Semantic Discrepancy-aware Detector for Image Forgery Identification

Ziye Wang[1], Minghang Yu[1], Chunyan Xu[1*], Zhen Cui[2*]
[1]Nanjing University of Science and Technology, Nanjing, Jiangsu, China
[2]Beijing Normal University, Beijing, China

Abstract

*With the rapid advancement of image generation techniques, robust forgery detection has become increasingly imperative to ensure the trustworthiness of digital media. Recent research indicates that the learned semantic concepts of pre-trained models are critical for identifying fake images. However, the misalignment between the forgery and semantic concept spaces hinders the model's forgery detection performance. To address this problem, we propose a novel **S**emantic **D**iscrepancy-aware **D**etector (SDD) that leverages reconstruction learning to align the two spaces at a fine-grained visual level. By exploiting the conceptual knowledge embedded in the pre-trained vision-language model, we specifically design a semantic token sampling module to mitigate the space shifts caused by features irrelevant to both forgery traces and semantic concepts. A concept-level forgery discrepancy learning module, built upon a visual reconstruction paradigm, is proposed to strengthen the interaction between visual semantic concepts and forgery traces, effectively capturing discrepancies under the concepts' guidance. Finally, the low-level forgery feature enhancemer integrates the learned concept-level forgery discrepancies to minimize redundant forgery information. Experiments conducted on two standard image forgery datasets demonstrate the efficacy of the proposed SDD, which achieves superior results compared to existing methods. The code is available at https://github.com/wzy1111111/SDD.*

1. Introduction

With the thriving of generative AI technologies, like Generative Adversarial Networks (GANs) [14] and diffusion models [2], the images generated by these methods can easily create confusion by passing off the spurious as genuine. Therefore, it is crucial to develop a universal method for detecting fake images to mitigate the widespread dis-
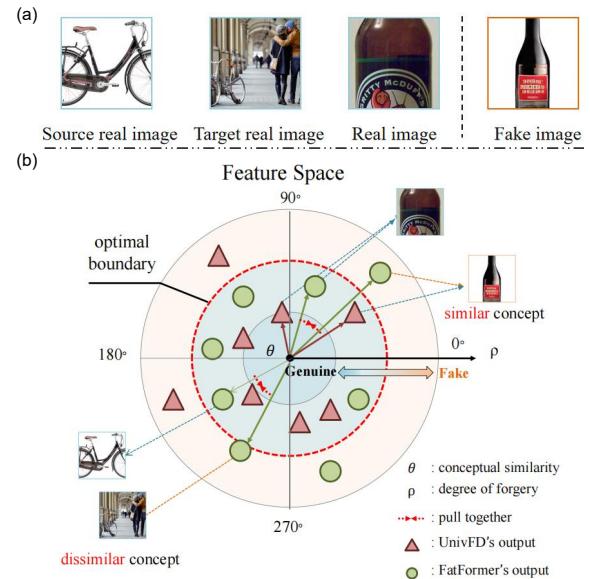

Figure 1. **The phenomenon of misalignment between semantic concept space and forgery space.** Since $\cos\theta$ can reflect the similarity of image descriptions, we model the feature space in polar coordinates. As the semantic concept space in [34] is frozen, fake samples sharing similar concepts with real ones can be easily misclassified. In the forgery-adaptive space like [26], the model can correctly distinguish between them based on re-learned forgery features. Nevertheless, due to the semantic concept bias introduced by coarse text prompts, the target samples may be projected into an inaccurate semantic concept dimension, causing them to drift away from the real source samples along the fake dimension.

semination of disinformation.

Pioneering research [26, 34] has shown that projecting images into a joint embedding space of texts and images can effectively capture discrepancies between fake and real images. In contrast, previous methods [6, 13, 46, 51] overlooking the interplay between forgery traces and semantic concepts perform poorly when confronted with unseen generative models. To investigate the visual semantic concepts of pre-trained models, we conduct a statistical analysis of the output features from CNNSpot [51] and CLIP: ViT-L/4 [34] (See Sec. A for more details). Under different categories, CNNSpot exhibits a synchronized difference be-

---

*Corresponding Authors: Z. Cui and C. Xu.
Email: {wzynjust,mhyu,cyx}@njust.edu.cn (Z. Wang, M. Yu and C. Xu), zhen.cui@bnu.edu.cn (Z. Cui).
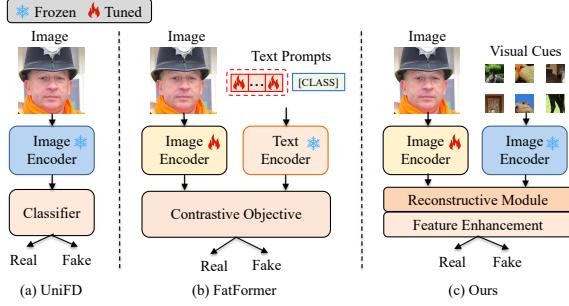
Figure 2. **Different paradigms of image forgery identification with pre-trained vision-language model.** (a) Fine-tune the frozen model only by fully connected (FC) layers [34]. (b) Prompt-based designs are tuned on text prompts and contrastive objectives [26]. (c) Our paradigm incorporating visual clues can capture fine-grained forgery traces by reconstructive learning.

tween real and fake features in its feature space. However, when transitioning to the CLIP's space, these differences become inconsistent. From this, we infer a nuanced relationship between semantic concepts and forgery traces: different semantic concepts may guide the model to uncover distinct forgery traces.

Intuitively, relying on a frozen pre-trained vision-language model like UnivFD [34] is essential to incorporate high-level semantic priors, but this tends to overlook fine-grained forgery details. Although FatFormer [26] achieves a substantial enhancement in generalization by employing the forgery-aware adaptive transformer, we observe that soft prompts based on simple [CLASS] embeddings have an intrinsic limitation in their semantic description granularity (See Sec. B for more details). The constrained breadth of the conveyed concepts may lead the detection toward incorrect predictions. This limitation highlights a misalignment between the visual semantic concept space and the target forgery space, as illustrated in Fig. 1. To address this issue, one empirical approach is to design more detailed text descriptions, but this method struggles to describe all visual details due to the limited length of texts and brings more computational overhead. Drawing from the aforementioned findings and analysis, we make a first attempt to align the CLIP's visual semantic concept space with the forgery space by reconstructing semantic features.

We develop a vision-based paradigm, as outlined in Fig. 2. First, employing a pre-trained model only with nearest neighbor or linear probing (*e.g.* UnivFD [34], Fig. 2 (a)) is suboptimal for image forgery detection. Second, modifying the pre-trained model with task-specific prompts (*e.g.* FatFormer [26] in Fig. 2 (b)) may favor models biased towards any particular semantic concept. These studies pave the way for exploring the semantic concepts space. Inspired by image reconstruction [45, 54], our paradigm amplifies the concept-level forgery discrepancies of forgery images, which empowers the model to detect suspicious forgery

traces with the assistance of semantic concepts.

In this work, we present a novel Semantic Discrepancy-aware Detector (SDD) to accurately align the semantic concept space and the forgery space. To mitigate interference from features unrelated to both semantic concepts and forgery traces, we divide the real images into non-overlapping blocks and feed them to the frozen CLIP [38] to obtain diverse semantic patch tokens. These tokens acting as visual cues smoothly align the two spaces. It is noteworthy that these tokens sampled by JS divergence are universally representative of the real semantic distribution. Then, the visual cues are fused into a concept-level forgery discrepancy module. Unlike FatFormer, LoRA layers are incorporated into the image encoder. The goal is to preserve the completeness and diversity of the learned semantic concepts of CLIP, while the forgery features sharing similar semantic concepts should be highlighted. During reconstruction, we only narrow the reconstruction gap for real samples to reinforce the reconstructed discrepancies of the synthetic images. Finally, we introduce the low-level forgery feature enhancer that leverages the reconstruction difference map to facilitate the extraction of the highly generalizable forgery features, while incurring minimal additional parameters. The main challenge is how to capture forgery features with strong semantic concept correlation and features with high forgery relevance but weak semantic concept ties to ensure the model converges to powerful features. Motivated by this, we apply convolutional modules and adaptive weight parameters to avoid over-relying on semantic concepts.

We thoroughly evaluate the generalization performance of our model on the UnivFD dataset [34] and the SynRIS dataset [5]. Surprisingly, our method achieves superior performance by a ap_m of 98.51% and a acc_m of 93.61% on the UnivFD benchmark [34] and an average AUROC of 95.1% on the SynRIS benchmark [5]. In summary, our contributions are as follows:

- We propose a robust Semantic Discrepancy-aware Detector (SDD) for forgery detection, specifically designed to align the semantic concept space and forgery space in terms of visual information.
- We sample semantic tokens to mitigate the space shifts and align the two spaces through reconstruction learning. Additionally, we strengthen low-level forgery features to enhance the model's robustness.
- Our method achieves superior performance on two benchmarks, demonstrating its superior capability in comparison to existing approaches.

