NS-Net: Decoupling CLIP Semantic Information through NULL-Space for Generalizable AI-Generated Image Detection

Jiazhen Yan[1*], Fan Wang[1*], Weiwei Jiang[1], Ziqiang Li[1†], Zhangjie Fu[1‡]

[1]School of Computer Science, Nanjing University of Information Science and Technology
247918horizon@gmail.com, wf71103@126.com,
weiwei.jiang@nuist.edu.cn, iceli@mail.ustc.edu.cn, fzj@nuist.edu.cn

Abstract

The rapid progress of generative models, such as GANs and diffusion models, has facilitated the creation of highly realistic images, raising growing concerns over their misuse in security-sensitive domains. While existing detectors perform well under known generative settings, they often fail to generalize to unknown generative models, especially when semantic content between real and fake images is closely aligned. In this paper, we revisit the use of CLIP features for AI-generated image detection and uncover a critical limitation: the high-level semantic information embedded in CLIP's visual features hinders effective discrimination. To address this, we propose NS-Net, a novel detection framework that leverages NULL-Space projection to decouple semantic information from CLIP's visual features, followed by contrastive learning to capture intrinsic distributional differences between real and generated images. Furthermore, we design a Patch Selection strategy to preserve fine-grained artifacts by mitigating semantic bias caused by global image structures. Extensive experiments on an open-world benchmark comprising images generated by 40 diverse generative models show that NS-Net outperforms existing state-of-the-art methods, achieving a 7.4% improvement in detection accuracy, thereby demonstrating strong generalization across both GAN- and diffusion-based image generation techniques.

Introduction

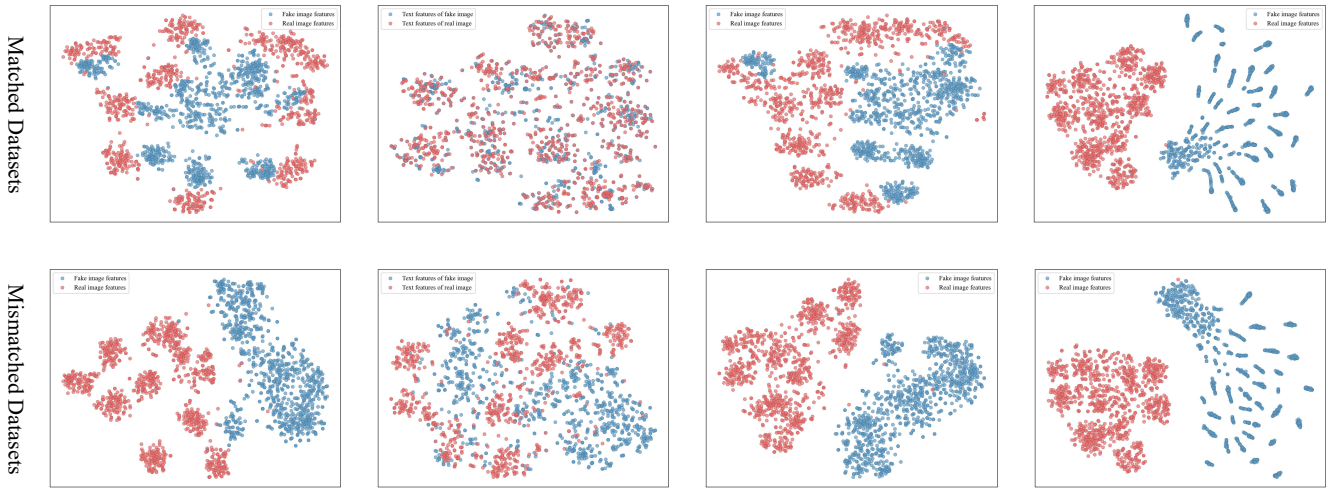
The rapid advancement of artificial intelligence has significantly accelerated the development of generative models, such as GANs (Karras et al. 2018; Park et al. 2019; Huang et al. 2024) and diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Wu et al. 2024b). While these models are widely employed to generate highly realistic images and have enabled numerous beneficial applications, their potential misuse (Li et al. 2025c) raises serious concerns related to economic stability, political manipulation, and social security. Consequently, the reliable detection of AI-generated images has emerged as a critical research focus. To date, several effective detection methods (Zhao et al. 2021; Wu et al. 2024a; Tan et al. 2024a; Tao et al. 2025; Li et al. 2025a) have been proposed, particularly in scenarios where the generative model is known. However, developing robust and generalizable detection techniques that can effectively identify images from unknown generative models remains a fundamental and unresolved challenge.

Recently, a number of studies on AI-generated image detection (Ojha, Li, and Lee 2023; Liu et al. 2024a; Fu et al. 2025) have sought to enhance generalization by directly leveraging or fine-tuning CLIP (Radford et al. 2021)—a multimodal model pre-trained on large-scale image-text datasets, known for its strong generalization capabilities across unseen visual and textual data. Although the effectiveness of CLIP in this context has been demonstrated through quantitative experiments, there remains a lack of in-depth investigation into the intrinsic properties of its extracted features and how these characteristics influence detection performance.

Inspired by VIB-Net (Zhang et al. 2025), which aims to decouple category-related features, we revisit the distinctions between features with different intrinsic properties extracted by CLIP. Specifically, we construct two datasets, each containing 1,000 real images from LSUN and 1,000 images generated by SDXL. In the *matched dataset*, the generated images are conditioned on BLIP-derived textual descriptions of the corresponding real images. In contrast, the *mismatched dataset* consists of randomly selected real images and generated images conditioned on unrelated text prompts. We extract visual features using CLIP's image encoder and visualize them using t-SNE, as shown in Figure 1(a). The features from the mismatched dataset exhibit clear clustering, suggestive of a linearly separable binary classification problem. In contrast, features from the matched dataset are more entangled and less distinguishable. This suggests that CLIP effectively separates real and generated images when semantic alignment is weak but struggles when the semantic content of both is closely aligned. Figure 1(b) presents the corresponding visualization of semantic features extracted by CLIP's text encoder. Based on these observations, we propose the following hypothesis: **The semantic information embedded in the visual features extracted by CLIP negatively impacts the performance of AI-generated image detection.**

To decouple the semantic information embedded in visual features, we construct the NULL-Space (Fang et al. 2024) corresponding to the semantic components in CLIP's image features. The NULL-Space, formally defined as the set of vectors mapped to the zero vector under a given linear

*These authors contributed equally.
†Corresponding author
‡Corresponding author



(a) Visual features extracted by CLIP's image encoder

(b) Semantic features extracted by CLIP's text encoder

(c) Visual features extracted by CLIP's image encoder in NULL-Space

(d) Visual features extracted by NS-Net

Figure 1: **T-SNE Visualization of Features Extracted from the Matched Dataset and the Mismatched Dataset**.

transformation, enables the removal of specific feature components. Given that CLIP is trained for text-image alignment, we leverage text features as a more tractable and semantically explicit representation, rather than attempting to disentangle the complex semantics directly from image features. Specifically, we extract features from the associated text prompts and compute their NULL-Space. Visual features extracted by CLIP's image encoder are then projected onto this NULL-Space, thereby eliminating semantically aligned information and retaining components more relevant to forgery detection. As illustrated in Figure 1(c), this projection—achieved without any additional training—results in a clear separation between real and generated images.hile reducing the influence of semantic content on forgery detection (Zheng et al. 2024).

