given fake input X_f, we calculate the gradient of the classification loss with respect to the input pixels. The enhanced sample X_{enh} is generated as:

$$X_{enh} = X_f + \epsilon \cdot \text{sign}(\nabla_{X_f} \mathcal{L}_{CE}(f_\theta(X_f), y)) \tag{3}$$

where ϵ is the scaling factor that controls the intensity of the perturbation. Crucially, the sign$(\cdot)$ function acts as the normalization operation, mapping varying gradient magnitudes into concentrated, discrete values: positive gradients are scaled to 1, negative gradients to -1, and zero gradients to 0. Because this normalized perturbation is added directly to the original sample, it actively reveals hidden forgery patterns and disrupts the fragile semantic shortcuts the model initially relies on. The model is then optimized to correctly categorize these enhanced samples by minimizing the NPE loss:

$$\mathcal{L}_{NPE} = \mathcal{L}_{CE}(f_\theta(X_{enh}), y) \tag{4}$$

By applying this normalized-gradient perturbation exclusively to synthetic images, NPE forces the detector to ignore superficial features and instead recognize a broader range of forensic signatures. This process improves the model's ability to distinguish fake from real samples, ensuring that the final decision is based strictly on the core categorical distinction rather than semantic content.

3.4 Optimization Objective

The entire BCNet framework is optimized in a joint end-to-end manner. We define the standard binary Cross-Entropy (CE) loss as our fundamental baseline objective for distinguishing between real and fake image categories:

$$\mathcal{L}_{CE}(p, y) = -[y \log(p) + (1 - y) \log(1 - p)] \tag{5}$$

For any input X, the baseline classification loss is computed as:

$$\mathcal{L}_{CE} = \mathcal{L}_{CE}(f_\theta(X), y) \tag{6}$$

To ensure the model training is dominated by the correction of its judgment basis, the total loss function is established as a weighted sum:

$$\mathcal{L}_{total} = \lambda_{CE}\mathcal{L}_{CE} + \mathbb{I}(y = 1) \cdot (\lambda_{ASE}\mathcal{L}_{ASE} + \lambda_{NPE}\mathcal{L}_{NPE}) \tag{7}$$

where $\mathbb{I}(y = 1)$ acts as an **asymmetric basis correction** mechanism, ensuring the correction penalties apply exclusively to synthesized samples. This asymmetric design preserves the natural statistical manifold of pristine images, allowing

the network to maintain an uncorrupted, stable reference for authenticity verification. Furthermore, we use an **asymmetric weighting strategy** where the correction weights (λ_{ASE} and λ_{NPE}) are set significantly higher than λ_{CE}. This forces the optimization process to prioritize solving the difficult, erased, and perturbation-enhanced cases [63]. By doing so, it compels the model to actively expose underlying forgery patterns, ensuring that it focuses on the core distinction between real and fake categories rather than relying on dataset-specific semantics. Detailed parameter settings are provided in Section 4.1.

4 Experiments

4.1 Settings

Datasets. To comprehensively evaluate the effectiveness of our proposed method, we conducted experiments on a diverse suite of benchmarks, including AIGI-Bench [38], Chameleon [77], WildRF [5], OpenSDI [72], and AIGCDetectBenchmark [83]. For the training protocols, we adopted Setting-II [38], using 144k images from ProGAN and SDv1.4 for all five datasets. The specific details of these five evaluation datasets are introduced as follows:

AIGI-Bench integrates 25 representative generative methods from four technical streams: (a) GAN-based noise-to-image (ProGAN [26], R3GAN [25], StyleGAN3 [27], StyleGAN-XL [61], StyleSwim [81], and WFIR [74]); (b) GAN-based DeepFakes (BlendFace [62], E4S [42], FaceSwap [31], InSwap [69], and SimSwap [6]); (c) T2I Diffusion models (DALLE-3 [51], FLUX1-dev [3], Midjourney-V6 [47], GLIDE [49], Imagen3 [20], SD3 [16], and SDXL [32]); and (d) Personalized Diffusion (BLIP [34], Infinite-ID [75], PhotoMaker [37], InstantID [70], and IP-Adapter [79]). Additionally, the benchmark incorporates CommunityAI and SocialRF for unconstrained evaluation.

WildRF aligns DeepFake evaluation with real-world scenarios by utilizing highly diverse and authentic imagery manually collected directly from widely-used social platforms, including Reddit, Twitter, and Facebook.

OpenSDI comprises five distinct generative models: SD1.5 [60], SD2.1 [60], SDXL [56], SD3 [16], and Flux.1 [32] for high-fidelity generation.

AIGCDetectBenchmark consists of 17 subsets derived from a diverse array of models, ranging from GAN-based architectures (ProGAN [26], CycleGAN [85], BigGAN [4], StyleGAN [28], StyleGAN2 [29], GauGAN [53], and StarGAN [11]) to advanced Diffusion Models (SDv1.4 [60], SDv1.5 [60], SDXL [56], ADM [13], GLIDE [49], Midjourney [46], Wukong [76], VQDM [22], DALLE2 [59], and WFIR [74]) for rigorous benchmarking.

Implementation Details. We employed DINOv3 ViT-H+/16 [15, 64] as the backbone for feature extraction. To efficiently fine-tune the image encoder, Low-Rank Adaptation (LoRA) [24] was integrated into the attention modules of all 32 Transformer layers, with a rank $r = 16$, a scaling factor $\alpha = 32$, and a dropout

Table 1: Quantitative comparison with state-of-the-art methods on the AIGI-Bench dataset. Red and Blue represent the best and second-best performance, respectively.

