Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection

Zhiyuan Yan [*1,2] **Jiangming Wang** [*2] **Peng Jin** [1] **Ke-Yue Zhang** [2] **Chengchun Liu** [1] **Shen Chen** [2] **Taiping Yao** [2] **Shouhong Ding** [2] **Baoyuan Wu** [3] **Li Yuan** [1]

Abstract

Detecting AI-generated images (AIGIs), such as natural images or face images, has become increasingly important yet challenging. In this paper, we start from a new perspective to excavate the reason behind the failure generalization in AIGI detection, named the *asymmetry phenomenon*, where a naively trained detector tends to favor overfitting to the limited and monotonous fake patterns, causing the feature space to become highly constrained and low-ranked, which is proved seriously limiting the expressivity and generalization. One potential remedy is incorporating the pre-trained knowledge within the vision foundation models (higher-ranked) to expand the feature space, alleviating the model's overfitting to fake. To this end, we employ Singular Value Decomposition (SVD) to decompose the original feature space into *two orthogonal subspaces*. By freezing the principal components and adapting only the remained components, we preserve the pre-trained knowledge while learning fake patterns. Compared to existing full-parameters and LoRA-based tuning methods, we explicitly ensure orthogonality, enabling the higher rank of the whole feature space, effectively minimizing overfitting and enhancing generalization. We finally identify a crucial insight: our method implicitly learns *a vital prior that fakes are actually derived from the real*, indicating a hierarchical relationship rather than independence. Modeling this prior, we believe, is essential for achieving superior generalization. Our codes are publicly available at GitHub.

[*]Equal contribution [1]Peking University Shenzhen Graduate School [2]Tencent Youtu Lab [3]The Chinese University of Hong Kong, Shenzhen. Correspondence to: Taiping Yao <taipingyao@tencent.cn>, Li Yuan <yuanli-ece@pku.edu.cn>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

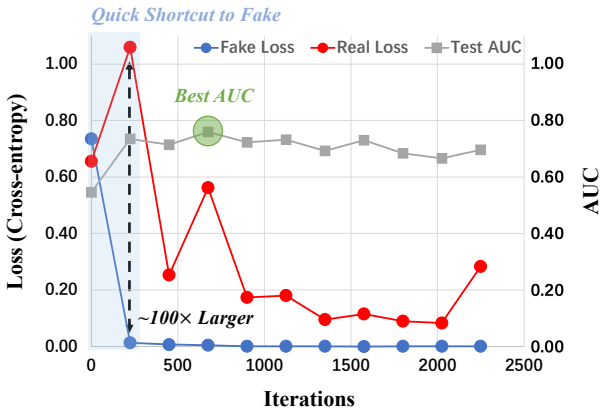

Figure 1. **Illustration of the asymmetry phenomenon in AI-generated image detection**. We show that the baseline detector (*i.e.,* Xception) tends to **quickly overfit to the fake patterns** in the training set (Rossler et al., 2019), causing the limited generalization when facing previously unseen fakes (Li et al., 2020b).

1. Introduction

The rapid development of AI generative technologies has significantly lowered the barrier for creating highly realistic fake images. As deep generative models advance and mature (Goodfellow et al., 2020; Ho et al., 2020; Rombach et al., 2022b; Yan et al., 2025), the proliferation of AI-generated images (AIGIs[1]) has drawn considerable attention, driven by their ability to produce high-quality content with relative ease. However, these advancements also introduce significant risks, if misused for malicious purposes such as deepfakes (mainly including face-swapping (Korshunov & Marcel, 2018) and face-reenactment (Thies et al., 2016)), which may violate personal privacy, spread misinformation, and erode trust in digital media. Consequently, there is an urgent need to develop a reliable and robust framework for detecting AIGIs.

Most existing studies in AIGI detection (Wang et al., 2020b; Rossler et al., 2019) typically approach the real/fake classification problem as a symmetric binary classification task, akin to the "cat versus dog" problem. A standard binary

[1]In the context of this research, AIGI primarily refers to deepfakes (face-swapping) and synthetic images (*e.g.,* nature or arts).