2. Related Work

AI-generated images detection: Extensive efforts have been devoted to enhancing the performance of AI-generated image detection. Early works like [25, 46, 47] tend to mine
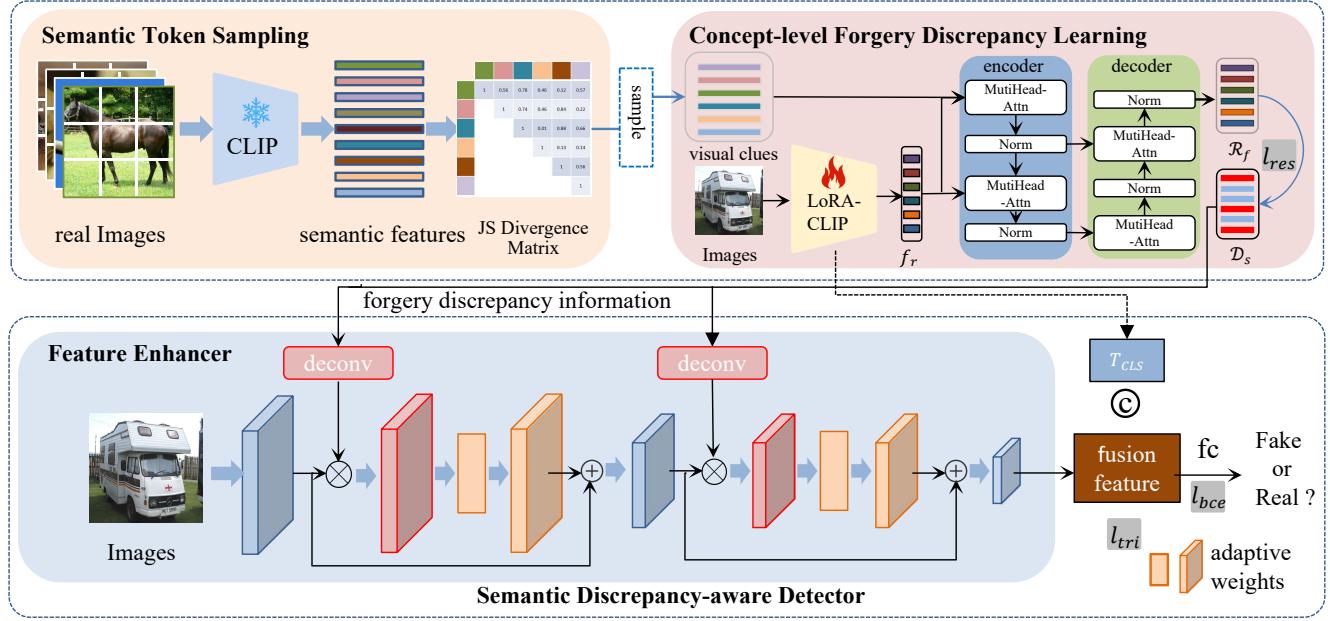
Figure 3. **The architecture of SDD.** We firstly sample semantic tokens from real images to learn features related to both semantic concepts and forgery. The input images are then mapped into a joint space of visual semantic concepts and forgery, which are transformed into learnable features V_H. We use transformer-based encoder and decoder to get reconstructed features $\mathcal{R}_f$. A reconstruction difference map $\mathcal{D}_S$ is obtained and goes through the multi-scale convolutional network to refine forgery features. Finally, we concatenate the CLIP's CLS token with this output along the same dimension for classification. The whole system is trained by jointly minimizing the binary cross-entropy loss L_{bce}, the reconstruction loss L_r, and the triplet loss L_{tri}.

the common forgery traces between all real and fake images, such as noise patterns, texture statistics, and frequency signals. As an illustration, Liu *et al*. [24] designed a network that learns the consistent noise patterns in images for fake detection. Liu *et al*. [29] proposed to leverage the gram matrix to discover the global anomalous texture of fake images. An effective approach [13] demonstrated that frequency representation is an important factor in improving fake detection performance. However, these differences are rigorously specific to the monotonous features, which contribute to the issue of overfitting. Cutting-edge research [26, 34] shifted attention toward the semantic properties of images. Ojha *et al*. [34] showed that projecting images into the feature space of pre-trained vision-language model enables strong generalization ability. To build generalized forgery representations, Liu *et al*. [26] constructed forgery adaptive space by a forgery-aware adapter. The above research [5, 26, 34] has suggested that concept attributes are vital in the image forgery detection task. Assuming that diffusion-based models leave distinct forgery traces that are characteristic of specific concept distributions, we aim to extract robust forgery features guided by semantic concepts, rather than suppressing them. Therefore, even "useless" information can be useful by providing significant certainty about the content of the image.

Reconstruction learning: Reconstruction learning has great potential in unsupervised representation learning [16,

27]. Some works [5, 28, 41] utilized reconstruction learning to reveal the nuances between real and fake images. For example, Wang *et al*. [52] found that reconstructing images by DDIM exposes an error between real images and their reconstructed replica. The new synthetic image detection method [5] used text-conditioned inversion maps to learn internal representations, which is conducive to predicting whether an image is fake. Ricker *et al*. [41] offered a simple detection approach by applying autoencoders to measure the reconstruction error. Notably, these works are committed to mining the representation discrepancy of images by treating the representation space of generative models as a forgery reference distribution. Unlike previous works, we follow UnivFD to use a pretrained vision-language model as the real reference space to mine the feature divergence of images.

3. Methodology

Our goal is to align the forgery and visual semantic concept spaces by reconstruction techniques for robust and generalizable synthetic image detection. To this end, we propose a fine-grained model named Semantic Discrepancy-aware Detector (SDD), which harness the generalization capability of vision-language models.

To better illustrate our method, we first introduce two key definitions involved in our framework. **Semantic con-**

cept space: the ideal joint embedding space of images and texts with four properties: semantic alignment, modality invariance, locality consistency, and structure preservation. **Forgery space**: this denotes an ideal discriminative representation space that highlights forgery-specific traces. Notably, we derive the semantic concept space by a vision-language model pre-trained solely on real images and thus treat the two spaces as inherently independent.

In SSD, the Semantic Tokens Sampling (STS) module utilizes Jensen-Shannon (JS) divergence to sample semantic patch tokens, facilitating the model in accurately associating real and fake images. The Concept-level Forgery Discrepancy Learning (CFDL) module employs reconstruction learning to capture the forgery discrepancies within the fine-tuned semantic concept space, which focuses on identifying subtle variations in reconstructed forgery features. Finally, the reconstruction difference map is fed into Feature Enhancer module, which aims to refine low-level forgery features with rich visual details.

3.1. Semantic Tokens Sampling

Initially, we consider to directly align the visual semantic concept space and forgery space by leveraging fine-grained reconstruction learning to model model a joint distribution over semantic concepts and forgery traces. However, this strategy would treat the differences in features unrelated to semantics and forgery as crucial factors for identifying image's authenticity. To eliminate these redundant features, we sample real semantic image patch tokens as visual cues to bridge the semantic gap between real and fake images. This module enables the model to focus on concept-related forgery traces and highlight the distinctions between real and fake images. In a tangible way, the image encoder of CLIP: ViT-L/14 is adapted to transform a real image x_r into a set of features f_r, without the image CLS token. We define the transformation as:

$$f_r = \phi(x_r), \tag{1}$$

where $\phi(\cdot)$ refers to the CLIP:ViT-L/14's visual encoder, $x_r \in \mathbb{I}_r^{H \times W \times 3}$ represents a real image characterized by a height of H and a width of W. Besides, $f_r \in \mathbb{R}^{N \times D}$, where N is the number of tokens and D denotes the dimension of each patch token.

Since integrating all real patch tokens into the image reconstruction module is computationally intensive and memory-consuming, it is urgent to select a suitable subset of these tokens. From a distribution perspective, the Jensen–Shannon (JS) divergence, derived from the Kullback–Leibler divergence [49], is a symmetric and finite metric that can effectively measure the similarity between tokens by quantifying differences in their distributions. To calculate the JS divergence between two tokens, both are converted into computable probability distribution space us-

ing the softmax function. Let $f_s \in \mathbb{F}_r^{M \times D}$ be the selected semantic patch tokens with the num of tokens $M = 1/\delta$ and the dimension D in terms of sampling rate δ ($0 \leq \delta \leq 1$, δ is user-defined). Once the initial token $\tilde{r}$ is determined, the JS divergence between $\tilde{r}$ and other tokens falls within the range $[0, 1]$. Subsequently, the sampling interval is split into M equal segments with one token r_j selected from each segment. As a consequence, the semantic tokens sampling module can be formulated as:

$$f_s = \mathcal{S}(\mathbb{R}^{N \times D}, \delta)$$
$$= A_c^{N_a \times M} \times \mathbb{R}^{N \times D},$$

$$s.t. \ a_{ij} = 1 \Rightarrow \text{JS}(\sigma(\tilde{r}), \sigma(r_j)) \in \left(\frac{j-1}{M}, \frac{j}{M} \right],$$

$$\sum_{i=1}^{N_a} \sum_{j=1}^{M} a_{ij} = M, \sum_{i=1}^{N_a} a_{ij} \in \{0, 1\}, \tag{2}$$

$$i = 1, \ldots, N_a, \ j = 1, \ldots, M,$$

where $S(\cdot, \cdot)$ represents the sampling process. $A_c^{N_a \times M}$ is a constraint matrix of size $N_a \times M$ whose element a_{ij} is constrained to the binary pattern of $\{0, 1\}$. Here JS $(\cdot)$ refers to the Jensen-Shannon divergence, N_a denotes the total number of real image patch tokens sampled from the training dataset of UnivFD and M represents the required subset size. $\sigma(\cdot)$ refers to the softmax function. The sampling tokens help the reconstruction module avoid becoming biased towards any particular forgery-unrelated distribution. Meanwhile, it avoids the semantic bias often introduced by text prompts, since the tokens are evenly distributed in a unified CLIP space.