Detector	CNNSpot		F3Net		UnivFD		FreqNet		NPR		SAFE		AIDE		FerretNet		VIB-Net		Effort		ours	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
ProGAN	97.6	99.9	97.8	99.9	98.4	99.9	99.3	100.0	99.4	100.0	100.0	100.0	97.2	99.6	99.9	100.0	99.2	99.9	100.0	100.0	99.8	100.0
R3GAN	50.4	52.7	77.4	87.6	83.5	91.2	62.3	56.8	50.8	61.1	93.8	97.4	92.9	97.1	90.2	98.5	66.3	65.9	97.1	99.2	96.0	100.0
StyleGAN3	55.8	73.1	59.5	71.3	79.6	84.5	83.0	92.4	78.4	91.7	92.9	97.9	88.1	91.4	92.8	98.9	76.1	71.4	94.8	98.5	95.7	99.2
StyleGAN-XL	52.8	64.2	61.4	72.1	84.6	93.3	79.8	84.1	60.3	75.3	95.7	98.8	88.7	93.2	86.8	98.3	76.8	83.4	87.8	96.8	96.3	100.0
StyleSwim	52.6	76.5	60.2	71.3	86.4	95.2	80.8	91.8	85.7	94.9	95.5	99.5	83.7	89.3	99.0	100.0	76.1	81.4	97.3	99.6	95.8	99.7
WFIR	49.8	50.0	50.1	51.3	70.0	82.0	58.5	48.9	51.6	65.5	51.9	59.7	71.4	90.8	50.0	91.8	83.0	95.7	85.6	100.0	95.7	99.9
BlendFace	52.4	73.4	43.8	40.4	35.0	35.3	23.3	34.1	44.5	34.7	45.1	41.6	51.5	54.2	48.0	39.6	51.8	49.2	54.9	73.1	64.0	79.4
E4S	51.1	68.9	43.7	40.3	57.0	57.1	25.8	34.7	45.0	34.4	44.4	42.6	44.3	44.3	48.4	38.0	67.4	69.4	61.2	81.1	83.8	91.1
FaceSwap	50.3	58.7	46.9	40.4	53.1	52.4	40.4	43.4	48.1	43.6	49.4	42.4	52.1	56.3	50.0	59.0	73.9	79.1	59.0	88.8	71.5	83.7
InSwap	54.5	77.9	46.5	43.3	43.7	40.2	37.5	42.1	47.8	40.7	49.8	46.2	50.9	54.6	50.0	46.1	63.1	60.4	62.8	89.5	70.9	84.1
SimSwap	52.1	70.0	46.2	44.3	43.7	40.4	36.5	41.9	47.4	42.7	48.8	47.3	54.9	62.7	49.3	52.1	68.4	68.4	70.1	92.5	73.4	86.0
FLUX1-dev	57.4	72.3	67.2	76.6	80.0	79.5	78.5	87.3	95.2	99.0	95.4	99.2	88.0	93.4	97.4	99.8	57.7	54.8	69.2	82.5	90.3	95.9
Midjourney-V6	52.3	59.8	50.8	50.2	65.3	61.5	53.9	55.9	68.8	76.9	87.3	96.1	76.4	83.0	93.5	98.3	49.5	46.1	87.1	94.7	87.5	94.8
GLIDE	51.1	60.0	51.2	54.5	76.7	80.3	75.8	77.4	82.5	94.3	90.8	96.9	93.4	97.7	97.9	99.8	65.8	62.4	83.1	93.8	96.1	99.8
DALLE-3	53.9	68.6	65.7	76.3	75.1	76.3	66.2	61.0	57.1	70.0	45.6	44.9	55.1	63.1	48.9	49.8	59.5	58.6	55.8	64.1	94.5	98.5
Imagen3	51.2	57.4	51.5	52.7	78.9	79.3	73.6	80.7	85.9	94.4	94.2	97.9	89.8	95.2	94.6	99.2	68.6	66.9	93.9	98.6	67.1	80.1
SD3	55.8	73.1	53.9	58.8	84.5	87.2	77.3	82.6	91.9	97.2	93.7	97.8	94.3	98.3	95.3	99.5	72.4	69.9	94.5	98.7	96.7	99.4
SDXL	52.8	64.2	68.0	81.1	84.7	88.0	82.7	95.2	86.6	94.4	96.7	99.2	93.5	95.7	98.8	99.9	75.9	75.6	95.5	99.2	97.1	99.8
BLIP	77.2	92.9	96.6	99.7	88.6	95.8	93.8	100.0	99.2	100.0	99.8	99.9	96.4	95.5	99.9	100.0	86.1	93.7	99.9	100.0	96.5	100.0
Infinite-ID	49.7	49.5	51.1	53.1	84.5	89.6	79.0	74.5	63.9	80.4	95.6	99.2	92.2	94.7	98.3	99.9	65.5	62.0	82.8	94.1	96.2	99.8
InstantID	53.2	80.2	88.6	95.0	85.4	93.5	79.8	86.3	63.8	79.2	95.9	99.0	91.8	96.3	99.0	100.0	76.0	83.3	87.2	95.9	96.0	100.0
IP-Adapter	52.0	65.8	57.5	66.4	82.6	87.3	78.8	79.9	82.4	91.7	94.5	97.9	90.0	95.4	98.9	99.9	73.6	71.1	95.0	98.9	92.3	97.5
PhotoMaker	50.1	58.2	58.9	70.5	69.3	72.3	77.0	74.9	48.1	43.6	95.5	99.5	91.7	95.6	98.7	99.9	65.8	66.7	69.2	82.6	83.7	93.0
SocialRF	51.1	50.6	59.3	68.9	54.4	55.2	54.2	58.1	59.1	68.4	58.1	64.9	57.8	65.0	57.5	67.5	58.8	67.6	57.9	62.0	95.0	97.8
CommunityAI	51.3	59.1	53.1	71.7	67.0	73.2	55.9	69.7	54.0	62.9	54.7	70.3	54.1	61.0	54.4	61.5	55.6	69.4	53.3	62.4	82.5	93.9
Average	55.1	67.1	60.3	65.5	72.5	75.6	66.2	70.1	67.9	73.5	78.6	81.4	77.6	82.5	79.9	83.9	69.3	70.9	79.8	89.9	88.6	94.9