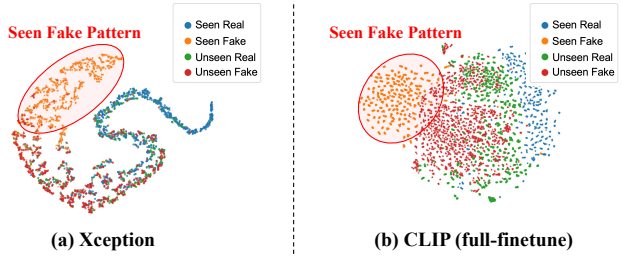

Figure 2. **t-SNE visualizations between Xception and CLIP (full-finetune)**. We show that both models only learn the specific fake patterns within the training set, treating samples with seen fake patterns as fake while other samples are all considered real, thereby limiting their generalization in detecting unseen fakes.

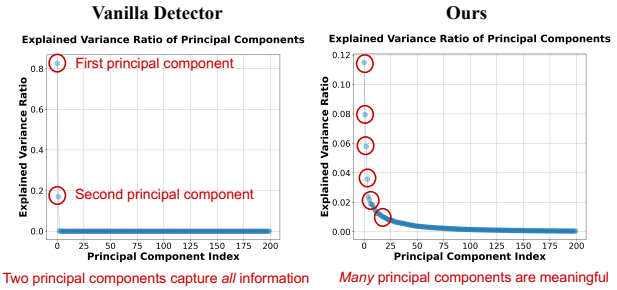

Figure 3. **Analysis for the *effective information* contained in the model's feature space.** We apply PCA for dimension reduction and visualize the explained variance ratio of principal components with high contribution. We show that the baseline model trained on the AIGI dataset can be highly constrained and low-ranked.

classifier, often based on deep neural networks, is trained to distinguish between real and fake images by predicting the likelihood of a given test image being fake during inference. Although this paradigm yields promising results when the training and testing distributions (in terms of fake generation methods) are similar, its performance tends to degrade significantly when applied to previously unseen fake methods, indicating the generalization issue (Ojha et al., 2023).

To understand the *underlying reasons* for the failure in generalization, we have conducted extensive preliminary investigations and identified an *asymmetry phenomenon* in AIGI detection: naively trained detectors tend to take the shortcut and very quickly overfit the limited fake patterns presented in the training set. Visualization in Fig. 1 corroborates this claim. Specifically, the vanilla detector (*i.e.,* Xception (Rossler et al., 2019)) quickly fits the fake patterns at the very early training stage (only a few iterations), resulting in a very low loss of fake, while the real loss is significantly higher than the fake loss ($\sim 100\times$ larger). This is likely because existing AIGI detection datasets (Rossler et al., 2019; Wang et al., 2020b) typically contain limited and homogeneous fake types, while real samples exhibit significantly greater diversity and variance between each other such as different categories and scenarios.

Consequently, the learned feature spaces become **fake-dominated and thus highly constrained.** As evidenced by the t-SNE visualization in Fig. 2, we see that the whole feature space is indeed dominated by the forgery patterns, where both the Xception (Vanilla CNN) and CLIP (Radford et al., 2021) detector group only the specific fake patterns within the training set into a single cluster, while all other data, including real samples and fake samples from unseen forgeries, are mapped into a separate cluster.

To quantify this, we analyze the *effective information*[2] contained in the model's feature space via **Principal Component Analysis (PCA)**. Specifically, we visualize the *explained variance ratio* of different principal components for the model's feature space in Fig. 3. The results show that the feature space of the naive detector can be highly constrained and low-ranked, with **only two principal components to capture all information**, resulting in limited generalization. This aligns with the previous theoretical analysis (Gunasekar et al., 2017) that *low-ranked*[3] *feature spaces hinder generalization by memorizing trivial patterns.*

One potential remedy to address the low-ranked problem is incorporating pre-trained knowledge within vision foundation models (VFMs), which provide higher-ranked feature representations, to **expand the low-ranked feature space**, thus alleviating the overfitting. However, naively fine-tuning a VFM (even CLIP) risks distorting the original rich representation feature space (Fig. 2), pushing the feature space to become low-ranked again (verified in Fig. 6).

To address this, we design a novel approach called **Effort**: **Eff**icient **ort**hogonal modeling for generalizable AIGI detection. Specifically, we employ Singular Value Decomposition (SVD) to **construct two orthogonal subspaces**. By freezing the principal components and adapting the remained components, we preserve the pre-trained knowledge while learning forgery-related patterns.

We have conducted extensive experiments on both deepfake detection and synthetic image detection benchmarks and find that our approach achieves significant superiority over other SOTAs with very little training cost. Compared to existing full-parameters and LoRA-based tuning methods, we explicitly ensure orthogonality, enabling the higher rank of the whole feature space, effectively minimizing overfitting to fake and enhancing generalization.