To this end, we propose an innovative method, termed NS-Net, designed to extract more generalizable artifact information. Specifically, we leverage NULL-Space projection to decouple high-level semantic information encoded in CLIP features, followed by contrastive learning to model the underlying distributional differences between real and generated images. This strategy improves generalization ability beyond simple classification accuracy. Moreover, traditional center-cropping operations (Li et al. 2025b) often discard critical forgery traces—especially when manipulations occur away from the image center. To mitigate this, we introduce a Patch Selection strategy. Each image is first divided into uniform patches, from which those with the highest and lowest texture richness (based on predefined metrics) are selected. These selected patches are then shuffled and rearranged to form a new input image of the desired size. This process reduces the dominance of semantic content and enhances the preservation of fine-grained forgery cues. As shown in Figure 1(d), our method effectively removes task-irrelevant semantic features while retaining the generator-specific artifact patterns crucial for reliable detection. The main contributions of this work are summarized as follows:

- We innovatively construct the NULL-Space of the semantic features of generated images, which simply and efficiently eliminates task-irrelevant semantic information embedded in visual features extracted by image encoders, thus improving the generalization detection ability of AI-generated images.
- We design a data preprocessing strategy called Patch Selection, which not only reduces the loss of forged information caused by center cropping operation adopted in previous work, but also utilizes patch rearrangement to further weaken the interference of irrelevant semantic information on detection performance.
- Comprehensive experimental results have verified that our method exhibits strong generalization ability across 40 different generative models. Compared with the state-of-the-art methods, the detection accuracy is improved by 7.4%, which further highlights its superiority and universality.

Related Work

Image Generation

With the rapid advancement of image generation technologies, the quality of synthetic images has significantly improved, enabling a wide range of practical applications. Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) introduces adversarial training between a generator and discriminator, leading to substantial gains in image realism (Karras 2019; Park et al. 2019). More recently, diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) have emerged as a powerful alternative, generating high-resolution and diverse images through a progressive

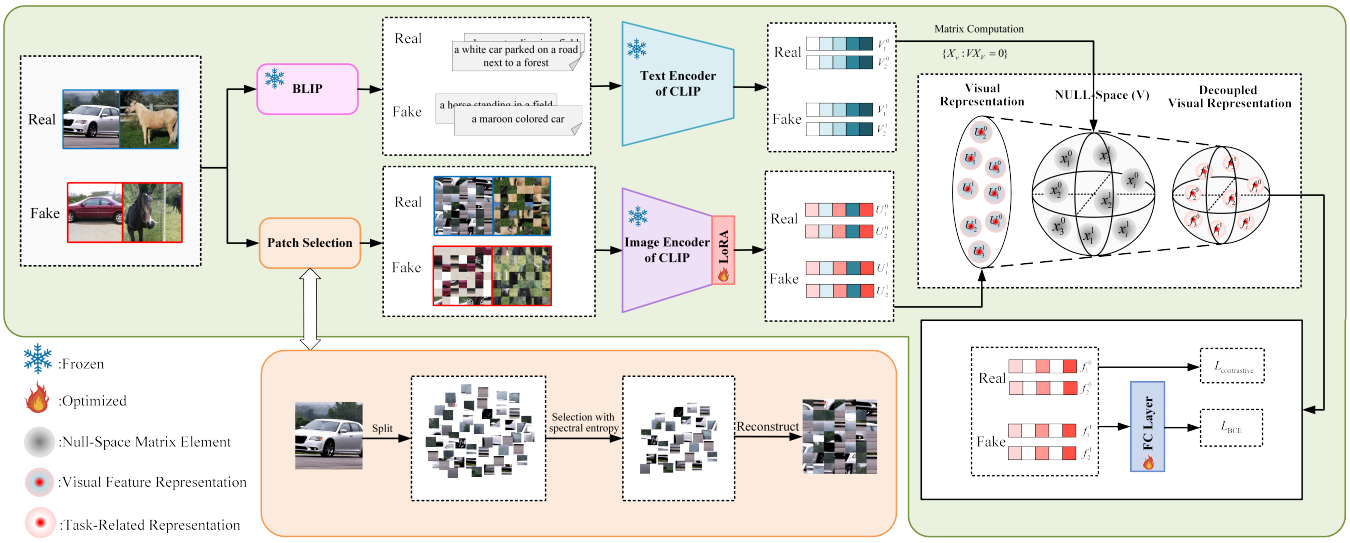


Figure 2: **Architecture of NS-Net for Generalizable AI-Generated Image Detection**. Specifically, we first employ the Patch Selection strategy adjusted for CLIP's input size to preserve potential forgery-related artifacts. Subsequently, the visual features extracted by the CLIP's image encoder are projected onto the NULL-Space of the semantic information, effectively removing task-irrelevant semantic components. The resulting features, tailored for the detection task, are then utilized in a contrastive learning framework, which can not only guide the linear classification layer but also capture the intrinsic distributional differences between real and AI-generated images, enhancing the model's ability to generalize beyond simple classification.

denoising process (Ramesh et al. 2022; Rombach et al. 2022), while offering greater training stability than GANs. Beyond unconditional synthesis, techniques such as customized generation (Wu et al. 2024b; Ye et al. 2023) and image inpainting (Brooks, Holynski, and Efros 2023) have gained traction, enabling users to guide the generation process using textual or conditional inputs, which enhance content fidelity and better accommodate user-specific requirements.

AI-Generated Image Detection

To mitigate the societal risks associated with AI-generated images, researchers (McCloskey and Albright 2018; Farid 2022; Haliassos et al. 2021; Chen et al. 2022) have increasingly focused on developing effective detection methods. As generative models have advanced, recent researches have shifted toward improving the generalization capabilities of detection models. Some methods extract subtle artifacts from fine-grained spatial features, utilizing limited receptive fields (Chai et al. 2020; Cavia et al. 2024), modeling local pixel relationships (Tan et al. 2024b), or employing patch-based analysis (Chen and Yang 2020; Baraldi et al. 2025). Others explore frequency-domain artifacts to enhance generalization across generative models (Luo et al. 2021; Tan et al. 2024a; Yan et al. 2025). In addition, LGrad (Tan et al. 2023) leverages gradient signals from pre-trained models, while SAFE (Li et al. 2025b) combines cropping with data augmentation techniques such as ColorJitter, RandomRotation, and RandomMask to improve detection performance against sophisticated generation methods.

In recent years, many researchers have explored leveraging CLIP (Radford et al. 2021) for detecting AI-generated images. UnivFD (Ojha, Li, and Lee 2023) utilize CLIP's image-text features for binary classification, demonstrating promising generalization capabilities. However, subsequent studies (Liu et al. 2024a,b; Fu et al. 2025) reveal that CLIP's native features are not fully optimized for forgery detection, prompting efforts to fine-tune the model to better capture subtle artifacts. To this end, Fatformer (Liu et al. 2024a) introduces a forgery-aware adapter that integrates both spatial and frequency-domain extractors. C2P_CLIP (Tan et al. 2025) further enhances CLIP's detection capability by injecting category-consistent cues into the text encoder. Additionally, VIB-Net (Zhang et al. 2025) addresses the issue of category interference in CLIP's features by employing a Variational Information Bottleneck framework.