rate of 0.3. To enhance model robustness, several widely used data augmentation techniques [78] were applied during training, such as Gaussian blur and image compression. For the basis correction strategy, the significance threshold τ for ASE was set to 0.75, and the scaling factor for NPE was fixed at $\epsilon = 0.0005$. The overall objective function jointly optimized the baseline CE, ASE, and NPE losses, weighted by $\lambda_{CE} = 0.1$, $\lambda_{ASE} = 0.45$, and $\lambda_{NPE} = 0.45$, respectively. The model was trained for a single epoch using the AdamW [43] optimizer with a learning rate of 5×10^{-5} and a weight decay of 0.001. Model performance was evaluated using Average Precision (A.P.) and classification Accuracy (Acc.), where Acc was computed using a decision threshold of 0.5. All experiments were implemented using the PyTorch framework [54] on dual NVIDIA RTX 4090 GPUs with a batch size of 32 to ensure stable model convergence.

Baselines. We evaluated 10 existing detection methods, including CNNSpot (CVPR 2020) [71], F3Net(ECCV 2020) [57], UnivFD(CVPR 2023) [50], FreqNet(AAAI 2024) [67], NPR(CVPR 2024) [65], AIDE(ICLR 2025) [77], SAFE (KDD 2025) [35], FerretNet(NeurIPS 2025) [39], VIB-Net(CVPR 2025) [82] and Effort(ICML 2025) [78]. To ensure fairness, we used the official experimental code. The training sets of all models were consistent with our method.

4.2 Performance Evaluation

Performance Comparison on AIGI-Bench. As summarized in Table 1, extensive evaluation across AIGI-Bench's 25 subsets [38] confirms BCNet's broad generalization. It achieves an average accuracy of 88.6%, surpassing the runner-up FerretNet (79.9%) by 8.7%. Notably, in semantic-heavy scenarios like DALLE-

Table 2: Performance comparison on challenging real-world and high-fidelity benchmarks. Red and Blue represent the best and second-best performance, respectively.

Method	Chame		WildRF						OpenSDI										Average	
			Facebook		Reddit		Twitter		SD1.5		SD2.1		SD3		SDXL		Flux.1			
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
CNNSpot	57.3	42.7	50.0	70.3	55.5	80.4	49.9	55.7	50.5	56.0	49.9	43.8	49.9	41.9	49.9	40.9	49.9	39.6	51.4	52.4
F3Net	59.0	63.7	63.8	85.0	66.6	79.7	60.2	76.5	66.8	79.2	63.4	77.9	56.1	73.2	55.1	71.6	53.6	68.3	60.5	75.0
UnivFD	51.7	42.7	55.9	59.5	64.1	71.3	62.8	70.4	63.7	75.0	63.3	81.4	59.8	78.3	60.8	79.3	56.1	73.3	59.8	70.1
FreqNet	59.6	55.9	64.4	78.7	66.3	73.3	48.5	50.2	50.4	51.4	50.0	43.8	47.8	46.2	48.0	45.9	47.0	45.0	53.6	54.5
NPR	59.1	47.4	53.1	84.9	73.9	85.1	53.4	59.6	53.8	60.3	51.7	54.0	50.6	49.6	50.5	47.8	50.1	46.1	55.1	59.4
SAFE	59.4	45.2	50.3	50.4	72.3	80.3	52.1	52.1	50.0	50.9	50.0	50.1	50.0	52.0	50.0	51.8	50.0	50.2	53.8	53.7
AIDE	57.8	44.3	65.0	79.0	70.8	78.9	63.1	68.6	59.7	68.2	62.8	72.1	57.8	65.8	58.1	65.5	54.3	58.9	61.0	66.8
FerretNet	59.3	44.2	50.0	53.4	70.6	86.7	51.7	56.3	50.3	52.1	49.9	46.7	50.0	47.8	50.0	42.5	50.0	44.5	53.5	52.7
VIB-Net	60.8	60.5	61.3	76.7	70.5	81.2	59.6	73.4	70.3	87.2	62.1	90.3	56.3	87.9	58.0	89.1	53.5	84.2	61.4	81.2
Effort	58.9	57.0	52.5	54.6	72.3	82.6	52.9	56.2	51.9	72.8	50.0	69.0	50.0	65.9	50.0	68.0	50.0	62.1	54.3	65.4
ours	82.8	90.3	95.0	98.6	97.8	99.8	97.4	99.9	77.0	86.1	78.0	88.6	75.1	88.3	76.8	88.3	63.0	81.1	82.5	91.2

Table 3: Generalization performance evaluation on the AIGCDetectBenchmark. Red and Blue represent the best and second-best performance, respectively.