3.2. Concept-level Forgery Discrepancy Learning

Coarse text prompts lack the semantic diversity, coverage, and detail necessary to convey the rich visual content of images, which in turn impairs the performance of image forgery detection. We argue that fine-grained visual details can uncover more forgery traces hidden in the images. As such, we mix sampling tokens with extracted features and capitalize on reconstruction learning to compensate for the omission of forgery traces brought by coarse prompts. As previous work [26] has shown that the pre-trained vision-language model necessitates fine-tuning to adapt to the forgery detection task. Therefore, we integrate LoRA [17] with the CLIP-ViT model to capture discriminative forgery features by making use of the bread semantic concepts. This method, denoted as LoRA-CLIP [55], is more streamlined and flexible. Given an input image $\mathcal{I} \in \mathbb{I}^{H \times W \times 3}$, we can get high-level visual features V_H, as follows:

$$V_H = \mathcal{F}_{LoRA}(\mathcal{I}). \tag{3}$$

Here $\mathcal{F}_{LoRA}$ refers to the CLIP image encoder fine-tuned by LoRA. The reconstruction module of CFDL encompasses

two submodules, i.e., transformer-based encoder and decoder. Thanks to the transformer's capability of long-range relationship modeling, we capitalize on the multi-head attention (MHA) mechanism, the core mechanism of transformer, to obtain more discriminative features by effectively incorporating contextual information. The MHA is set as:

$$
\begin{cases}
head_i = \text{Attn}(QW_i^Q, KW_i^K, VW_i^V) \\
\quad = \sigma\left(\dfrac{QW_i^Q(KW_i^K)^\top}{\sqrt{d}}\right)VW_i^V, \\
\text{MHA}(Q, K, V) = \text{Concat}(head_1, \ldots, head_h)W^O,
\end{cases}
\tag{4}
$$

where Q(Query), K(Key), V(Value) refer to the input, W_i^Q, W_i^K, W_i^V separately denote the corresponding weights of linear projection, Attn($\cdot$) denotes the function of the scaled dot product, d refers to the dimension of input, Concat ($\cdot$) represents the concatenation used to stitch the discrete attention outputs of head $1 \sim h$ together.

To amplify the discrepancy between a fake image and its reconstructed counterpart, the sampled visual clues are employed for the initial processing by the encoder. The encoding process can be formulated as follows:

$$
R_1 = \text{LN}(\text{MHA}(f_s, V_H, V_H)),
\tag{5}
$$
$$
R_2 = \text{LN}(\text{MHA}(R_1, V_H, V_H)),
\tag{6}
$$

where $\text{LN}(\cdot)$ denotes the Layer Normalization. Then, the encoder's outputs used as queries are injected into the decoder to get the final reconstructed features, which are similar to the encoder process and perform the following operation:

$$
R_3 = \text{LN}(\text{MHA}(R_2, R_2, R_2)),
\tag{7}
$$
$$
R_e = \text{LN}(\text{MHA}(R_1, R_3, R_3)).
\tag{8}
$$

During the reconstruction process, we just calculate the reconstruction loss $\mathcal{L}_r$ between the real input features and their reconstructed counterparts $\mathcal{R}_e$ within a mini-batch as follows:

$$
\mathcal{L}_r = \frac{1}{B}\sum_{i=0}^{B} \text{MSE}(R_e, V_H),
\tag{9}
$$

where $\text{MSE}(\cdot, \cdot)$ is the mean squared error. $\mathcal{L}_r$ encourages preserving the completeness and richness of the visual semantic concept space and highlighting the concept-related forgery features. Given the reconstructed features R_f and the original feature f_r, the reconstruction difference map can be formally expressed as:

$$
\mathcal{D}_s = |R_f - f_r|,
\tag{10}
$$
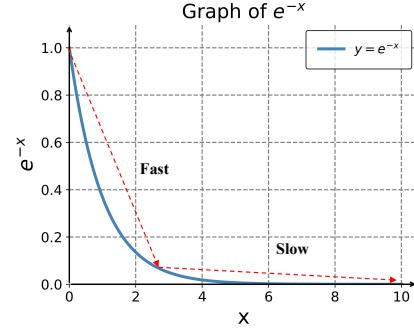
where $|\cdot|$ denotes the absolute value function.

Figure 4. The curve of exponential inverse. In the "fast" interval, the value drops sharply. In the "slow" interval, the curve flattens out, indicating a decay toward 0.

3.3. Low-level Forgery Feature Enhancer

Existing methods based on pre-trained vision-language models [26, 34] overlook the importance of features with weak semantic relevance. We believe that a thorough alignment between the visual semantic concept space and the forgery space should include the exploration of forgery features that are weakly related to semantic concepts. Meanwhile, to eliminate redundant forgery features, we come up with a novel feature enhancer that refines low-level forgery features. Empowered by the reconstruction difference map, our detector orchestrates the extraction of multi-scale features with exceptional robustness and markedly enhanced effectiveness. As shown in Fig. 3, the enhancer follows the typical architecture of a convolutional network. It involves the repeated application of convolutions, each followed by a batch normalization (BN) and a rectified linear unit (ReLU). For a given stage n, $F(n)$ ($n = 1, 2, 3$) corresponds to its output features. We then deconvolve the semantic difference map D_s to match the shape of $F(n)$ and perform pixel-wise multiplication with $F(n)$ to get $F'(n)$ as:

$$
F'(n) = F(n) \otimes \text{deconv}(D_s),
\tag{11}
$$

where $\otimes$ is the element-wise multiplication, deconv($\cdot$) represents deconvolution operation and $F'(n)$ is the low-level feature aggregated with semantic information. To further enhance the reliability of the extracted features, we compute an adaptive weight coefficient $\frac{1}{e_n}$ to indicate the importance of D_s to $F(n)$:

$$
\frac{1}{e_n} = \frac{1}{e^{|F'(n)-F(n)|}}.
\tag{12}
$$

Here we explain the role of the exponential inverse through Fig. 4. As x grows large, the curve of e^{-x} becomes flatter. In the "fast" interval, forgery features with a significant divergence from the semantic difference map will be assigned smaller weights, which mobilizes the network to capture concept strongly-related features. However, in the "slow" interval, features strongly associated with forgery

Methods	Ref	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAP
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 Steps	200 w/cfg	100 Steps	100 27	50 27	100 10		
CNN-Spot	CVPR2020	100.0	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
PatchFor	ECCV2020	80.88	72.84	71.66	85.75	65.99	69.25	76.55	76.19	76.34	74.52	68.52	75.03	87.10	86.72	86.40	85.37	83.73	78.38	75.67	77.73
Co-occurrence	Elect.Imag.	99.74	80.95	50.61	98.63	53.11	67.99	59.14	68.98	60.42	73.06	87.21	70.20	91.21	89.02	92.39	89.32	88.35	82.79	80.96	78.11
Freq-spec	WIFS2019	55.39	100.0	75.08	55.11	66.08	100.0	45.18	47.46	57.12	53.61	50.98	57.72	77.72	77.25	76.47	68.58	64.58	61.92	67.77	66.21
Dire	ICCV2023	100.0	83.59	81.50	96.50	81.70	99.88	95.73	62.51	69.98	97.31	98.62	79.53	75.52	73.42	76.45	86.28	89.00	88.34	51.35	83.54
UnivFD	CVPR2023	100.0	99.46	99.59	97.24	99.98	99.60	82.45	61.32	79.02	96.72	99.00	87.77	99.14	92.15	99.17	94.74	95.34	94.57	97.15	93.38
NPR	CVPR2024	100.0	99.50	96.50	99.80	96.80	100.0	92.20	73.10	78.70	87.20	64.80	65.80	99.80	99.80	99.80	99.70	99.80	99.80	98.60	92.19
FatFormer	CVPR2024	100.0	100.0	99.98	99.75	100.0	100.0	97.99	97.94	81.23	99.84	99.93	91.92	99.83	99.22	99.89	99.27	99.50	99.33	99.84	98.18
Ours		100.0	99.77	99.93	99.48	99.98	99.97	97.23	97.91	93.10	99.79	99.96	92.06	99.88	98.95	99.92	98.06	98.29	97.73	99.81	98.51

Table 1. Average precision comparisons on the UnivFD dataset. We replicate the results of CNNSpot, Patchfor, Co-occurrence, Freq-spec, and UnivFD from the work [34]. Additionally, the results of Dire and NPR are obtained from models retrained by ourselves, whereas those of FatFormer come from the official pre-trained model. Red and underline indicate the best and the second best result, respectively.