Despite ongoing advances, the intrinsic properties of CLIP features remains insufficiently understood. We hypothesize that the entanglement of semantic information within these features impairs the model's ability to capture task-specific artifact cues essential for forgery detection.

Methodology

We propose the **NULL-Space Network (NS-Net)**, a novel framework designed to decouple semantic information embedded in CLIP features, preserving only the components relevant to forgery detection, which significantly improves the model's generalization ability. As illustrated in Figure 2, our architecture begins with a Patch Selection strategy tailored to accommodate the fixed input size of CLIP, ensuring that potential forgery traces are retained. We then project the visual features extracted from CLIP's image encoder onto the NULL-Space of the semantic information, thereby effectively removing irrelevant or confounding semantics. The

resulting task-specific features are subsequently leveraged in a contrastive learning setup, which not only guides the linear classification layer but also models the intrinsic distributional differences between real and AI-generated images, beyond simple categorical prediction. The following sections elaborate on the proposed methodology in detail.

Patch Selection

Inspired by SAFE (Li et al. 2025b), which highlights that the resizing operations commonly used during training and testing can suppress artifact information, we incorporate a cropping strategy into the image pre-processing pipeline. However, conventional center cropping may inadvertently discard crucial forgery traces—especially when manipulated regions lie away from the image center. To address these issues, we propose the Patch Selection strategy that preserves task-relevant regions while adapting images to the fixed input size required by the network.

Specifically, we first divide the image into multiple non-overlapping patches using a fixed window size, denoted as $I = \{x_1, x_2, ..., x_n\}, x_i \in \mathbb{R}^{N\times N\times 3}$. In this work, we set the patch size to $N = 32$. Each patch is then transformed into the frequency domain using the Fast Fourier Transform (FFT), resulting in $I_f = \{x_1^f, x_2^f, ..., x_n^f\}, x_i^f \in \mathbb{R}^{N\times N\times 3}$. We compute the spectral entropy of each patch to quantify its texture complexity and sort all patches based on this value in descending order. Given the network's required input image size $M \times M$, we determine the number of patches to be selected as $2K = (\frac{M}{N})^2$. We then select the top K patches with the highest spectral entropy and the bottom K patches with the lowest spectral entropy. These patches are randomly shuffled and reconstructed into a new image $X_{\text{reconstruct}} \in \mathbb{R}^{M\times M\times 3}$ to match the input requirements. The proposed patch selection strategy not only ensures compatibility with networks requiring fixed-size inputs, but also preserves both the texture-rich and texture-poor regions, thereby reducing the loss of critical artifact information. Moreover, the shuffling process disrupts the original spatial semantics of the image, helping the model focus on subtle artifacts rather than high-level semantic content—thus enhancing its ability to detect forgeries.

NULL-Space Decoupling

Preliminarily, the NULL-Space of a given matrix A, denoted as NULL-Space(A) , is defined as:

$$\text{NULL-Space}(A) = \{X : AX = 0\}. \tag{1}$$

We also define $\text{Proj}_S(f)$ as the projection of a feature vector f onto a subspace S. In particular, if S is the NULL-Space associated with f, such as the orthogonal complement of f, then the projection satisfies $\text{Proj}_S(f) = 0$.

Analogously, our primary objective is to construct the NULL-Space corresponding to the semantic information embedded within the visual features extracted by CLIP's image encoder. Specifically, we use the CLIP's image encoder to extract visual features U from each image, which inherently encode both semantic information U_s and artifact-related information U_t, as discussed in Section I. However, the entanglement and complex interactions between U_s and U_t hinder the direct construction of the NULL-space X_{U_s} associated with semantic content alone. Given that CLIP is pre-trained for image-text alignment, the textual features, extracted by CLIP's text encoder from the corresponding image description, should ideally align with the semantic component U_s, *i.e.*, $V \approx U_s$. Consequently, we shift our goal to constructing the NULL-Space X_V of the textual features V, and then projecting the visual features U onto X_V in order to suppress semantic content and achieve effective disentanglement. During training, we freeze both CLIP's image and text encoders and introduce LoRA (Low-Rank Adaptation) layers to fine-tune the image encoder. This design improves the encoder's ability to emphasize artifact-related cues relevant for distinguishing AI-generated images.

To begin, we consider a dataset D consisting of both real and fake images, defined as:

$$D = \{\text{img}_j, \text{label}_j\}_{j=1}^N, \text{label} \in \{0, 1\}, \tag{2}$$

where label $= 1$ denotes a fake image, and label $= 0$ denotes a real image. For each image img_j, we employ BLIP (Li et al. 2022) to generate a corresponding text description text_j. This yields an enriched dataset with paired visual and textual information, defined as:

$$\tilde{D} = \{\text{img}_j, \text{text}_j, \text{label}_j\}_{j=1}^N, \text{label} \in \{0, 1\}. \tag{3}$$

Next, we extract visual features U using CLIP's image encoder and semantic features V using the text encoder, as follows:

$$\begin{aligned} U = U_t + U_s &= \text{E}_\text{I}^\text{lora}(\text{img}), \\ V &= \text{E}_\text{T}(\text{text}), \end{aligned} \tag{4}$$

where $\text{E}_\text{I}^\text{lora}$ denotes the image encoder augmented with LoRA layers, and E_T represents the original text encoder. Since V serves as a semantic approximation of U_s, we substitute V in place of U_s and compute the NULL-space X_V of the textual features V:

$$\text{NULL-Space}(V) = \{X_V : VX_V = 0\}. \tag{5}$$

Specifically, we first apply Singular Value Decomposition (SVD) (Wang et al. 2021) to V:

$$\{M, \Sigma, N^\mathsf{T}\} = \text{SVD}\{V\}, \tag{6}$$

where M and N are orthogonal matrices containing the left and right singular vectors, respectively, and Σ is a diagonal matrix with singular values on the diagonal. We remove the columns of N corresponding to non-zero singular values in Σ, and denote the remaining submatrix as $\tilde{N}$, which spans the NULL-Space of V. Based on this, we define the projection matrix P_V as:

$$P_V = \tilde{N}(\tilde{N})^\mathsf{T}. \tag{7}$$

The NULL-space of V is then given by:

$$\{X_V : X_V = P_V \cdot v,\ v \in \mathbb{R}^n\}, \tag{8}$$

indicating that the projection matrix P_V maps any input vector v into the NULL-space X_V. Notably, projecting the original text features V yields zero, *i.e.*,

$$\text{Proj}_{X_V}(V) = VP_V = 0. \tag{9}$$

We then project the visual features U onto the NULL-space X_V of the text features:

$$\text{Proj}_{X_V}(U) = UP_V. \tag{10}$$

This projection effectively suppresses the semantic content aligned with the text modality. Importantly, linear operations such as projection onto a NULL-space can transform the feature space while preserving orthogonal (i.e., non-semantic) information. Therefore, $\text{Proj}_{X_V}(U)$ achieves semantic disentanglement by filtering out semantic components, while retaining information that is potentially discriminative for forgery detection.