Detector	CNNSpot		F3Net		UnivFD		FreqNet		NPR		SAFE		AIDE		FerretNet		VIB-Net		Effort		ours	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
ADM	50.7	66.6	55.4	82.0	53.9	65.2	64.9	73.9	67.5	77.6	68.4	92.5	95.9	99.4	69.6	92.4	75.6	91.2	78.2	98.4	91.5	97.9
DALLE2	52.2	87.8	62.7	92.3	78.1	89.8	82.0	92.3	98.1	99.8	95.7	99.9	99.3	99.9	99.4	100.0	78.6	94.6	85.9	99.9	93.2	99.1
GLIDE	52.3	83.1	55.8	85.2	85.5	95.3	85.4	94.1	93.5	98.7	94.6	99.3	99.1	99.9	98.6	100.0	84.0	96.4	85.2	99.7	99.7	100.0
Midjourney	52.5	80.4	64.8	91.3	78.3	91.7	86.4	94.4	91.2	98.9	97.1	100.0	81.5	90.3	94.7	99.7	81.0	94.3	80.6	99.5	85.3	97.5
VQDM	50.2	57.0	53.2	73.2	70.8	87.2	62.9	74.9	60.6	74.8	94.8	99.8	94.8	98.9	71.3	97.6	89.7	97.4	99.7	100.0	99.7	100.0
BigGAN	50.6	59.2	57.9	72.0	85.9	94.7	72.7	87.9	58.9	72.8	91.5	97.5	75.9	91.6	79.0	97.1	92.7	97.1	99.5	100.0	99.9	100.0
CycleGAN	59.2	91.0	78.4	87.8	88.0	96.2	80.4	97.5	70.4	98.0	88.3	99.5	92.9	98.2	78.8	99.1	98.0	99.7	100.0	100.0	98.0	99.8
GauGAN	51.2	86.1	61.1	72.9	92.1	97.6	51.8	54.0	51.6	65.6	93.8	98.1	63.8	74.1	60.7	95.8	97.3	99.2	99.7	100.0	99.9	100.0
ProGAN	97.6	99.9	97.8	99.9	98.4	99.9	99.3	100.0	99.9	99.9	99.9	100.0	97.2	99.6	99.9	100.0	99.2	99.9	100.0	100.0	99.8	100.0
SDXL	50.9	76.6	70.0	94.8	72.9	96.3	79.3	89.0	95.0	99.5	99.8	100.0	83.0	99.4	99.9	100.0	83.2	99.8	94.1	100.0	99.6	99.9
SDv1.4	77.7	97.6	99.2	100.0	96.1	99.3	95.6	99.9	99.7	100.0	100.0	100.0	99.5	99.9	98.2	100.0	99.3	100.0	100.0	100.0	99.8	100.0
SDv1.5	77.7	97.6	99.3	99.9	95.7	99.1	95.6	99.9	99.6	99.9	100.0	99.9	99.6	99.9	98.3	99.9	99.1	99.9	99.9	100.0	99.7	100.0
StarGAN	60.5	82.5	59.3	99.7	97.5	99.7	94.2	100.0	96.7	100.0	100.0	100.0	97.8	99.7	99.9	100.0	97.2	99.7	100.0	100.0	90.2	98.2
StyleGAN	68.2	94.8	68.7	91.1	75.4	90.9	86.9	95.9	88.9	98.0	97.0	99.9	94.5	99.4	93.2	99.8	85.6	93.9	91.4	99.9	98.2	99.9
StyleGAN2	61.0	94.2	60.9	85.6	72.7	91.1	83.1	94.2	88.4	99.1	98.9	100.0	98.2	99.9	97.1	100.0	86.5	98.1	87.1	99.4	92.4	97.4
WFIR	50.2	65.5	50.1	51.3	70.4	83.0	61.5	57.0	60.8	78.3	59.7	51.9	71.4	90.8	50.0	91.9	83.0	95.7	85.6	100.0	95.4	99.9
Wukong	63.7	88.9	97.0	99.8	90.4	97.4	94.7	99.1	98.8	99.9	100.0	99.8	99.2	99.9	94.5	99.9	98.9	99.9	100.0	100.0	99.8	100.0
Average	60.4	82.9	70.1	87.0	82.5	92.6	81.0	88.5	83.5	91.8	92.9	96.4	90.8	96.5	87.2	98.4	89.9	97.5	93.3	99.8	96.6	99.4

3 and E4S, where baselines like SAFE drop to near-random (∼45%) due to semantic bias, BCNet retains high accuracy (94.5% and 83.8%). This proves that ASE successfully mitigates semantic interference by forcing the model to focus on intrinsic traces. Furthermore, on wild-simulated subsets like SocialRF and CommunityAI, BCNet outperforms Effort, scoring 95.0% and 82.5% respectively. This confirms NPE effectively amplifies forgery artifacts, ensuring stable detection even in diverse, unconstrained environments.

Performance Comparison on Challenging Real-World and High-Fidelity Benchmarks. To evaluate BCNet in diverse and challenging scenarios, we test it on Chameleon [77] (High-Fidelity), WildRF [5] (Social Media), and OpenSDI [72] (Open-World Editing). As reported in Table 2, our framework achieves an 82.5% mean accuracy, surpassing VIB-Net (61.4%) by 21.1%. Specifically, on Chameleon, where inherent semantic bias limits VIB-Net to 60.8%, BCNet reaches 82.8%. This demonstrates ASE successfully suppresses semantic interference to isolate intrinsic synthesis traces. Furthermore, BCNet exhibits

Table 4: Ablation study of the individual proposed components on various representative benchmarks, with results quantitatively reported in Accuracy (Acc.).

Method	ASE	NPE	AIGI-Bench	Chameleon	WildRF	OpenSDI
Baseline			84.2	76.3	90.1	70.5
Baseline + ASE	✓		85.4(1.2↑)	79.5(3.2↑)	95.2(5.1↑)	71.2(0.7↑)
Baseline + NPE		✓	85.7(1.5↑)	79.9(3.6↑)	93.1(3.0↑)	71.8(1.3↑)
Baseline + ASE + NPE	✓	✓	**88.6**(4.4↑)	**82.6**(6.3↑)	**96.7**(6.6↑)	**74.0**(3.5↑)

strong resilience on WildRF; while baselines like CNNSpot drop to near-random (49.9%) on Twitter, our method sustains 97.4%. This confirms NPE effectively preserves discriminative forgery patterns despite complex transcoding and real-world degradations. Finally, achieving 63.0% on the advanced Flux.1 model (OpenSDI) validates BCNet's generalization to unseen generative architectures.