Methods	Ref	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	Avg-acc
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 Steps	200 w/cfg	100 Steps	100 27	50 27	100 10		
CNN-Spot	CVPR2020	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
PatchFor	ECCV2020	75.03	68.97	68.47	79.16	64.23	63.94	75.54	75.14	75.28	72.33	55.30	67.41	76.50	76.10	75.77	74.81	73.28	68.52	67.91	71.24
Co-occurrence	Elect.Imag.	97.70	63.15	53.75	92.50	51.10	54.70	57.10	63.06	55.85	65.65	65.80	60.50	70.70	70.55	71.00	70.25	69.60	69.90	67.55	66.86
Freq-spec	WIFS2019	49.90	99.90	50.50	49.90	50.30	99.70	50.10	50.00	48.00	50.60	50.10	50.90	50.40	50.40	50.30	51.70	51.40	50.40	50.00	55.45
Dire	ICCV2023	99.86	73.47	60.68	72.39	65.15	93.60	88.86	52.78	56.39	90.07	94.05	61.05	59.35	59.95	60.65	69.30	72,70	71.00	52.75	71.19
UnivFD	CVPR2023	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
NPR	CVPR2024	99.80	92.00	89.50	96.30	87.60	99.70	79.40	61.40	70.60	74.50	57.10	55.23	97.40	98.70	97.90	97.00	97.90	97.00	88.80	86.20
FatFormer	CVPR2024	99.89	99.36	99.50	97.12	99.43	99.75	93.25	81.39	68.04	69.47	69.47	76.00	98.55	94.85	98.60	94.30	94.60	94.15	98.70	90.86
Ours		99.88	95.76	96.70	98.08	98.46	99.17	91.82	83.61	77.45	95.40	96.47	79.55	98.05	94.60	98.25	92.20	93.35	91.80	98.00	93.61

Table 2. Accuracy comparisons with different methods on the UnivFD dataset.

can avoid being misguided by semantic concepts, which indicates that the order of importance is reversed. Next, we obtain the attended low-level features F_{low} by the residual connection:

$$F_{low}(n) = F(n) + \frac{F^(n)}{e_n}. \tag{13}$$

For optimizing the anchor features f_a of the enhancer, the following triplet loss [37] is employed to bring positive samples f_p closer while pushing negative samples f_n apart:

$$\mathcal{L}_{tri} = \max(0, d(f_p, f_a) - d(f_n, f_a) + \alpha), \tag{14}$$

where $d(\cdot)$ represents the Euclidean distance between samples and α is the margin. On top of that, we concatenate the LoRA-CLIP's CLS token T_{CLS} with F_{low} along the same dimension to yield the refined representation F_{out}. This ensures that forged features exhibit distinctiveness across different semantic identities while preserving their uniformity within similar semantic identities. The process is formulated as:

$$F_{out} = F_{low} || T_{CLS}, \tag{15}$$

where F^{out} is strategically integrated with a linear classifier to enable the execution of binary classification. Eventually,

the total loss function $\mathcal{L}$ of the proposed framework can be defined as:

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{tri} + \lambda_2 \mathcal{L}_r, \tag{16}$$

where $\mathcal{L}_{bce}$ and $\mathcal{L}_{tri}$ refer to the binary cross-entropy loss and the triplet loss. $\mathcal{L}_{tri}$ and $\mathcal{L}_r$ are scaled by the hyperparameters λ_1 and λ_2, respectively.

4. Experiments

4.1. Experiment Setups

Datasets: We follow the protocol described in [34], using ProGAN's images as training data. Additionally, we adopt the protocol from [5], where the training data is composed of fake Stable Diffusion v1 images [53] and random real LAION images [43]. The UnivFD dataset [34] covers a broad range of generative models, including GANs and diffusion models, such as ProGAN [18], StyleGAN [19], Big-GAN [4], CycleGAN [60], StarGAN [10], GauGAN [48], CRN [9], IMLE [22], SAN [11], SITD [7], DeepFakes [20], Guided [12], Glide [33], LDM [42] and DALL-E [39]. The SynRIS dataset [5] is designed to avoid bias toward any specific topic, theme, or style and contains high-fidelity images generated by text-to-image models, such as Kandinsky2 [40], Kandinsky3 [1], PixArt-α [8], SDXL-DPO [50],

Methods	CNN-Spot	Freq-spec	Dire	Univ FD	NPR	FatFormer	FakeInversion	Patch For	**Ours**
Kandinsky2	60.0	57.0	71.6	56.2	97.5	75.6	69.9	53.5	97.1
Kandinsky3	65.9	45.7	74.9	61.4	93.7	80.1	74.3	51.4	94.8
PixArt-α	62.7	56.4	81.5	64.7	89.5	75.3	73.0	49.6	90.2
SDXL-DPO	84.3	69.8	69.9	70.2	97.6	86.0	88.1	54.5	95.5
Segmind-Vega	74.2	65.3	81.0	62.3	97.1	82.4	81.1	53.1	97.0
SDXL	81.4	61.2	86.2	66.3	96.0	85.1	80.7	63.9	97.6
Seg-MoE	66.3	54.6	71.9	62.0	93.6	70.8	71.3	49.8	97.6
SSD-1B	72.6	67.8	79.8	62.8	99.6	70.1	79.4	61.2	99.8
Stable-Cascade	70.5	62.1	74.1	68.2	97.6	81.6	74.9	57.4	99.2
Würstchen2	61.0	63.3	74.2	69.7	90.9	72.9	70.5	47.2	98.2
Midjourney	63.0	50.9	72.4	59.2	58.5	73.6	66.4	53.7	90.0
Playground	58.2	52.3	67.9	58.7	93.1	81.4	62.5	54.1	92.8
DALL·E3	71.6	59.9	80.8	48.0	69.1	79.2	75.9	50.1	85.9
Average	68.6	58.9	75.9	62.3	90.3	78.0	74.5	53.8	95.1

Table 3. AUROC comparisons with different methods on the Syn-RIS dataset. We retrieve the results of CNNSpot, UnivFD, and FakeInversion from [5] and obtain the results for Dire, NPR, Fat-Former, and PatchFor using re-implemented models. Red and underline indicate the best and the second best result, respectively.

#	STS module	CFDL module	Feature Enhancer	UnivFD ap_m	Dataset acc_m
1		✓		97.37	81.64
2		✓	✓	97.41	90.17
3	✓	✓		97.39	89.98
4	✓	✓	✓	98.52	93.61

Table 4. Ablation study of the proposed modules on the UnivFD Dataset. We show the mean accuracy (acc_m) and average precision (ap_m). Red and underline indicate the best and the second-best result, respectively.

Pytorch on 2 Nvidia GeForce RTX A6000 GPUs.

4.2. Comparision Results

The UnivFD dataset includes a diverse range of models, allowing for a comprehensive evaluation of our method across both GAN and diffusion generative models. In addition, the SynRIS dataset provides images generated by cutting-edge generative models.

Results on the UnivFD dataset: As reported in Table 1 and Table 2, experimental results show that our proposed method achieves superior performance compared to exiting methods. Notably, without the biased interpretation introduced by coarse-grained text prompts, SDD surpasses the latest state-of-the-art method FatFormer by the mean AP (ap_m) of 0.34% and the mean acc (acc_m) of 2.75%. Compared with methods based on relatively monotonous forgery features, our approach can outperform all of them with a large improvement. The above evidence indicates effective combination of visual concepts and forgery features can contribute model to extract sufficient forgery traces and eliminate the superfluous features.

Results on the SynRIS dataset: As shown in Table 3, when confronted with high-fidelity images generated by text-to-image models, methods leveraging pre-trained vision-language models, such as UnivFD and FatFormer, lose their competitiveness. In contrast, NPR, which focuses on neighboring pixel relationships, retains its edge. We assume that current generative methods grasp the relationships between visual information and semantic concepts in images but cannot refine local visual details at the pixel level. Considering that excessive reliance on concepts misses abnormal pixel arrangements and focusing on monotonous forgery patterns can cause overfitting, our detector, which emphasizes low-level features with visual concepts, is trained on lower-fidelity fake images generated by Stable Diffusion [53] to capture concept-specific lacunae. We follow the evaluation protocol from SynRIS [5]. In comparison, our detector achieves an impressive mean AU-ROC of 95.1%, surpassing the state-of-the-art method by 4.8%. This demonstrates its superior ability to tackle the challenges posed by evolving generative models.