Contrastive Learning and Classification

We believe that direct binary classification will cause the network to overfit to a specific feature structure, which will inhibit the generalization effect of the detector. In order to make the detector more focused on learning the difference between the features of real and generated images, we use contrastive loss (Chen et al. 2020) $\mathcal{L}_{\text{contrastive}}$ as the dominant and linear classifier loss $\mathcal{L}_{\text{BCE}}$ as the auxiliary for training.

Specifically, for the decoupled features, we define features with the same label as positive samples and features with different labels as negative. We hope that positive samples are closer, while negative samples are gradually farther away. Thus, let $i \in I \equiv \{1...N\}$ be the index of the batch sample, where N is the batchsize. The self-supervised contrastive loss of decoupled features can be formulated as follow:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \left(\frac{\exp(f_i \cdot f_p/\tau)}{\sum_{j=1, j\neq i}^{N} \exp(f_i \cdot f_j/\tau)} \right), \tag{11}$$

where f are the decoupled features, the index i called the anchor, the index p called the positive, $\mathcal{P}_i = \{j \mid \text{label}_i = \text{label}_j,\ j \neq i\}$ is the set of positive samples, τ is a temperature hyperparameter and $|\cdot|$ denotes the number of vectors.

The final loss function is obtained by the weighted sum of the above loss functions as follows:

$$\mathcal{L} = (1-\lambda) * \mathcal{L}_{\text{contrastive}} + \lambda * \mathcal{L}_{\text{BCE}}, \tag{12}$$

where λ is the hyper-parameter for balancing two losses.

Experiments

Evaluation Setup

Training datasets. We utilize the dataset from AIGIBench (Li et al. 2025d), specifically adopting **Setting-II**, which involves training on 144k images generated by ProGAN (Karras et al. 2018) and SDv1.4 (Rombach et al. 2022). These images span four object categories: car, cat, chair, and horse. In contrast to prior approaches that trained on images from a single generation technique, such as ProGAN or SDv1.4, our method incorporates multiple known generation techniques into the training dataset, aiming to improve the model's generalization across various generation methods.

Testing datasets. To comprehensively evaluate the effectiveness of our method, we conduct generalization experiments on three different datasets, including GenImage (Zhu et al. 2024) (with deepfakes from the ForenSynths (Wang et al. 2020) dataset added), UniversalFakeDetect (Ojha, Li, and Lee 2023), and AIGIBench (Li et al. 2025d). The test set comprises 19 subsets from various generative models, including: ProGAN (Karras et al. 2018), StyleGAN, StyleGAN2, StyleGAN3 (Karras et al. 2021), StyleGAN-XL (Sauer, Schwarz, and Geiger 2022), StyleSwim (Zhang et al. 2022), BigGAN (Brock 2018), CycleGAN (Zhu et al. 2017), StarGAN (Choi et al. 2018), GauGAN (Park et al. 2019), R3GAN (Huang et al. 2024), Deepfake (Rossler et al. 2019), WFIR, LDM (Rombach et al. 2022), Midjourney, SDv1.4 (Rombach et al. 2022), SDv1.5, SD-3, SD-XL, ADM (Dhariwal and Nichol 2021), GLIDE (Nichol et al. 2021), Wukong, VQDM, Midjourney-v6, FLUX.1-dev, Imagen-3, Infinite-ID (Wu et al. 2024b), BLIP (Li et al. 2022), IP-Adapter (Ye et al. 2023), and 2 general subsets from open-source platforms, which called CommunityAI and SocialRF.

Evaluation Metrics. Following the established evaluation paradigm (Li et al. 2025d), we adopt the accuracy of detecting fake images (F.Acc.) as the primary evaluation metrics. Additionally, we calculate the accuracy of detecting real images (R.Acc.) at the beginning of the table to provide a complete assessment of the model's performance. Additional metric about A.P. is shown in Appendix II.

Comparison Baselines. We compare our method with existing state-of-the-art detection methods, including CNN-Spot (CVPR 2020) (Wang et al. 2020), UnivFD (CVPR 2023) (Ojha, Li, and Lee 2023), FreqNet (AAAI 2024) (Tan et al. 2024a), NPR (CVPR 2024) (Tan et al. 2024b), Ladeda (arXiv 2024) (Cavia et al. 2024), AIDE (ICLR 2025) (Yan et al. 2024), DFFreq (arXiv 2025) (Yan et al. 2025), SAFE (KDD 2025) (Li et al. 2025b), and VIB-Net (CVPR 2025) (Zhang et al. 2025). More details are shown in Appendix I. To ensure a fair comparison, we retrain all methods on the AIGIBench training set.

Implementation Details. We use the Adam optimizer (Kingma 2014) with a learning rate of 2×10^{-4}. The batchsize is set to 32, and we train the model for only 2 epoches. The ViT-L/14 model of CLIP is adopted as the pre-trained model following the baseline UnivFD (Ojha, Li, and Lee 2023). In addition, we add Low-Rank Adaptation to fine-tune the CLIP's image encoder, following (Zanella and Ben Ayed 2024). For the setting of the LoRA layers, we configure the hyperparameters as follows: *lora_r = 6*, *lora_alpha = 6*, and *lora_dropout = 0.8*. The hyperparameters λ is set to 0.2. The proposed method is implemented using PyTorch (Paszke et al. 2019) on an NVIDIA RTX 4090 GPU.

Comparison with the State-of-the-Art

Performance on GenImage. As shown in Table 1, our method demonstrates superior cross-domain generalization on GAN- and Diffusion-generated images, achieving an average accuracy of 93.2%. While its accuracy on real images (97.7%) is slightly lower than some methods (e.g., LadeDA at 99.8%), it surpasses the state-of-the-art AIDE (Yan et al. 2024) by 7.6% in mean accuracy (mAcc.). All methods achieve near-perfect detection (e.g., 100% on ProGAN and SDv1.4) on in-domain datasets; however, their performance degrades significantly on out-of-domain datasets. In con-

Method	Real Image	Generative Adversarial Networks							Other
		ProGAN	CycleGAN	BigGAN	StyleGAN	StyleGAN2	GauGAN	StarGAN	WFIR
CNN-Spot	99.0	95.3	18.7	1.8	36.5	22.0	2.5	23.1	1.1
UnivFD	92.3	98.9	90.5	79.2	55.7	48.7	91.1	96.9	92.8
FreqNet	89.9	99.4	57.1	51.0	75.1	67.5	9.9	88.4	95.4
NPR	99.3	98.9	29.3	16.5	67.1	58.7	1.7	6.5	7.9
Ladeda	**99.8**	99.7	7.2	29.0	95.6	98.3	4.9	0.0	19.2
AIDE	93.0	95.3	88.0	96.9	89.6	97.0	90.0	97.0	42.9
DFFreq	98.0	98.0	70.9	93.6	**99.0**	**98.9**	93.8	97.4	3.5
SAFE	99.2	99.7	85.6	83.5	88.1	96.7	89.0	99.9	8.4
VIB-Net	93.5	99.4	97.3	91.3	71.3	73.3	97.8	97.3	**97.3**
Ours	97.7	**100.0**	**99.9**	**100.0**	95.7	98.6	**98.5**	**100.0**	68.9