Performance Comparison on AIGCDetectBenchmark. Evaluation on the AIGCDetectBenchmark [83] (Table 3) confirms BCNet's robust generalization, achieving a 96.6% average accuracy and surpassing Effort (93.3%) by 3.3%. The performance gap is clear on Diffusion models like WFIR and ADM; while CNNSpot drops to near-random (∼50%) and Effort struggles (78.2% on ADM), BCNet maintains 95.4% and 91.5%. This indicates ASE effectively extracts intrinsic synthesis traces even from realistic diffusion outputs. Concurrently, BCNet achieves 99.9% on GauGAN, where NPR fails significantly (51.6%). This consistent success across both Diffusion and GAN paradigms confirms NPE actively isolates fundamental forgery characteristics, proving BCNet's broad applicability across established foundational generative architectures.

4.3 Ablation Study

To isolate individual component contributions, we conducted an ablation study across four diverse benchmarks (Table 4). Integrating **ASE** alone provides consistent performance gains, boosting accuracy by **5.1%** on WildRF and **3.2%** on Chameleon. This confirms that masking prominent semantic regions effectively forces the network to mine intrinsic synthesis traces. Similarly, **NPE** independently improves performance (e.g., +**3.6%** on Chameleon), demonstrating that normalized perturbations successfully amplify latent forgery characteristics. Crucially, the unified framework exhibits strong synergy. Combining both modules yields an overall increase of **6.3%** on Chameleon and **6.6%** on WildRF, culminating at **88.6%** on AIGI-Bench. This proves our complementary basis correction strategy is essential for achieving robust and optimal detection results.

4.4 Robustness Evaluation

We further assessed the model's resilience against common image degradations, specifically JPEG compression and Gaussian blur (Table 5). Under JPEG com-

Table 5: Robustness evaluation against common image degradations (JPEG compression and Gaussian blur), with results quantitatively reported in Accuracy (Acc.).

Type	Setting	AIGI-Bench	Chameleon	WildRF	OpenSDI	AIGCBench
Original		**88.6**	82.8	96.7	74.0	**96.7**
JPEG Compression	QF=95	88.3 (0.3↓)	82.9 (0.1↑)	96.8 (0.1↑)	73.9 (0.1↓)	96.4 (0.3↓)
	QF=75	86.3 (2.3↓)	82.2 (0.6↓)	95.1 (1.6↓)	71.9 (2.1↓)	95.3 (1.4↓)
	QF=60	82.4 (6.2↓)	77.2 (5.6↓)	91.8 (4.9↓)	69.1 (4.9↓)	93.9 (2.8↓)
Gaussian Blur	$\sigma = 0.5$	88.4 (0.2↓)	82.8 (-)	96.9 (0.2↑)	73.4 (0.6↓)	96.7 (-)
	$\sigma = 1.5$	84.9 (3.7↓)	**83.7** (0.9↑)	**97.3** (0.6↑)	**74.9** (0.9↑)	94.1 (2.6↓)
	$\sigma = 2.5$	84.8 (3.8↓)	83.1 (0.3↑)	96.9 (0.2↑)	74.8 (0.8↑)	90.2 (6.5↓)

pression ($QF = 60$), accuracy on the WildRF dataset remains strong at **91.8%**, and AIGCBench retains **93.9%** (a 2.8% drop). This proves NPE preserves critical discriminative features even when destructive signal distortions erase standard visual cues. Notably, under Gaussian blur ($\sigma = 1.5$), performance improves on Chameleon (+**0.9%**) and OpenSDI (+**0.9%**). This demonstrates ASE avoids relying on inherently fragile high-frequency details. Instead, moderate blur acts as a low-pass filter, stripping away superficial pixel-level distractions and compelling the network to focus on degradation-resistant structural inconsistencies. Ultimately, this confirms our strategy establishes a robust forensic foundation for highly complex and unpredictable real-world environments.

4.5 Extended Analysis

Impact of Asymmetric Basis Correction. In our default configuration, BCNet applies Attention-Guided Semantic Erasure (ASE) and Normalized-Gradient Perturbation Enhancement (NPE) exclusively to synthesized samples. To validate this asymmetric design, we investigate activating these modules on real images (Table 6). Modifying pristine samples inevitably degrades overall detection accuracy. Applying both drops the Average Accuracy from 87.7% to 82.9%. Activating NPE alone causes a decline to 82.8%, indicating that injecting normalized perturbations into real images corrupts their natural statistical manifold, confusing the model's authenticity baseline. Similarly, applying only ASE reduces accuracy to 86.1% by unnecessarily masking pristine visual content, disrupting the structural integrity of natural images. These findings confirm real samples must serve as an unaltered reference anchor. Restricting basis correction entirely to fake images prevents distorting natural data distributions while effectively isolating forgery artifacts across diverse generative domains.

Extended Analysis on Perturbation Versatility. To validate the universality of our perturbation strategy, we replaced the default generation method with advanced adversarial algorithms, specifically PGD [45] and MI-FGSM [14] (Table 7). BCNet sustains highly consistent performance, with overall accuracy

Table 6: Impact of asymmetric basis correction: a performance comparison when activating the ASE and NPE modules on original real samples.