SDXL [36], SegMoE [23], SSD-1B [44], Stable-Cascade [35], Segmind-Vega [15], and Würstchen2 [35], Midjourney [31], DALL.E 3 [3] and playground [21]. As standard evaluation metrics, the average precision (AP), the accuracy (ACC) and AUCROC are considered to measure the effectiveness of different methods.

Baselines: We perform comparisons with state-of-the-art methods, as follows: 1) CNNSpot [51] relies only on a CNN. 2) PatchFor [6] performs detection on a patch level. 3) Co-occurrence [32] converts input images into co-occurrence matrices for detection. 4) Freq-spec [58] employs the frequency spectrum of images. 5) Dire [52] exploits the error between input images and its reconstruction counterpart. 6) UnivFD [34] uses the pre-trained language-vision model to determine the authenticity of images. 7) NPR [47] captures the generalized artifacts according to the local interdependence among image pixels. 8) FatFormer [26] aims at extracting forgery-adaptive features based on UnivFD. 9) Fakeinversion [5] employs text-conditioned inversion maps extracted from Stable Diffusion.

Implement details: Our training settings are adapted from the approach outlined in the previous study [26] with several key modifications. Specifically, early stopping is employed during model training, with an initial learning rate of 1×10^{-5} and a batch size of 32. Additionally, the Lora layers are configured with hyperparameters $lora_r = 6$, $lora_\alpha = 6$, and a dropout rate of 0.8, while α is set to 8.0. The proposed method is implemented using
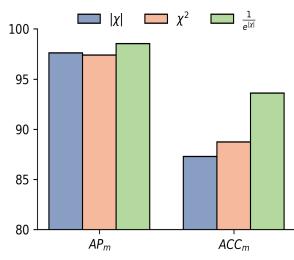
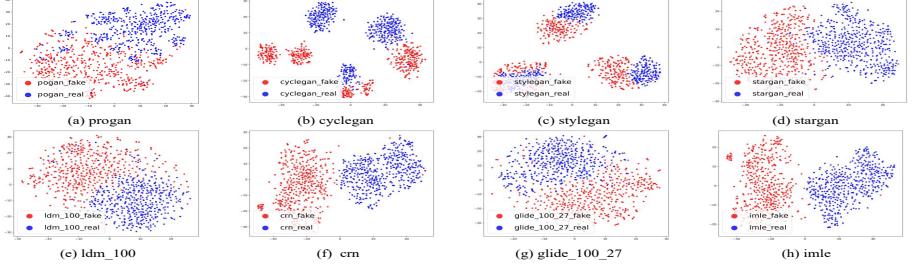Figure 5. Performance of different functions on adaptive weights.

Figure 6. T-SNE visualization of real and fake images [30]. The feature space is based on our classifier. Each randomly samples 500 real and 500 fake images.
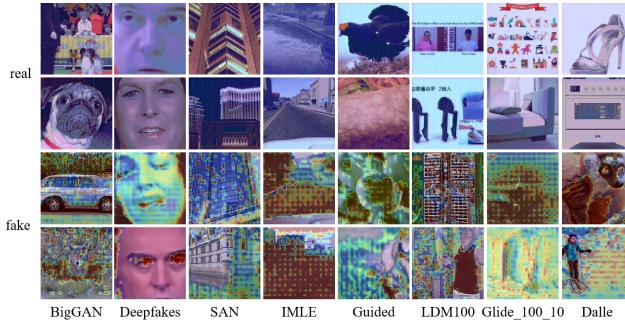

Figure 7. The showcase of attention maps for the input images.

4.3. More analysis

We perform comprehensive ablation studies on the UnivFD dataset under the original experimental configurations, reporting the mean accuracy (acc_m) and mean average precision (ap_m) as the primary evaluation metrics.

Effect of each component: We study the effects of removing STS module, CFDL module, and Feature Enhancer in our method. The results, presented in Table 4, demonstrate that these components are essential for improving performance in generalization on unseen models. This empirical finding suggests that CFDL effectively captures forgery discrepancies associated with semantic concepts, while the enhancer plays a crucial role in identifying robust forgery artifacts. The collaboration of all modules enhances the model's ability to distinguish between real and fake images.

Effect of function on adaptive weights: To check how well the proposed function works in the SDD, we select two conventional functions for comparison: $f(x) = |x|$ and $f(x) = x^2$. The corresponding results are presented in Table 5. We find our proposed function yields improvements in both ap_m and acc_m compared to the selected functions. These results demonstrate that our proposed function is effective in capturing robust and distinctive forgery features.

Visualization of learned latent space: As shown in Fig. 6, the input images can be distinctly categorized into two clusters: real and fake. Nevertheless, why does the divided boundary of ProGAN appear ambiguous in contrast to other models? Additionally, why do the real clusters of CycleGAN and StyleGAN separate from each other? We attribute these to the influence of visual semantic concepts. Perceptively, with the supervision of visual semantic concepts, the learned boundary of ProGAN is more complex and nuanced, rather than just simple straight lines or curves. Similarly, the images generated by StyleGAN and CycleGAN are projected into the corresponding semantic concept distribution and then separated from the real images based on the visual semantic concepts.

Visualization of attention on images: Fig. 7 illustrates the attention maps for the input images, where we apply class activation mapping [59] to visualize the learned representations. With the aid of semantic information, our model can focus on different regions of fake images, including the background, local object regions, and marginal details. This suggests that our fine-grained model is capable of capturing intricate discrepancies generalized to unseen models. Notably, the real images nearly always show no forgery discrepancy regions, which demonstrates the effectiveness of the reconstruction loss in the forgery detection task.

5. Conclusion

In this paper, we propose a novel method, SDD, for generalizable forgery image detection. The findings show that our method establishes a new state-of-the-art in detecting images generated by generative models from different periods, which underscores its robustness and superior generalization capability. To the best of our knowledge, in pretrained vision-language paradigms, our approach is the first to rely solely on visual information, without text prompts. Based on experimental results, we conclude that leveraging sampled tokens and reconstruction techniques effectively aligns the visual semantic concept space with the forgery space. Additionally, refining low-level forgery features under the supervision of visual semantic concepts enhances the performance of forgery detection. Although SDD performs well across various generative methods, there is still room for improvement as generative technologies continue to advance.

Acknowledgement

References

[1] Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 3.0 technical report. *arXiv preprint arXiv:2312.03511*, 2023. 6

[2] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021. 1

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 7

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 6

[5] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024. 2, 3, 6, 7

[6] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pages 103–120. Springer, 2020. 1, 7

[7] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018. 6

[8] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart-{\delta}: Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024. 6

[9] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 6

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 6

[11] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6

[13] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 1, 3

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63:139 – 144, 2014. 1

[15] Yatharth Gupta, Vishnu V Jaddipal, Harish Prabhala, Sayak Paul, and Patrick Von Platen. Progressive knowledge distillation of stable diffusion xl using layer level loss. *arXiv preprint arXiv:2401.02677*, 2024. 7

[16] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Multi-angle point cloud-vae: Unsupervised feature learning for 3d point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10441–10450. IEEE, 2019. 3

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[20] Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2589–2597, 2020. 6

[21] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 7

[22] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4220–4229, 2019. 6

[23] Zhenghong Li, Hao Chen, Jiangjiang Wu, Jun Li, and Ning Jing. Segmind: Semisupervised remote sensing image semantic segmentation with masked image modeling and contrastive learning method. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–17, 2023. 7

[24] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022. 3

[25] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in

frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021. 2

[26] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10770–10780, 2024. 1, 2, 3, 4, 5, 7

[27] Xinhai Liu, Xinchen Liu, Yu-Shen Liu, and Zhizhong Han. Spu-net: Self-supervised point cloud upsampling by coarse-to-fine reconstruction with self-projection optimization. *IEEE Transactions on Image Processing*, 31:4213–4226, 2022. 3

[28] Xiyao Liu, Jiaxin Hu, Qingying Yang, Ming Jiang, Jianbiao He, and Hui Fang. A divide-and-conquer reconstruction method for defending against adversarial example attacks. *Visual Intelligence*, 2(1):30, 2024. 3

[29] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020. 3

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008. 8

[31] Midjourney. Midjourney, n.d. Accessed: 2025-03-04. 7

[32] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019. 7

[33] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 6

[34] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 1, 2, 3, 5, 6, 7

[35] Pablo Pernias, Dominic Rampas, and Marc Aubreville. Wuerstchen: Efficient pretraining of text-to-image models. 2023. 7

[36] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. 7

[37] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 783–792, 2021. 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 6

[40] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023. 6

[41] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140, 2024. 3, 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 6