Method	Other	Diffusion Models								mAcc.
	Deepfake	SDv1.4	SDv1.5	ADM	GLIDE	Midjourney	Wukong	VQDM	DALLE2	
CNN-Spot	29.3	55.9	55.6	1.8	4.8	5.2	27.6	0.7	4.5	29.0
UnivFD	26.9	96.3	96.0	12.7	75.6	61.2	84.7	45.6	62.3	72.7
FreqNet	35.8	99.9	99.8	37.7	78.9	80.8	98.0	34.1	88.8	71.7
NPR	0.1	100.0	99.9	26.5	69.2	71.0	97.7	15.4	89.8	53.1
Ladeda	0.1	100.0	99.9	27.3	79.7	88.8	97.9	15.5	92.4	58.6
AIDE	9.2	99.8	99.7	**92.4**	**98.7**	63.7	98.9	90.1	**98.9**	85.6
DFFreq	76.0	99.9	99.8	65.9	87.8	**94.8**	99.1	76.0	95.9	85.2
SAFE	17.3	99.8	99.6	36.8	90.5	86.3	98.5	84.0	92.0	80.8
VIB-Net	**90.6**	100.0	99.8	52.8	69.3	63.7	99.3	80.9	58.5	84.7
Ours	60.6	**100.0**	**99.9**	85.3	97.2	77.4	**100.0**	**98.9**	98.8	**93.2**

Table 1: **Cross-model Accuracy (Acc.) Performance on the GenImage (Zhu et al. 2024) Dataset.** The first column represents the accuracy of detecting real images (R.Acc.), and the others are the accuracy of detecting fake images (F.Acc.).

Method	Guided	Glide_50_27	Glide_100_10	Glide_100_27	LDM_100	LDM_200	LDM_200_cfg	DALLE	mAcc.
CNN-Spot	3.0	4.6	4.6	5.1	21.4	24.1	36.1	3.5	12.8
UnivFD	35.5	79.1	80.0	79.6	75.9	76.0	57.7	49.1	66.4
FreqNet	40.9	80.4	81.1	78.2	96.6	97.6	96.1	40.1	76.4
NPR	29.5	61.9	60.6	59.9	76.1	73.9	89.3	27.3	59.8
Ladeda	39.5	69.4	71.6	72.4	74.3	75.7	85.3	9.2	62.2
AIDE	81.7	**97.3**	96.5	**97.0**	98.6	98.4	99.7	96.1	95.6
DFFreq	80.7	90.7	92.5	88.6	99.9	99.8	99.7	91.4	92.9
SAFE	45.5	92.9	94.7	90.2	99.8	99.9	99.7	96.1	89.9
VIB-Net	58.1	75.5	76.6	72.4	97.1	97.0	86.5	93.9	82.1
Ours	**85.3**	96.7	**98.0**	96.4	**100.0**	**100.0**	**100.0**	**100.0**	**97.0**

Table 2: **Cross-model Fake Accuracy(F.Acc.) Performance on the UniversalFakeDetect (Ojha, Li, and Lee 2023) Dataset.**

trast, our method maintains high accuracy across most cross-domain datasets, such as 100.0% on BigGAN and 98.9% on VQDM. This generalizability highlights the effectiveness of our approach in decoupling semantic information to focus on features critical for forgery detection.

Performance on UniversalFakeDetect. Table 2 underscores our method's capability in detecting unknown Diffusion-based fake images, achieving an average accuracy of 97.0%. It attains perfect detection (100%) on LDM_200, LDM_200_cfg, and DALLE, outperforming competing methods such as AIDE (95.6%) and DFFreq (92.9%). This consistent high performance across diverse diffusion models demonstrates the method's strong generalization ability.

Performance on AIGIBench. Given the rapid evolution of generative techniques, existing detection datasets often fall short in evaluating contemporary AI-generated image detection methods. We address this by utilizing the AIGIBench dataset (Li et al. 2025d), which includes a wide range of advanced generative models (Table 3). Our method still maintains the highest detection accuracy, with the improvement of 2.8% compared with the state-of-the-art methods. It significantly outperforms UnivFD by over 10% on most datasets, with a 21.4% improvement on GLIDE, owing to its ability to eliminate semantic information from CLIP's upstream

Generator	CNN-Spot	UnivFD	FreqNet	NPR	Ladeda	AIDE	DFFreq	SAFE	VIB-Net	Ours
Real Image	**98.0**	73.2	65.8	93.8	97.1	88.1	91.7	92.2	60.8	88.0
R3GAN	2.3	94.1	59.9	8.4	19.5	**99.0**	78.4	93.0	79.3	98.5
StyleGAN3	9.1	82.6	98.2	63.6	93.2	91.1	95.5	94.6	89.8	**99.6**
StyleGAN-XL	0.7	96.7	95.5	28.2	80.5	91.7	15.6	89.4	**99.8**	99.3
StyleSwim	6.9	98.1	97.1	77.7	97.3	82.0	99.8	**99.9**	98.4	99.8
FLUX1-dev	16.3	86.6	92.4	97.2	99.3	90.0	96.1	**99.5**	60.5	96.2
Midjourney-V6	5.8	80.6	83.6	53.8	83.4	79.8	**95.8**	94.1	80.2	94.8
GLIDE	4.6	75.2	79.7	70.3	81.8	**98.4**	86.9	89.2	69.2	96.6
Imagen3	4.2	84.2	81.5	78.2	92.6	93.9	51.9	94.8	82.5	**99.5**
SD3	13.3	90.6	88.1	89.7	99.0	99.3	92.1	94.4	82.1	**99.9**
SDXL	7.2	88.0	98.9	79.0	98.3	97.6	98.0	99.9	89.3	**99.9**
BLIP	56.5	92.1	100.0	99.9	100.0	100.0	100.0	100.0	99.6	**100.0**
Infinite-ID	1.1	93.8	92.7	34.6	32.2	97.5	93.9	**99.2**	72.4	92.2
IP-Adapter	6.0	92.0	92.0	71.8	90.6	93.5	97.8	97.2	93.1	**99.8**
SocialRF	7.5	**55.5**	39.3	21.9	19.4	18.4	18.4	17.1	27.4	18.4
CommunityAI	5.4	**51.2**	12.2	8.2	9.0	9.3	9.2	9.1	13.8	8.9
Average	15.3	83.4	79.8	61.0	74.5	83.1	76.3	85.2	74.9	**87.0**

Table 3: **Cross-model Accuracy (Acc.) Performance on the AIGIBench (Li et al. 2025d) Dataset.** The first row represents the accuracy of detecting real images (R.Acc.), and the others are the accuracy of detecting fake images (F.Acc.).

training while retaining fine-grained features essential for detection. However, performance on SocialRF (18.4%) and CommunityAI (8.9%) remains low, consistent with other methods, likely due to the heterogeneous and unknown generative methods in these datasets, suggesting further research.

Ablation Studies

To thoroughly assess the effectiveness of our proposed method for AI-generated image detection, we conducted an ablation study evaluating the contributions of three key modules: NULL-Space Decoupling, Patch Selection, and Contrastive Learning. The results, illustrated in Table 4, quantify the impact of each module individually and in combination, with mean Acc. serving as the primary metric.

Baseline Performance. The baseline CLIP model, which adds Lora for fine-tuning only, achieves a mean Acc. of 81.2%, serving as the reference point for evaluating the contributions of each module.