Finally, we arrive at **a key insight for generalizable AIGI detection**: there exists an important prior of the detection

[2] In PCA, it refers to variance captured by principal components. Larger eigenvalues indicate components explaining more variance and contributing more to data representation.

[3] The "rank" here means the number of significant principal components.

task, where fake images are generated from real ones, establishing a **hierarchical relationship** rather than an independent or symmetric one. When aligning semantic information-such as distinguishing a fake dog from a real dog-this prior allows the model to focus discrimination within a smaller, semantically consistent subspace, e.g., only among dogs (see Fig. 4 for illustration). This focused discrimination simplifies the task and aligns with theoretical results from Rademacher complexity (Mohri & Rostamizadeh, 2008), which states that reducing model complexity leads to tighter generalization bounds. In contrast, naively trained detectors that treat real and fake data as independent fail to capture this structure, resulting in limited generalization performance. Therefore, modeling this hierarchical prior, we believe, is vitally crucial for AIGI detection.

Our work makes the following key contributions:

- **Asymmetry phenomenon in AIGI detection**: We introduce the concept of *asymmetry phenomenon*, where a naively trained detector tends to quickly fit the seen fake methods well but, in doing so, it often overfits to specific fake patterns in the training set, limiting its generalization ability to detect unseen fake methods.

- **New perspective to explain the failure reasons behind generalization**: We use PCA to quantitatively assess the *effective information* within the learned feature space, and we find that the overfitting to fake, results in a *highly low-ranked and constrained feature space*, thus leading to the limited generalization capability.

- **Novel method via orthogonal subspace decomposition**: To address the overfitting, we propose a novel approach, *Effort*, with two careful designs: (1) incorporating the pre-trained knowledge (proving higher-ranked feature space) within the vision foundation models to expand the previous feature space, improving the model's expressivity and alleviating the overfitting; and (2) employing SVD to *explicitly* construct two *orthogonal subspaces*, where the principal one for preserving pre-trained knowledge and the remained one for learning new forgeries, avoiding the distortion of original rich feature space during learning fakes.

- **Key insight toward generalizable detection**: We reveal that fake data is actually generated from real data, forming a hierarchical relationship rather than being independent. Our method effectively models this prior by maintaining the pre-trained semantic components while adapting to fake detection effectively, enabling the detector to make discrimination on the semantic-aligned subspaces, reducing model complexity and thus improving generalization.

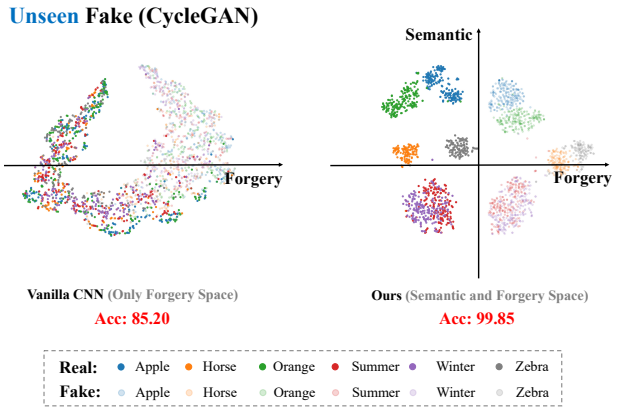

Figure 4. **t-SNE visualizations of the latent feature spaces between vanilla CNN (Wang et al., 2020b) and ours.** Our method achieves orthogonal learning between the dimensions of semantic and forgery, allowing the model to capture fake patterns on the semantically-aligned subspace, simplifying the discrimination and thereby improving the generalization.

2. Related Work

Our work focuses on detecting AI-generated images (AIGIs), especially **deepfake images** (*e.g.,* face-swapping) and **synthetic images** (*e.g.,* nature or art), following Yan et al. (2024a). As the majority of recent works specifically focus on dealing with the generalization issue, where the training and testing distribution differ (in terms of fake methods), we will briefly introduce the classical and recent detection methods toward generalization in deepfake and synthetic images, respectively.