Setting on Real	AIGI-Bench		Chameleon		WildRF		OpenSDI		AIGCBench		Avg. Acc.	Avg. A.P.
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.		
ASE + NPE	83.4	92.9	76.9	90.7	89.3	**99.5**	70.6	85.9	94.4	99.3	82.9	93.7
ASE	86.3	93.4	81.2	89.6	94.2	99.4	72.4	85.6	96.4	**99.4**	86.1	93.5
NPE	83.3	92.8	76.8	**91.0**	89.2	**99.5**	70.5	**86.2**	94.1	99.3	82.8	**93.8**
None (ours)	**88.6**	**94.9**	**82.8**	90.3	**96.7**	99.4	**74.0**	84.5	**96.6**	**99.4**	**87.7**	93.7

Table 7: Gradient enhancement versatility analysis: performance comparison when replacing the default NPE perturbation with alternative adversarial methods.

Method	AIGI-Bench		Chameleon		WildRF		OpenSDI		AIGCBench		Avg. Acc.	Avg. A.P.
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.		
MI-FGSM	86.6	94.1	80.7	**91.0**	93.2	**99.4**	72.7	**87.0**	96.2	**99.6**	85.9	**94.2**
PGD	87.9	93.8	**83.0**	88.8	95.4	98.6	73.3	84.9	94.6	98.6	86.8	92.9
ours	**88.6**	**94.9**	82.8	90.3	**96.7**	**99.4**	**74.0**	84.5	**96.6**	99.4	**87.7**	93.7

fluctuations bounded within 2%. This proves its robustness stems from the core perturbation paradigm rather than a specific underlying algorithm. For instance, PGD's stable accuracy on Chameleon (83.0%) and MI-FGSM's strong overall performance (94.2% average A.P.) confirm that applying normalized-gradient perturbations consistently amplifies latent artifacts regardless of the implementation. This highlights NPE's broad generalizability, demonstrating it reliably isolates synthesis traces independent of specific generative mechanisms.

Impact of Asymmetric Loss Weighting. To evaluate this strategy, we varied the ratios of λ_{ASE}, λ_{NPE}, and λ_{CE} (Table 8). The framework achieves peak Average Accuracy (87.7%) when basis correction components dominate the optimization objective (0.45:0.45:0.1). Conversely, increasing the baseline λ_{CE} weight from 0.1 to 0.9 causes a performance decline to 84.6%. Specifically, accuracy on Chameleon drops from 82.8% to 79.4%, and WildRF degrades from 96.7% to 91.9%. This illustrates that diluting basis correction supervision weakens the model's ability to rectify inherent semantic bias. These findings confirm assigning higher weights to ASE and NPE losses is essential to prevent overfitting to superficial shortcuts and establish a reliable forensic representation.

Impact of Backbone Model Capacity. We systematically evaluated BCNet's scalability across varying backbone capacities, from ViT-L (300M) to a 7B parameter model (Table 9). Performance exhibits consistent improvement as model capacity expands. The 7B variant achieves a 90.3% average accuracy, featuring an increase on Chameleon from 79.3% to 90.6%. This indicates that larger models provide diverse visual priors, which BCNet effectively leverages to isolate

Table 8: Ablation study of loss weighting ratios on representative benchmarks.

$\lambda_{ASE} : \lambda_{NPE} : \lambda_{CE}$	AIGI-Bench		Chameleon		WildRF		OpenSDI		AIGCBench		Avg. Acc.	Avg. A.P.
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.		
0.05 : 0.05 : 0.9	85.0	94.2	79.4	**91.9**	91.9	**99.6**	71.1	85.5	95.7	99.4	84.6	94.1
0.15 : 0.15 : 0.7	85.5	93.9	80.3	91.8	93.4	**99.6**	71.6	86.5	95.9	99.4	85.3	94.2
0.25 : 0.25 : 0.5	85.6	93.9	79.6	90.9	93.6	**99.6**	71.6	86.5	95.9	99.4	85.3	94.1
0.35 : 0.35 : 0.3	85.8	93.6	80.3	91.5	94.0	99.5	72.2	**87.2**	96.3	**99.6**	85.7	**94.3**
0.45 : 0.45 : 0.1	**88.6**	**94.9**	**82.8**	90.3	**96.7**	99.4	**74.0**	84.5	**96.6**	99.4	**87.7**	93.7

Table 9: Performance comparison across different backbone model scales.

Backbone Scale	AIGI-Bench		Chameleon		WildRF		OpenSDI		AIGCBench		Avg. Acc.	Avg. A.P.
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.		
300M	87.9	94.4	79.3	87.0	91.4	97.7	72.9	85.4	94.7	98.8	85.2	92.7
840M	**88.6**	**94.9**	82.8	90.3	96.7	99.4	74.0	84.5	96.6	99.4	87.7	93.7
7B	88.3	94.4	**90.6**	**96.9**	**98.2**	**99.7**	**76.6**	**89.0**	**97.9**	**99.7**	**90.3**	**95.9**

complex synthesis traces. These gains highlight the scalability of our asymmetric strategy, proving that increased foundational capacity further enhances the framework's ability to counter highly sophisticated, real-world forgeries.

t-SNE Visualization. The t-SNE [44] visualization across **10 representative subsets** of AIGI-Bench qualitatively illustrates BCNet's discriminative capability. As depicted in Figure 3, we evaluate a diverse spectrum, spanning legacy GANs (ProGAN), advanced Diffusion models (**FLUX.1-dev**, **SD3**), and unconstrained scenarios (**SocialRF**, **CommunityAI**). Comprehensive visualizations for all 25 subsets of AIGI-Bench are provided in the supplementary material. The plots exhibit consistent separation between real (blue) and fake (red) clusters across all domains. This intra-class compactness paired with distinct inter-class margins visually validates our complementary basis correction. The cluster isolation within high-fidelity datasets proves ASE successfully mitigates semantic interference to extract intrinsic traces. Concurrently, the robust boundaries in real-world samples confirm NPE effectively amplifies discriminative artifacts. Ultimately, BCNet maintains clear categorical separation despite complex real-world degradations or diverse generative architectures.