NULL-Space Decoupling. Adding only NULL-Space Decoupling to the baseline increases the mean Acc. to 85.3%, a 4.1% improvement. This enhancement, achieved without additional training, suggests that NULL-Space Decoupling effectively removes task-irrelevant semantic information from the feature space, allowing the model to focus on features more pertinent to distinguishing real and fake images. Its significance is further underscored when removed from the full model (mean Acc. 93.2%), resulting in a substantial 5.7% drop to 87.5%, which indicates the critical role in maintaining high detection performance.

Patch Selection & Contrastive Learning. Incorporating only Patch Selection improves the mean Acc. to 82.8%, a 1.6% increase over the baseline, which reflects Patch Selection's ability to retain artifact information while mitigating the influence of irrelevant semantics. What's more, Contrastive Learning improves detection accuracy by 3%, indicating that Contrastive Learning effectively encourages the model to learn the distributional differences between real and fake image features, enhancing its discriminative ability.

Patch Selection	Null-Space Decoupling	Contrastive Learning	mean Acc.
✗	✗	✗	81.2
✓	✗	✗	82.8
✗	✓	✗	85.3
✗	✗	✓	84.2
✓	✗	✓	87.5
✓	✓	✓	93.2

Table 4: **Ablation Study on the GenImage (Zhu et al. 2024) Dataset.**

Plug & Play Application In order to further verify the wide application capability of our method, we applied the NULL-Space Decoupling module and the Patch Slelection module to the existing detection framework. The results are shown in Table 7 and Table 8 of the Appendix. They all improve the detection ability of the model. More details can be found in Appendix IV.

Conclusion

In this paper, we we revisit the distinctions between features with different intrinsic properties extracted by CLIP, and innovatively proposed that the semantic information embedded in the visual features extracted by CLIP seriously interferes with the detector's extraction of artifact features. Accordingly, we proposed an innovative method, NS-Net. Specifically, we use the feature homogeneity extracted by the text encoder to replace the semantic information of the features extracted by the image encoder, and use NULL-Space to decouple

the semantic information, retaining the artifact information related to the forgery detection task. At the same time, we proposed Patch Selection and used contrastive learning to retain artifact information and promote the model to learn the underlying distributional differences between real and generated images, thereby jointly improving the generalization effect of the model. Extensive experiments on 40 diverse generative models strongly demonstrate the generalization capability of our method.

References

Baraldi, L.; Cocchi, F.; Cornia, M.; Nicolosi, A.; and Cucchiara, R. 2025. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *European Conference on Computer Vision*, 199–216. Springer.

Brock, A. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Cavia, B.; Horwitz, E.; Reiss, T.; and Hoshen, Y. 2024. Real-Time Deepfake Detection in the Real-World. *arXiv preprint arXiv:2406.09398*.

Chai, L.; Bau, D.; Lim, S.-N.; and Isola, P. 2020. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 103–120. Springer.

Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; and Wang, J. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18710–18719.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, Z.; and Yang, H. 2020. Manipulated face detector: Joint spatial and frequency domain attention network. *arXiv preprint arXiv:2005.02958*, 1(2): 4.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fang, J.; Jiang, H.; Wang, K.; Ma, Y.; Jie, S.; Wang, X.; He, X.; and Chua, T.-S. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355*.

Farid, H. 2022. Lighting (in) consistency of paint by text. *arXiv preprint arXiv:2207.13744*.

Fu, X.; Yan, Z.; Yao, T.; Chen, S.; and Li, X. 2025. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3040–3048.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5039–5049.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, N.; Gokaslan, A.; Kuleshov, V.; and Tompkin, J. 2024. The GAN is dead; long live the GAN! A Modern GAN Baseline. *Advances in Neural Information Processing Systems*, 37: 44177–44215.

Karras, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv preprint arXiv:1812.04948*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34: 852–863.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, M.; Tao, R.; Liu, Y.; Tan, C.; Qin, H.; Li, B.; Wei, Y.; and Zhao, Y. 2025a. Pay Less Attention to Deceptive Artifacts: Robust Detection of Compressed Deepfakes on Online Social Networks. *arXiv preprint arXiv:2506.20548*.

Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Feng, F. 2025b. Improving Synthetic Image Detection Towards Generalization: An Image Transformation Perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2405–2414.

Li, Y.; Shao, S.; He, Y.; Guo, J.; Zhang, T.; Qin, Z.; Chen, P.-Y.; Backes, M.; Torr, P.; Tao, D.; et al. 2025c. Rethinking data protection in the (generative) artificial intelligence era. *arXiv preprint arXiv:2507.03034*.

Li, Z.; Yan, J.; He, Z.; Zeng, K.; Jiang, W.; Xiong, L.; and Fu, Z. 2025d. Is Artificial Intelligence Generated Image Detection a Solved Problem? *arXiv preprint arXiv:2505.12335*.

Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024a. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 10770–10780.

Liu, Z.; Wang, H.; Kang, Y.; and Wang, S. 2024b. Mixture of low-rank experts for transferable ai-generated image detection. *arXiv preprint arXiv:2404.04883*.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16317–16326.

McCloskey, S.; and Albright, M. 2018. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Sauer, A.; Schwarz, K.; and Geiger, A. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tan, C.; Tao, R.; Liu, H.; Gu, G.; Wu, B.; Zhao, Y.; and Wei, Y. 2025. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7184–7192.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024a. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5052–5060.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

Tao, R.; Tan, C.; Liu, H.; Wang, J.; Qin, H.; Chang, Y.; Wang, W.; Ni, R.; and Zhao, Y. 2025. SAGNet: Decoupling Semantic-Agnostic Artifacts from Limited Training Data for Robust Generalization in Deepfake Detection. *IEEE Transactions on Information Forensics and Security*.

Wang, S.; Li, X.; Sun, J.; and Xu, Z. 2021. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 184–193.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

Wu, J.; Zhu, Y.; Jiang, X.; Liu, Y.; and Lin, J. 2024a. Local attention and long-distance interaction of rPPG for deepfake detection. *The Visual Computer*, 40(2): 1083–1094.

Wu, Y.; Li, Z.; Zheng, H.; Wang, C.; and Li, B. 2024b. Infinite-ID: Identity-preserved Personalization via ID-semantics Decoupling Paradigm. In *European Conference on Computer Vision*, 279–296. Springer.

Yan, J.; Li, Z.; He, Z.; and Fu, Z. 2025. Generalizable Deepfake Detection via Effective Local-Global Feature Extraction. *arXiv preprint arXiv:2501.15253*.

Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2024. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zanella, M.; and Ben Ayed, I. 2024. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1593–1603.

Zhang, B.; Gu, S.; Zhang, B.; Bao, J.; Chen, D.; Wen, F.; Wang, Y.; and Guo, B. 2022. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11304–11314.

Zhang, H.; He, Q.; Bi, X.; Li, W.; Liu, B.; and Xiao, B. 2025. Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network. In *Proceedings*

of the Computer Vision and Pattern Recognition Conference, 23828–23837.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2185–2194.

Zheng, C.; Lin, C.; Zhao, Z.; Wang, H.; Guo, X.; Liu, S.; and Shen, C. 2024. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems*, 37: 59570–59596.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2024. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36.