[44] Segmind. Announcing ssd-1b: A leap in efficient t2i generation., 2023. 7

[45] Minghe Shen, Hongping Gan, Chao Ning, Yi Hua, and Tao Zhang. Transcs: A transformer-based hybrid architecture for image compressed sensing. *IEEE Transactions on Image Processing*, 31:6991–7005, 2022. 2

[46] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 1, 2

[47] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2, 7

[48] Devavrat Tomar, Manana Lortkipanidze, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-attentive spatial adaptive normalization for cross-modality domain adaptation. *IEEE transactions on medical imaging*, 40(10):2926–2938, 2021. 6

[49] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. 4

[50] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 6

[51] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1, 7, 2, 3

[52] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. 3, 7

[53] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 6, 7

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1934–1948, 2022. 2

[55] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024. 4

[56] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023. 5

[57] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7579–7590, 2023. 5

[58] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019. 7

[59] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 8

[60] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 6

Semantic Discrepancy-aware Detector for Image Forgery Identification

Supplementary Material

A. Statistical analysis of feature values from detectors

Previous works have validated the effectiveness of the CLIP model in the forgery detection task. Unlike conventional detection models, CLIP distinguishes itself by jointly learning from both visual and textual modalities, enabling it to understand and align images with natural language. This enables CLIP to better understand the semantic relationships between images and text, allowing for more nuanced detection of subtle forgery traces. In contrast, traditional detectors typically focus exclusively on the visual features of the images themselves, without leveraging additional semantic conceptual information. We attribute our preliminary findings to the influence of conceptual semantic factors, which help to distinguish real from fake images more effectively.

In Section 2, we briefly introduce the differences between the features extracted by CNNSpot [51] and CLIP [34]. In this section, we provide a more detailed discussion of these differences and investigate the characteristics of these differences and the reasons behind them. To further validate the distinguishing role of concepts in differentiating real and fake images, we observe the feature spaces of different categories of real and fake images. Both sets of features originate from the inputs of the detectors' final fully connected (FC) layers. This setup enables us to explore the distinction between concept-related features and those typically extracted by general detectors.

As illustrated in Fig. 14, we observe a notable difference in how real and fake images are represented in the visual semantic concept space. The gap between real and fake images in CLIP space is more pronounced across various categories, suggesting that semantic concepts help separate these two types of images more effectively. In contrast, in the feature space of CNNSpot, the distinction between real and fake images becomes much less obvious and more uniform, indicating that the learned features tend to exhibit monotonic patterns, which may lead to overfitting and limiting the model's ability to generalize to unseen data. This highlights the importance of incorporating conceptual semantic understanding into the feature extraction process.

From these observations, we conclude that the use of concept-based features can significantly alleviate the problem of overfitting and improve a model's ability to generalize to unseen generative models.

B. Analysis of Semantic Description Granularity in FatFormer

In the introduction, we point out that the soft prompts based on simple [CLASS] embeddings of FatFormer have an intrinsic limitation in their semantic description granularity. This concern arises from our observation that FatFormer achieves a significantly lower $racc_m$ compared to UnivFD. The $racc_m$ is shown in Table 5. This indicates that, when faced with real images, FatFormer is more likely to misclassify them as fake images. In more extreme terms, compared to its backbone, FatFormer appears to have lost the ability to recognize authentic images, which is clearly an anomalous behavior.

Our intuitive explanation is that the coarse-grained soft prompts used in FatFormer weaken its ability to perceive varying visual semantic details in real images. To validate this hypothesis, we randomly sampled 5,000 pairs of images and computed the cosine similarity between them based on the output vectors of FatFormer's text encoder and the final-layer features of UniVFD's image encoder. As shown in Fig. 11 and Fig. 10, the cosine similarity scores for UnivFD show a wider range of variation compared to FatFormer's. It indicates that FatFormer's soft prompts fail to distinguish semantic differences between images, indicating a significant decline in semantic discrimination capability.

Furthermore, we compared the semantic similarity between real images that were misclassified as fake and those that were correctly classified. As shown in Fig. 9, despite the higher semantic similarity among real images, the UnivFD is still able to correctly determine their authenticity. In contrast, FatFormer, while eliminating semantic information interference, fails to make accurate authenticity judgments.

These findings suggest that although forgery-adaptive mechanisms improve FatFormer's sensitivity to forgery traces, the lack of adequate semantic-guided information provided by the soft prompts hinders the model's generalization ability in real-world scenarios.

C. Training details

In this section, we provide the details regarding the training process of our work. We use the official code repository provided by [34]. We train the CLIP:ViT variant of this baseline with Blur and JPEG augmentations applied with a probability of 0.5. The network is trained with a batch size of 32 and a learning rate of 1×10^{-4}. The random seed is set to 46. For the loss function, the hyper-parameters λ_1 and λ_2 are set to $\frac{1}{9}$ and $\frac{1}{3}$, respectively. During testing, no

(a) all
(b) cat
(c) dog
(d) person

(e) all
(f) cat
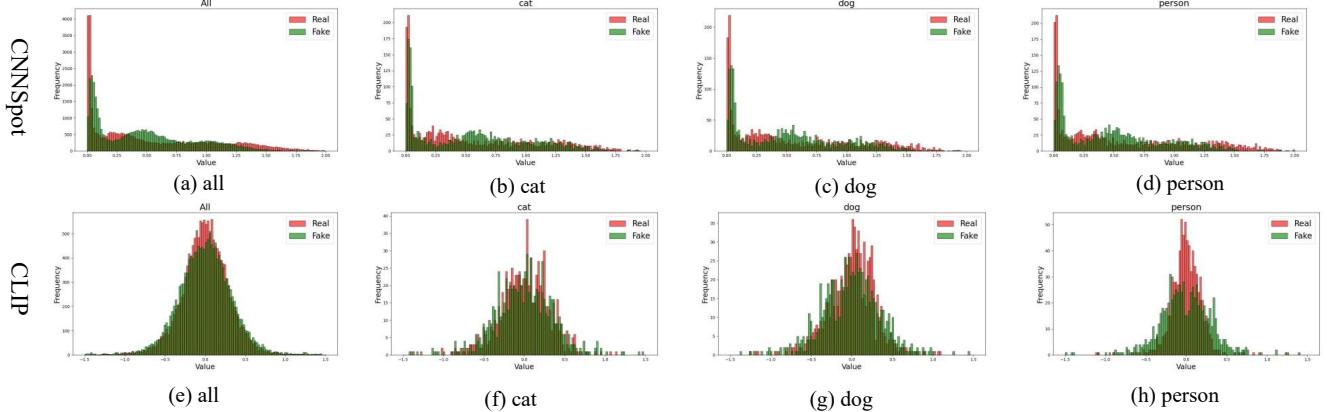(g) dog
(h) person

CNNSpot

CLIP

Figure 8. **Value statistics of extracted features.** We compare the input features from the last FC layer of CNNSpot [51] and CLIP [34], both of which are fed with ProGAN [41] data. Three classes from ProGAN's testing data are considered: cat, dog, and person. We also present the results for data from all classes.

source real Images

target real Images

FatFormer's semantic similarity | 0.9344 | 0.9345 | 0.9343 | 0.9346 | 0.9348

CLIP's semantic similarity | 0.8511 | 0.7245 | 0.8314 | 0.7903 | 0.6608

Figure 9. **Semantic similarity comparison of real images.** Inside the red dashed box, the source real images are correctly classified, while the target real images are misclassified. Inside the blue dashed box, both the source real images and the target real images are correctly classified.
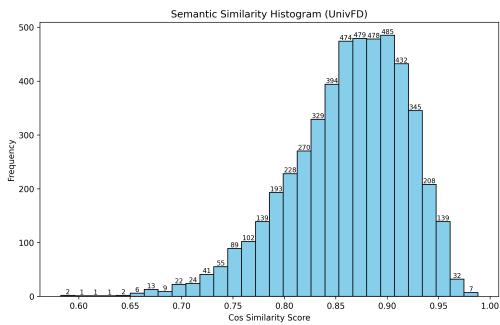

Figure 10. **Semantic similarity histogram of UnivFD.** The data is primarily concentrated in the cosine similarity range of 0.85 to 0.90, with the overall data falling within the range of 0.582 to 0.988.
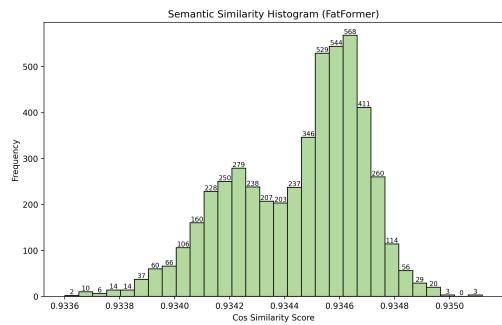

Figure 11. **Semantic similarity histogram of FatFormer.** The data is primarily concentrated in the cosine similarity range of 0.9344 to 0.9348, with the overall data falling within the range of 0.9336 to 0.9351.