Grad-CAM Visualization. To qualitatively evaluate the decision basis, we randomly select one real image and 13 fake samples generated by different generative models from the test set and employ Grad-CAM [18] to visualize their attention maps. As shown in Figure 4, the baseline DINOv3 exhibits a significant semantic bias, with its resulting attention predominantly concentrated on salient foreground objects like human faces, food, or books. In contrast, our proposed BCNet effectively disperses the attention across broader and more decentralized regions, such as background structures and environmental textures. This visual

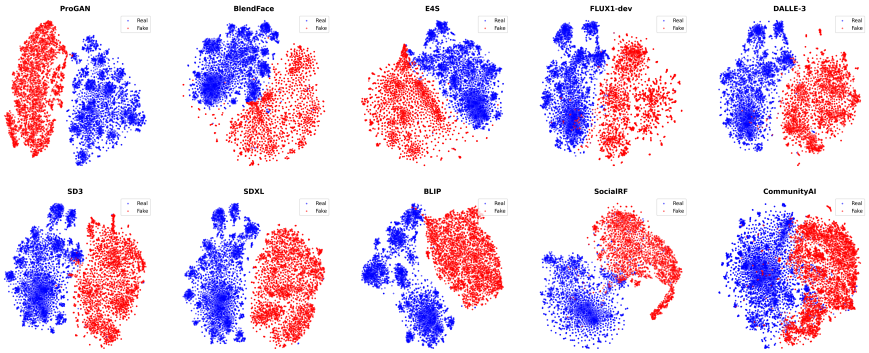

Fig. 3: t-SNE visualization on 10 representative subsets of AIGI-Bench. We select diverse benchmarks covering legacy GANs (ProGAN), advanced Diffusion models (FLUX.1-dev, SD3), and challenging wild scenarios (SocialRF, CommunityAI). The distinct separation between real (blue) and fake (red) clusters across this broad spectrum validates that BCNet effectively learns highly discriminative feature representations for robust forgery detection in various complex real-world scenarios.

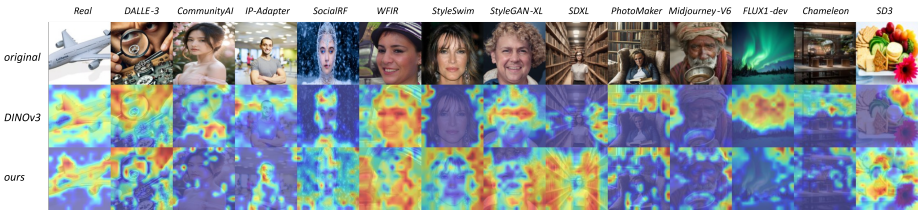

Fig. 4: Grad-CAM visualization of attention maps. While the baseline DINOv3 focuses almost exclusively on salient foreground objects, our **BCNet** effectively disperses attention across broader regions, successfully mitigating semantic dependence.

shift clearly indicates that our framework successfully suppresses the reliance on prominent semantic shortcuts, compelling the model to capture more intrinsic and subtle forgery patterns distributed throughout the entire image.

5 Conclusions

In this paper, we propose the Basis Correction Network (BCNet) to address the severe semantic dependence that limits the generalization of Vision Foundation Models (VFMs) in synthetic image detection. By introducing a complementary basis correction strategy, BCNet efficiently adapts pre-trained features via LoRA. Specifically, Attention-Guided Semantic Erasure (ASE) adaptively identifies and masks prominent semantic regions, compelling the model to correct its decision basis. Meanwhile, Normalized-Gradient Perturbation Enhancement (NPE) scales the gradients of fake samples to concentrated values and adds them as small perturbations. This actively helps the network recognize more forgery

patterns, ensuring it strictly focuses on the core distinction between real and fake categories rather than misleading semantic information. Extensive experiments across diverse AI-generated datasets confirm that BCNet overcomes semantic bias, achieving robust generalization and offering a fresh, highly effective perspective on semantic generalization for real-world deepfake detection.

References

1. Amerini, I., Barni, M., Battiato, S., Bestagini, P., Boato, G., Bruni, V., Caldelli, R., De Natale, F., De Nicola, R., Guarnera, L., et al.: Deepfake media forensics: Status and future challenges. Journal of Imaging **11**(3), 73 (2025)
2. Babaei, R., Cheng, S., Duan, R., Zhao, S.: Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. Journal of Sensor and Actuator Networks **14**(1), 17 (2025)
3. Black Forest Labs: Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev` (2024)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Cavia, B., Horwitz, E., Reiss, T., Hoshen, Y.: Real-time deepfake detection in the real-world. ArXiv **abs/2406.09398** (2024), `https://api.semanticscholar.org/CorpusID:270440423`
6. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2003–2011 (2020)
7. Chen, R., Gao, J., Lin, K., Zhang, K., Zhao, Y., Guan, I., Yao, T., Ding, S.: Task-model alignment: A simple path to generalizable ai-generated image detection. arXiv preprint arXiv:2512.06746 (2025)
8. Chen, R., Xi, J., Yan, Z., Zhang, K.Y., Wu, S., Xie, J., Chen, X., Xu, L., Guan, I., Yao, T., et al.: Dual data alignment makes ai-generated image detector easier generalizable. arXiv preprint arXiv:2505.14359 (2025)
9. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
10. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2219–2228 (2019)
11. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
12. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the bar of ai-generated image detection with clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4356–4366 (2024)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
14. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)