Appendix

Details of Detection Models for Comparison

In order to effectively evaluate the advancement and generalization of our proposed method in the field of AI-generated detection, we selected several state-of-the-art detection methods from recent years. The following provide detailed descriptions of these detection approaches:

1) CNN-Spot (CVPR 2020) (Wang et al. 2020). CNN-Spot uses CNN to identify synthetic content by analyzing common spatial artifacts in AI-generated images. It extracts hierarchical features from raw pixel data by stacking convolutional layers, effectively capturing generation anomalies.

2) UnivFD (CVPR 2023) (Ojha, Li, and Lee 2023). UnivFD demonstrates that CLIP effectively extracts artifacts from images. By training a classifier on these features, they achieve strong generalization performance.

3) FreqNet (AAAI 2024) (Tan et al. 2024a). FreqNet isolates high-frequency components of each image using an FFT-based high-pass filter, and introduces a plug-in frequency-domain learning block that transforms intermediate feature maps via FFT, applies learnable magnitude and phase transformations, and then performs an inverse FFT (iFFT), enabling optimization directly in the frequency domain.

4) NPR (CVPR 2024) (Tan et al. 2024b). NPR targets the universal structural artifacts introduced by up-sampling layers in generative models. The method transforms each input image into NPR maps to capture signed intensity differences between each pixel and its four immediate neighbors. These maps make local pixel-dependency patterns explicit, revealing artifacts characteristic of synthetic up-sampling operations.

5) Ladeda (arxiv 2024) (Cavia et al. 2024). LaDeDa is a patch-level deepfake detector that partitions each input image into 9×9 pixel patches and processes them using a BagNet-style ResNet-50 variant with its receptive field constrained to the same 9×9 region. The model assigns a deepfake likelihood to each patch, and the final prediction is obtained by globally pooling the patch-level scores.

6) AIDE (ICLR 2025) (Yan et al. 2024). AIDE simultaneously incorporates low-level patch statistics and high-level semantics for AI-generated image detection. It employs two expert branches: i) a Semantic Feature Extractor, which utilizes CLIP-ConvNeXt embeddings to detect high-level content inconsistencies, and ii) a Patchwise Feature Extractor, which ranks image patches by spectral energy, selects the highest- and lowest-frequency regions, and applies a lightweight CNN to capture fine-grained noise and artifact patterns.

7) DFFreq (arxiv 2025) (Yan et al. 2025). DFFreq first utilizea a sliding window to restrict the attention mechanism to a local window, and reconstruct the features within the window to model the relationships between neighboring internal elements within the local region. Then, it designs a dual frequency domain branch framework consisting of four frequency domain subbands of DWT and the phase part of FFT to enrich the extraction of local forgery features from different perspectives.

8) SAFE (KDD 2025) (Li et al. 2025b). SAFE replaces conventional resizing with random cropping to better preserve high-frequency details, applies data augmentations such as Color-Jitter and RandomRotation to break correlations tied to color and layout, and introduces patch-level random masking to encourage the model to focus on localized regions where synthetic pixel correlations typically emerge.

9) VIB-Net (CVPR 2025) (Zhang et al. 2025). VIB-Net finds that the general features extracted by current methods based on large-scale pre-trained models contain irrelevant features that are unrelated to the task of distinguishing real from fake images, and proposes VIB-Net, which uses Variational Information Bottlenecks to enforce authentication task-related feature learning.

Additional Metrics

In order to further study the effectiveness of our method and to be consistent with other research evaluation paradigms, we additionally use average precision as our evaluation metric. Average Precision (AP) provides a comprehensive evaluation of a model's performance by integrating precision and recall across all decision thresholds, making it particularly effective for imbalanced datasets where the minority class is of primary interest. The specific results are shown in Table 5 and 6. Specifically, our method achieves the average precision of 99.2% and 99.9% on GenImage and UniversalFakeDetect respectively, surpassing the best methods 1.1% and 0.2%, further verifying the excellent detection effect and generalization ability of our proposed method.

Qualitative Analysis

To further assess the generalization ability of our method, we visualize the feature distribution before binary classification, as shown in Table 3. This visualization highlights how effectively each trained model distinguishes between real and fake images, showcasing our method's ability to generalize across diverse fake representations. From the visualization, it is evident that VIB-Net struggles with unseen GAN or diffusion models, as the feature distributions of real and generated images have a high overlap, indicating that the model often misclassifies fake images as "real". In contrast, our method

Method	Generative Adversarial Networks							Other	
	ProGAN	CycleGAN	BigGAN	StyleGAN	StyleGAN2	GauGAN	StarGAN	WFIR	Deepfake
CNN-Spot	99.9	91.0	59.2	94.8	94.2	86.1	82.5	65.5	74.2
UnivFD	99.9	90.5	79.2	55.7	48.7	91.1	96.9	92.8	26.9
FreqNet	99.4	97.5	87.9	95.9	94.2	54.0	100.0	57.0	88.5
NPR	100.0	98.6	79.4	98.6	99.6	71.3	97.6	65.5	92.5
Ladeda	100.0	91.5	93.2	100.0	100.0	81.8	93.0	86.9	72.8
AIDE	99.6	98.2	91.6	99.4	99.9	74.1	99.7	90.8	66.5
DFFreq	99.7	99.0	98.4	**100.0**	100.0	98.1	99.8	89.0	87.5
SAFE	100.0	99.5	97.8	99.9	**100.0**	96.6	100.0	73.8	91.0
VIB-Net	99.9	99.7	97.1	93.9	98.1	**99.2**	99.7	95.7	83.6
Ours	**100.0**	**100.0**	**99.7**	99.8	99.9	97.0	**100.0**	**99.5**	**95.2**

Method	Diffusion Models								mA.P.
	SDv1.4	SDv1.5	ADM	GLIDE	Midjourney	Wukong	VQDM	DALLE2	
CNN-Spot	97.6	97.6	66.6	83.1	80.4	88.9	57.0	87.8	84.3
UnivFD	96.3	96.0	65.2	95.3	91.7	97.4	87.2	62.3	91.9
FreqNet	99.9	99.9	73.9	94.1	94.4	99.1	74.9	92.3	88.4
NPR	100.0	99.9	74.1	97.4	97.8	99.9	81.5	99.3	91.4
Ladeda	100.0	99.9	77.3	99.1	99.7	100.0	89.3	99.8	93.2
AIDE	100.0	99.9	99.4	99.9	90.3	99.9	99.3	99.9	94.5
DFFreq	100.0	99.9	98.7	99.5	99.7	100.0	99.3	99.9	98.1
SAFE	100.0	100.0	98.1	**99.9**	**99.9**	100.0	99.9	100.0	97.4
VIB-Net	100.0	99.9	91.2	96.4	94.3	99.9	97.4	94.6	96.5
Ours	**100.0**	**100.0**	**100.0**	99.8	97.0	**100.0**	**100.0**	**100.0**	**99.2**

Table 5: **Cross-model Average Precision (A.P.) Performance on the GenImage (Zhu et al. 2024) Dataset.**