Blur or JPEG augmentation is applied. Lastly, when training our classifier, we make use of Blur + JPEG data augmentations, any real or fake image is first augmented before being passed to the CLIP:ViT encoder (φ).

D. Effect of sampling rate δ of SDL

In this section, we investigate the effect of the parameter δ on forgery detection performance. We set δ to various values of $\frac{1}{500}, \frac{1}{1000}, \frac{1}{2000}, \frac{1}{4000}, \frac{1}{5000}$ to explore how the number of sampled tokens impacts the detection task. Notably, our sampled dataset is drawn from the entire training set in [34]. Despite the large size of the training set, the number of sampled tokens remains below expectations — some segments contain no patch tokens at all.

Intuitively, increasing the number of tokens should allow the model to better reflect the true distribution of visual semantic concepts, as more tokens provide a more comprehensive representation of the whole image. As shown in Fig. 12, when δ changes from $\frac{1}{500}$ to $\frac{1}{1000}$, although the change in Average Precision (AP_m) is not significantly large, there is a noticeable improvement in Accuracy (ACC_m), demonstrating that the additional tokens help the model better differentiate between real and fake images.

Beyond this point, as δ continues to increase, the changes in both AP_m and ACC_m increase gradually, suggesting that after a certain threshold, increasing the number of sampled tokens yields diminishing returns in performance. These findings underscore the effectiveness of the sampled tokens in enhancing the model's ability to detect forgery traces.

Moreover, even with a relatively small number of tokens, the model achieves significant performance improvements. This characteristic is especially valuable as it reduces both computational costs and memory usage, making it a more efficient solution for real-world applications. In conclusion, this finding highlights that our method is both effective in detecting real and fake images and computationally efficient, even with fewer tokens.

E. Robustness

In order to evade a fake detection system, an attacker may apply certain low-level post-processing operations to the fake images. To evaluate the robustness of our classifier against such operations, we follow prior work and assess its performance under different post-processing conditions. As shown in Fig. 13, our method demonstrates general robustness to both blur and JPEG compression artifacts compared to the baseline [34].

It is worth noting that as the Gaussian blur sigma value changes, the average precision (AP) for different generative models consistently remains above 75%, with the exception of the SAN model. This indicates that our method is quite robust to Gaussian blur, effectively detecting forgery traces

even under varying levels of blur. In particular, the AP remains stable across most generative models, suggesting that our approach maintains strong performance in the presence of noise or degradation typically introduced by Gaussian blur. However, for the SAN model, a noticeable drop in AP suggests that certain models, such as SAN, might be more sensitive to this type of distortion.

On the other hand, when the JPEG compression quality is varied, the AP for all forgery models remains consistently above 80%, indicating that our method is highly resistant to JPEG compression artifacts. This is a strong indication of the model's ability to maintain accuracy even under common image compression techniques that often degrade the quality of forged images. Notably, our models exhibit minimal degradation in AP, which demonstrates their capability to accurately distinguish between real and fake images, even when compression artifacts are present. In contrast, models that are not robust to such distortions may experience significant drops in AP, reflecting their vulnerability to such post-processing operations.

F. Accuracy breakdown of real and fake classes

Lastly, we break down the performance of different methods into performance on real (Table 5) and fake images (Table 6) associated with different generative models. This breakdown helps us understand the specific ways in which a detection method may fail. In particular, we observe that an image-level classifier, such as CNNSpot [51], works well in detecting real and fake images when they belong to the GAN domain. However, when tested on images from latent diffusion models, the network tends to classify almost all images as real. Consequently, while the classification accuracy on real images remains high, the accuracy on fake images drops drastically.

In contrast to other models, our method strikes a remarkable balance between performance on real and fake images, as evidenced by the results in Table 6 and Table 5, where the fake image classification accuracy ($facc_m$) and real image classification accuracy ($racc_m$) are 93.16% and 94.06%, respectively. This indicates that our model excels at distinguishing between real and fake images. Furthermore, our model has learned a feature space that effectively differentiates between these two categories. This ability to maintain consistent performance across both real and fake images highlights the robustness and effectiveness of our model in real-world applications. Our approach demonstrates its capability to detect subtle forgery traces, irrespective of the generative model used to create the fake images.

Methods	Ref	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	Avg-acc
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 Steps	200 w/cfg	100 Steps	100 27	50 27	100 10		
CNN-Spot	CVPR2020	100.0	98.64	99.05	99.95	99.40	99.30	99.45	100.0	100.0	99.22	99.22	99.14	99.61	99.61	99.61	99.61	99.61	99.61	99.61	99.50
PatchFor	ECCV2020	95.30	65.56	61.35	85.95	49.88	75.83	89.21	43.48	47.24	12.25	12.25	61.34	84.86	84.86	84.86	84.86	84.86	84.86	84.86	68.08
Freq-spec	WIFS2019	99.80	99.80	99.10	99.90	99.80	99.30	100.0	100.0	100.0	99.80	99.80	99.60	99.40	99.40	99.50	99.40	99.50	99.40	99.60	99.60
UnivFD	CVPR2023	99.08	87.21	92.55	99.63	95.88	99.35	96.0	61.0	95.0	96.47	96.47	93.34	92.39	92.39	92.39	92.39	92.39	92.39	92.39	92.56
FatFormer	CVPR2024	100	98.71	99.1	100	98.88	99.50	99.45	63.3	98.17	38.94	38.94	97.90	99.30	99.30	99.30	99.30	99.30	99.30	99.30	90.95
SDD		100.0	100.0	100.0	99.95	100.0	100.0	89.47	99.44	60.27	99.86	100.0	65.50	99.40	92.50	99.80	87.70	90.0	86.90	99.30	93.16

Table 5. **Accuracy of detecting real images.** For each generative model (column), we consider its corresponding real images and test how frequently a classifier (row) correctly predicts it as real.

Methods	Ref	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	Avg-acc
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 Steps	200 w/cfg	100 Steps	100 27	50 27	100 10		
CNN-Spot	CVPR2020	100.0	62.91	18.90	38.52	59.10	62.58	2.52	13.89	0.0	75.95	88.92	4.67	3.05	4.26	2.96	9.25	12.34	9.1	4.9	30.20
PatchFor	ECCV2020	93.45	69.20	67.90	78.56	64.51	84.74	21.31	85.70	54.90	96.33	97.96	68.94	73.32	67.48	73.86	49.26	52.23	51.22	54.02	68.68
Freq-spec	WIFS2019	0.20	100.0	1.80	0.0	0.90	100.0	0.0	0.0	0.4	1.30	0.50	0.40	1.30	1.40	1.10	3.90	3.30	1.30	0.50	11.50
UnivFD	CVPR2023	100.0	99.77	84.7	61.88	98.34	98.6	73.0	82.0	27.0	42.06	61.94	48.77	90.2	51.65	90.2	85.7	88.94	87.77	70.55	75.95
FatFormer	CVPR2024	99.78	100	99.90	94.24	99.98	100	87.83	99.44	37.90	100	100	54.10	97.80	97.90	90.40	89.30	89.90	89.00	98.10	90.78
SDD		99.75	91.52	93.40	90.59	96.92	98.35	96.16	67.78	96.35	92.95	92.95	93.60	96.70	96.70	96.70	96.70	96.70	96.70	96.70	94.06

Table 6. **Accuracy of detecting fake images.** For each generative model (column), we consider its corresponding fake images and test how frequently a classifier (row) correctly predicts it as fake.
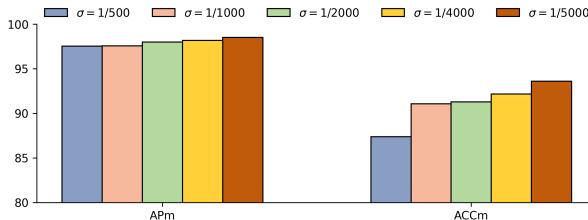

Figure 12. Performance of sampling rate δ of SDL.

Method	AP_m	Acc_m
BLIP (VIT-L/16)	95.61	84 .46
CLIP (VIT-L/14)	98.52	93.61

Table 7. Comparisons with different backbones on the UnivFD dataset.

G. Comparisons with different backbones on the UnivFD dataset.

To further investigate the role of semantic concepts, we adopt the BLIP: VIT-L/16 as the backbone for forgery detection. We hypothesize that BLIP provides stronger fine-grained perception over the entire image, potentially making it more suitable for capturing semantic-level inconsistencies in manipulated content. Unlike CLIP, which primarily focuses on contrastive learning, BLIP is trained using vision-language pretraining tasks such as image-text matching and image captioning, leading to improved vision-language alignment and a more detailed semantic understanding. During the experiment, we observed that the number of patch tokens sampled by BLIP is fewer than that by CLIP. This seems to suggest the incompleteness and inadequacy of BLIP's visual semantic concept space.