15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
17. Fu, X., Yan, Z., Yang, Z., Yao, T., Zhao, Y., Ding, S., Li, X.: Pid: Generalized ai-generated images detection with pixelwise decomposition residuals. In: Forty-second International Conference on Machine Learning (2025)
18. Gildenblat, J., contributors: Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam` (2021)
19. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
20. Google DeepMind: Imagen3. `https://deepmind.google/technologies/imagen-3` (2024)
21. Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L.: Are gan generated images easy to detect? a critical analysis of the state-of-the-art. arXiv preprint arXiv:2104.02617 (2021)
22. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10696–10706 (2022)
23. Guan, W., Wang, W., Dong, J., Peng, B.: Improving generalization of deepfake detectors by imposing gradient regularization. IEEE Transactions on Information Forensics and Security **19**, 5345–5356 (2024)
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. ArXiv **abs/2106.09685** (2021), `https://api.semanticscholar.org/CorpusID:235458009`
25. Huang, N., Gokaslan, A., Kuleshov, V., Tompkin, J.: The gan is dead; long live the gan! a modern gan baseline. Advances in Neural Information Processing Systems **37**, 44177–44215 (2024)
26. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
27. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in neural information processing systems **34**, 852–863 (2021)
28. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
29. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
30. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
31. Kowalski, M.: Faceswap. `https://github.com/MarekKowalski/FaceSwap` (2020), gitHub Repository
32. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz,

D., Muller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. ArXiv **abs/2506.15742** (2025), https://api.semanticscholar.org/CorpusID:279464475

33. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
34. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
35. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: An image transformation perspective. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (2024), https://api.semanticscholar.org/CorpusID:271860186
36. Li, S., Li, X., Chiariglione, L., Luo, J., Wang, W., Yang, Z., Mandic, D., Fujita, H.: Introduction to the special issue on ai-generated content for multimedia. IEEE Transactions on Circuits and Systems for Video Technology **34**(8), 6809–6813 (2024)
37. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked id embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8640–8650 (2024)
38. Li, Z., Yan, J., He, Z., Zeng, K., Jiang, W., Xiong, L., Fu, Z.: Is artificial intelligence generated image detection a solved problem? arXiv preprint arXiv:2505.12335 (2025)
39. Liang, S., Liu, J., Chen, R., Guan, Q.: Ferretnet: Efficient synthetic image detection via local pixel dependencies. ArXiv **abs/2509.20890** (2025), https://api.semanticscholar.org/CorpusID:281526218
40. Liu, H., Tan, Z., Tan, C., Wei, Y., Zhao, Y., Wang, J.: Forgery-aware adaptive transformer for generalizable synthetic image detection. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10770–10780 (2023), https://api.semanticscholar.org/CorpusID:266573117
41. Liu, M.H., Cheng, H., Luo, X., Xu, X.S.: Suppressing gradient conflict for generalizable deepfake detection. arXiv preprint arXiv:2507.21530 (2025)
42. Liu, Z., Li, M., Zhang, Y., Wang, C., Zhang, Q., Wang, J., Nie, Y.: Fine-grained face swapping via regional gan inversion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8578–8587 (2023)
43. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
44. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
45. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
46. Midjourney: Midjourney. https://www.midjourney.com/home/ (2022)
47. Midjourney Team: Midjourney v6.1. https://www.midjourney.com/home (2024)
48. Nguyen, D., Mejri, N., Singh, I.P., Kuleshova, P., Astrid, M., Kacem, A., Ghorbel, E., Aouada, D.: Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17395–17405 (2024)

49. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
50. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 24480–24489 (2023), https://api.semanticscholar.org/CorpusID:257038440
51. OpenAI Team: Dall-e 3 ai image generator. https://dalle3.ai/ (2024)
52. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. ArXiv **abs/2304.07193** (2023), https://api.semanticscholar.org/CorpusID:258170077
53. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
54. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
55. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
56. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
57. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision (2020), https://api.semanticscholar.org/CorpusID:220647499
58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
59. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
61. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
62. Shiohara, K., Yang, X., Taketomi, T.: Blendface: Re-designing identity encoders for face-swapping. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7634–7644 (2023)
63. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016)
64. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A.,

Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104
65. Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 28130–28139 (2024), https://api.semanticscholar.org/CorpusID:266348433
66. Tan, C., Tao, R., Liu, H., Gu, G., Wu, B., Zhao, Y., Wei, Y.: C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7184–7192 (2025)
67. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space learning. ArXiv **abs/2403.07240** (2024), https://api.semanticscholar.org/CorpusID:268890333
68. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
69. Wang, H., Kleynhans, A., Estrada, A.: Inswap. https://github.com/haofanwang/inswapper (2023), gitHub Repository
70. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
71. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
72. Wang, Y., Huang, Z., Hong, X.: Opensdi: Spotting diffusion-generated images in the open world. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4291–4301 (2025), https://api.semanticscholar.org/CorpusID:277313522
73. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22445–22455 (2023)
74. West, J., Bergstrom, C.: Which face is real? https://www.whichfaceisreal.com/ (2019)
75. Wu, Y., Li, Z., Zheng, H., Wang, C., Li, B.: Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In: European Conference on Computer Vision. pp. 279–296. Springer (2024)
76. Wukong: Wukong. Available at https://xihe.mindspore.cn/modelzoo/wukong (2022)
77. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for ai-generated image detection. arXiv preprint arXiv:2406.19435 (2024)
78. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable ai-generated image detection. arXiv preprint arXiv:2411.15633 (2024)
79. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
80. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)

81. Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., Guo, B.: Styleswin: Transformer-based gan for high-resolution image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11304–11314 (2022)
82. Zhang, H., He, Q., Bi, X., Li, W., Liu, B., Xiao, B.: Towards universal ai-generated image detection by variational information bottleneck network. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 23828–23837 (2025), https://api.semanticscholar.org/CorpusID:280069753
83. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: Patchcraft: Exploring texture patch for efficient ai-generated image detection. arXiv preprint arXiv:2311.12397 (2023)
84. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13001–13008 (2020)
85. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)