Method	Guided	Glide_50_27	Glide_100_10	Glide_100_27	LDM_100	LDM_200	LDM_200_cfg	DALLE	mA.P.
CNN-Spot	70.8	76.9	78.0	76.8	88.7	90.1	93.8	62.7	79.7
UnivFD	81.0	92.5	92.1	92.3	91.4	91.4	84.7	80.3	88.2
FreqNet	78.5	97.8	97.9	97.6	99.7	99.8	99.7	93.1	95.5
NPR	83.0	99.3	99.0	99.3	99.9	99.9	100.0	98.9	97.4
Ladeda	89.0	99.6	99.4	99.7	99.9	99.9	99.9	96.4	98.0
AIDE	98.5	99.9	99.8	99.8	99.9	100.0	100.0	99.5	99.6
DFFreq	99.2	99.2	99.2	99.0	100.0	100.0	100.0	99.5	99.5
SAFE	98.9	99.6	99.7	99.4	100.0	100.0	100.0	99.9	99.7
VIB-Net	94.2	97.2	96.8	96.9	99.9	99.8	98.6	99.5	97.9
Ours	**99.4**	**99.9**	**99.9**	**99.8**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**

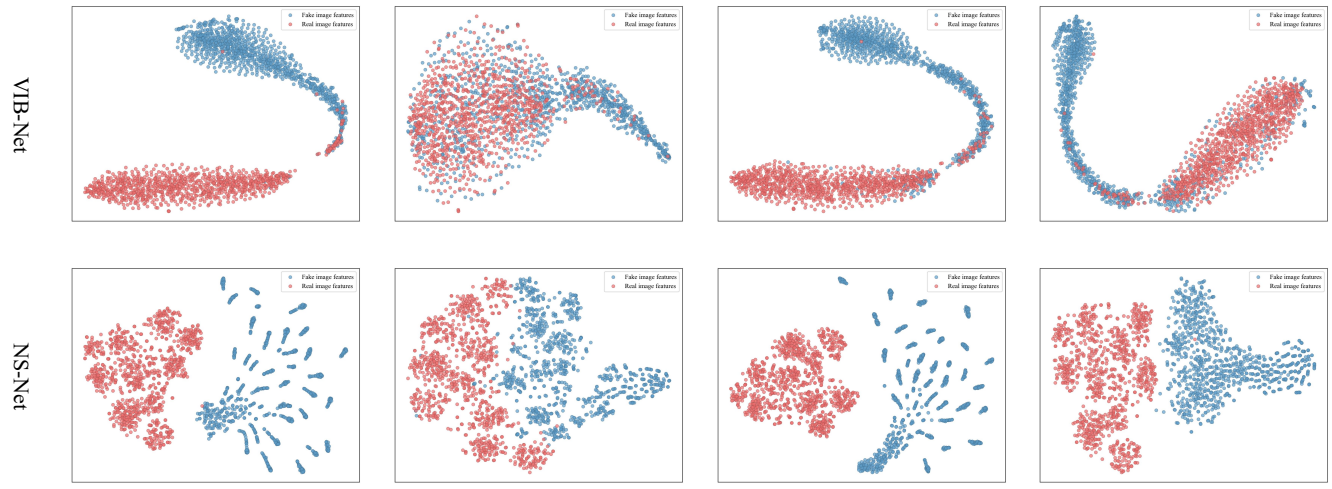
Table 6: **Cross-model Average Precision (A.P.) Performance on the UniversalFakeDetect (Ojha, Li, and Lee 2023) Dataset.**

demonstrates superior discrimination, effectively separating "real" and "fake" categories, even when encountering unseen sources.

Plug & Play Application

NULL-Space Decoupling. To evaluate the broad applicability of our NULL-Space decoupling module within CLIP-based feature extraction frameworks, we integrated it into the established detection models UnivFD and VIB-Net. As shown in Table 7, our NULL-Space decoupling module effectively improves the detection performance. Specifically, our decoupling module improves the average accuracy by 13.9% and 2.2% on UnivFD and VIB-Net, respectively, indicating that the features extracted by CLIP contain rich semantic information, and our reasonable decoupling of semantic information helps the model focus more on features related to the detection task.

Patch Selection. We also performed a plug-and-play evaluation on our data preprocessing operation Parch Selection. The comparison results, illustrated in Table 8, demonstrate a consistent improvement in detecting synthetic images from various generative models, including the proposer of using crop during detection SAFE (Li et al. 2025b). The results demonstrate that Patch Selection enables the detector to extract more generalizable features indicative of artifacts, while mitigating the adverse effects of semantic content on artifact detection. This approach enhances the detector's ability to identify subtle artifacts in input samples and improves its



(a) SDXL (b) FLUX (c) R3GAN (d) Guided

Figure 3: **T-SNE Visualization of Features Extracted before Classifier**. We compare the VIB-Net and our NS-Net. A total of four testing GANs and diffusion models are considered, including SDXL, FLUX, R3GAN, and Guided.

Method	SDv1.4	SDv1.5	ADM	GLIDE	Midjourney	Wukong	VQDM	DALLE2	mAcc.
UnivFD (Ojha, Li, and Lee 2023)	96.3	96.0	12.7	75.6	61.2	84.7	45.6	62.3	66.8
+NULL-Space	97.8	97.4	59.5	74.9	65.9	95.4	81.4	73.0	**80.7**$_{+13.9}$
VIB-Net (Zhang et al. 2025)	100.0	99.8	52.8	69.3	63.7	99.3	80.9	58.5	78.0
+NULL-Space	100.0	99.8	55.1	72.3	68.0	99.5	83.3	63.3	**80.2**$_{+2.2}$
Ours w/o NULL-Space	100.0	99.9	73.1	90.7	72.3	99.3	91.0	90.4	89.6
Ours	100.0	99.9	85.3	97.2	77.4	100.0	98.9	98.8	**94.7**$_{+5.1}$

Table 7: **Plug & Play Application of NULL-Space Decoupling on Existing Detectors.** Cross-model Fake Accuracy (F.Acc.) Performance on the GenImage (Zhu et al. 2024) Dataset.

Method	SDv1.4	SDv1.5	ADM	GLIDE	Midjourney	Wukong	VQDM	DALLE2	mAcc.
DFFreq (Ojha, Li, and Lee 2023)	99.9	99.8	65.9	87.8	94.8	99.1	76.0	95.9	89.9
+Patch Selection	99.9	99.9	69.1	93.2	89.5	99.6	84.2	96.3	**91.5**$_{+1.6}$
SAFE (Li et al. 2025b)	99.8	99.6	36.8	90.5	86.3	98.5	84.0	92.0	85.9
+Patch Selection	100.0	99.8	54.1	95.3	95.8	99.4	84.4	97.7	**90.7**$_{+4.8}$
VIB-Net (Zhang et al. 2025)	100.0	99.8	52.8	69.3	63.7	99.3	80.9	58.5	78.0
+Patch Selection	100.0	99.9	73.9	85.3	66.8	99.0	83.9	79.6	**86.1**$_{+8.1}$
Ours w/o Patch Selection	100.0	99.8	74.7	78.8	54.7	99.9	94.8	92.4	86.9
Ours	100.0	99.9	85.3	97.2	77.4	100.0	98.9	98.8	**94.7**$_{+7.8}$

Table 8: **Plug & Play Application of Patch Selection on Existing Detectors.** Cross-model Fake Accuracy (F.Acc.) Performance on the GenImage (Zhu et al. 2024) Dataset.

generalization performance.