However, as shown in Table 7, using CLIP as the backbone yields better performance than using BLIP, which deepened our understanding of the semantic concept space. Despite its stronger alignment at the image-caption level, BLIP appears to have a less comprehensive and diverse concept space compared to CLIP, resulting in concept-forgery misalignment.

We attribute this limitation primarily to the scale and diversity of pretraining data. BLIP is trained on 129M samples, while CLIP uses 400M samples. The broader and more diverse supervision in CLIP likely equips it with a more robust and generalizable semantic embedding space, especially under open-world or adversarial conditions such as image forgery. Furthermore, CLIP's contrastive training may emphasize discriminative concept boundaries, which could be inherently more beneficial for tasks requiring semantic-level anomaly detection.

In summary, although BLIP possesses advantages in fine-grained alignment and descriptive representation, its current pretraining scale and objectives may limit its effectiveness in tasks like forgery detection, where broad semantic coverage and discriminative representation are critical.
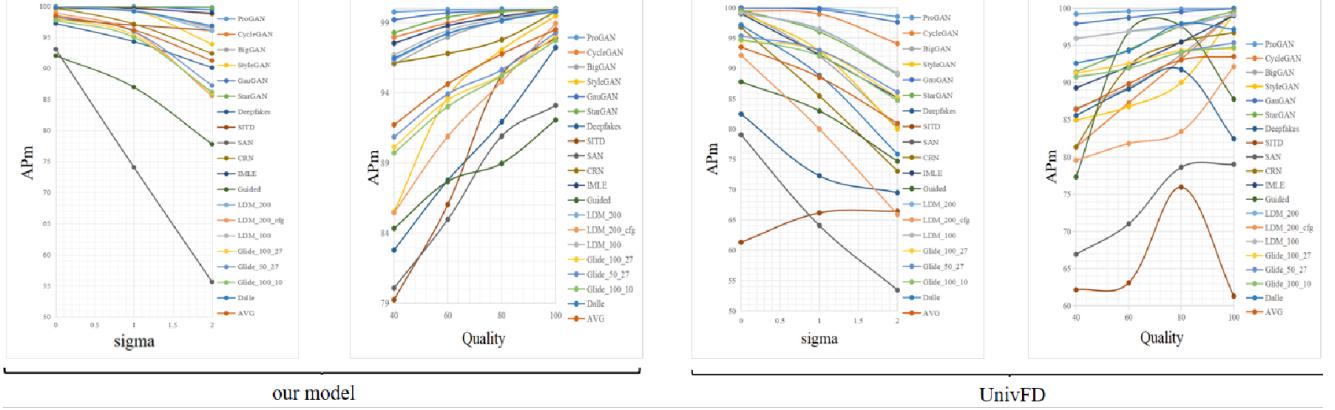
Figure 13. Robustness to different image processing operations. Both our detector and the trained baseline [34] demonstrate general robustness to these artifacts, but our performance is notably superior on unseen models.
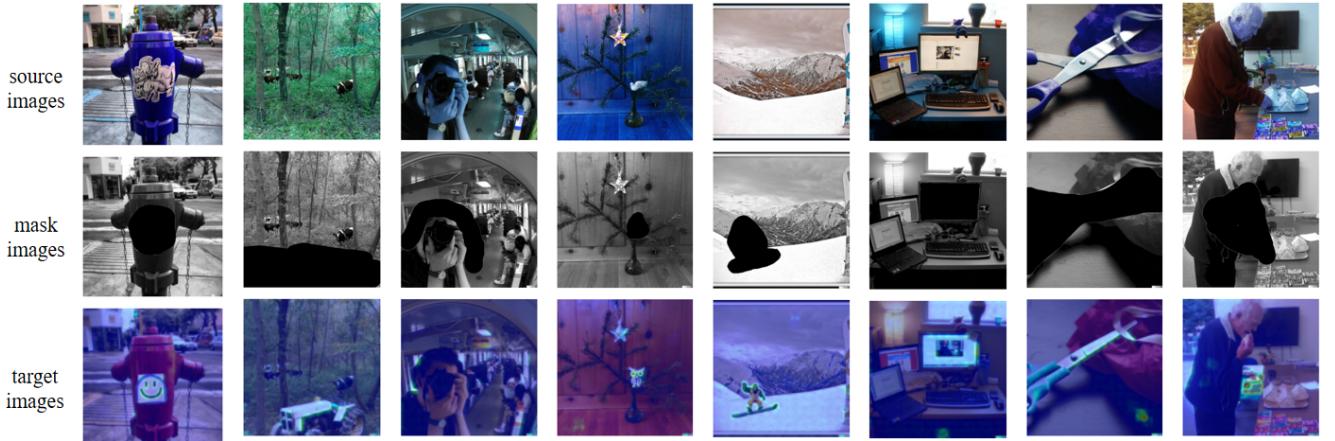

Figure 14. The visualization of attention on dataset [57]. The first row displays the original image, the second row shows the corresponding mask, and the third row presents the generated image within the masked region.

H. Effect of SDD on the tampered dataset[56]

Beyond simply classifying images as real or generated, numerous research efforts have sought to localize the edited regions within the tampered images. Since we emphasize the role of CFDL in localizing semantically relevant forgery regions in this work, we try to apply our pretrained model trained on Stable Diffusion v1 images and random real LAION images to detect manipulated regions in tampered images. Clearly, identifying the authenticity of a whole image becomes a significant challenge due to the increasing proportion of real content within a given image. To further investigate our model's ability to tackle this challenge, we conduct experiments on the MAGICBRUSH dataset [56]. MAGICBRUSH, finetuned by InstructPix2Pix, is a manually annotated dataset for instruction-guided real image editing that covers diverse scenarios: single-turn, multi-turn, mask-provided, and mask-free editing.

We input tampered images into our model and obtain the corresponding heatmaps using CAM. Although our model's forgery detection performance decreases on this dataset, by analyzing the heatmaps alongside the mask images, we are surprised to find that our model can still localize the manipulated regions, albeit with limited accuracy. This demonstrates the significant potential of our model in the field of image forgery detection. In the future, we plan to further explore methods for distinguishing fake images that have been manipulated from real images.

I. More analysis of learned latent space

We argue that the indistinct boundaries observed in generative models arise from applying t-SNE across multiple classes (i.e., semantic concepts), while the visualization itself is presented in a binary fashion (real vs. fake). To further support this claim, we perform a more fine-grained t-SNE analysis on the ProGAN test data (only the ProGAN dataset provides explicit class labels for each sample. Other

Model	FPS	Time (ms)	**GPU usage(MB)**
SDD	15	68	3555
-LORA	17	59	3252
-feature enhancement	16	61	3186
-LA	16	63	3510

Table 8. The computational cost of our model without different modules on the UnivFD dataset. The prefix '-' indicates the module is removed.

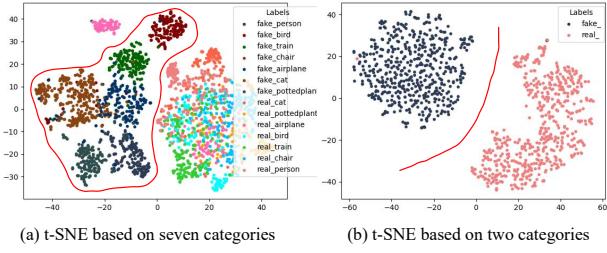(a) t-SNE based on seven categories   (b) t-SNE based on two categories

Figure 15. The t-SNE visualization of semantic concepts with different numbers of categories, where the size of samples is equal.

generative models do not offer such semantic annotations) using explicit class labels. In particular, we visualize the feature distribution of samples from a combined subset of categories — person, bird, train, chair, airplane, cat, and potted plant — as well ajhjhs from the airplane category alone.

As shown in Fig. 15, increasing concept diversity leads to blurrier global boundaries in the t-SNE projection. Nevertheless, real and fake samples within the same concept remain locally separable, suggesting that the observed structure is shaped by concept-aware organization.

J. The computational cost of our modules

We evaluate the computational cost introduced by our key components: LoRA fine-tuning (LORA), feature enhancement, and reconstruction-based alignment (RA). As shown in Table 8, all three modules introduce only a minor increase in inference-time cost, maintaining the model's efficiency while improving performance.

Specifically, the full model (SDD) runs at 15 FPS with an average inference time of 68 ms and a GPU memory footprint of 3555 MB. Removing LoRA slightly improves FPS to 17 and reduces memory usage by approximately 300 MB, indicating that LoRA contributes a small computational cost. Removing feature enhancement results in the lowest memory usage (3186 MB) and a slight FPS increase, showing that multi-scale feature fusion is lightweight in practice. Excluding the RA module also reduces the inference time slightly, suggesting that reconstruction-based

alignment introduces minimal cost while contributing important semantic consistency.

Overall, these results confirm that our proposed components are computationally efficient and practical for real-world deployment scenarios, striking a favorable balance between performance and resource consumption.