Generalizable Deepfake Image Detection. The task of deepfake detection grapples profoundly with the issue of generalization. To tackle the generalization issue, one mainstream approach is *fake pattern learning*. Most existing works are within this line. These methods generally design a "transformation function", *e.g.,* frequency transformation (Li et al., 2021; Luo et al., 2021; Liu et al., 2021a), blending operations (Li et al., 2020a; Zhao et al., 2021; Shiohara & Yamasaki, 2022; Chen et al., 2022a), reconstruction (Zhu et al., 2021; Cao et al., 2022), content/ID disentanglement (Yan et al., 2023a; Fu et al., 2025; Huang et al., 2023; Dong et al., 2023), to transform the original input x into x', where they believe that the more general fake patterns can be captured within the feature space of x' compared to x. However, given the ever-increasing diversity of forgery methods in the real world, it is unrealistic to elaborate all possible fake patterns and "expect" good generalization on unseen fake methods. Another notable direction is to *real distribution learning*, with a specific methodology involved: one-class anomaly detection (Khalid & Woo, 2020; Larue et al., 2023). Specifically, Khalid & Woo (2020) introduced a one-class-based anomaly detection, where "ab-

normal" data is detected by the proposed reconstruction error as the anomaly score. Larue et al. (2023) proposed a similar approach to create pseudo-fake "anomaly" samples by using image-level blending on different facial regions. However, it is challenging to ensure that the detector can learn a robust representation of real images by using the very limited real data in existing deepfake datasets (*e.g.,* the FF++ dataset (Rossler et al., 2019) contains only 1,000 real videos with imbalanced facial attribute distributions (Trinh & Liu, 2021)).

Generalizable Synthetic Image Detection. With the rapid advancement of existing AI generative technologies, the scope of forged content has expanded beyond facial forgeries to encompass a wide range of scenes. In this context, similar to the deepfake detection field, most existing works typically focus on *fake pattern learning* that mines the low-level forgery clues from different aspects. Specifically, several approaches have been proposed to capture low-level artifacts, including RGB data augmentations (Wang et al., 2020b), frequency-based features (Jeong et al., 2022), gradients (Tan et al., 2023), reconstruction artifacts (Wang et al., 2023a; Chen et al., 2024; Luo et al., 2024), and neighboring pixel relationships (Tan et al., 2024c), random-mapping feature (Tan et al., 2024a). To illustrate, BiHPF (Jeong et al., 2022) amplifies artifact magnitudes through the application of dual high-pass filters, while LGrad (Tan et al., 2023) uses gradient information from pre-trained models as artifact representations. NPR (Tan et al., 2024c) introduces a straightforward yet effective artifact representation by rethinking up-sampling operations. In addition to learning from scratch, there are also several research works (Ojha et al., 2023; Wu et al., 2023; Liu et al., 2024) that perform *fake pattern learning* by leveraging the existing vision foundation models. For instance, UniFD (Ojha et al., 2023) directly freezes the visual encoder of the pre-trained CLIP model and tunes only a linear layer for binary classification, demonstrating effective deepfake detection even with previously unseen sources. LASTED (Wu et al., 2023) proposes designing textual labels to supervise the CLIP vision model through image-text contrastive learning, advancing the field of synthetic image detection. These arts have shown notable improvement in generalization performance when facing previously unseen fake methods.

3. Methodology

The overall pipeline of the proposed *Effort* approach is illustrated in Fig. 5, aiming to address the asymmetry phenomena in AIGI detection. Our approach involves the SVD to construct explicit orthogonality for preserving pre-trained knowledge and learning forgery-related patterns, avoiding the distortion of well-learned pre-trained knowledge during learning forgeries.

Formally, given a pre-trained weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$ for a certain linear layer, we perform SVD to decompose W:

$$W = U\Sigma V^\top, \tag{1}$$

where $U \in \mathbb{R}^{d_1 \times d_1}$ and $V \in \mathbb{R}^{d_2 \times d_2}$ are orthogonal matrices containing the left and right singular vectors, respectively, and $\Sigma \in \mathbb{R}^{d_1 \times d_2}$ is a diagonal matrix with singular values on the diagonal. Since the linear layer of VFM generally has the same input and output dimensions, we consider the case of SVD with $d_1 = d_2 = n$ in the following discussion.

To obtain a rank-r approximation of the pre-trained weight matrix, we retain only the top r singular values and corresponding singular vectors:

$$W \approx W_r = U_r \Sigma_r V_r^\top, \tag{2}$$

where $U_r \in \mathbb{R}^{n \times r}$, $\Sigma_r \in \mathbb{R}^{r \times r}$, and $V_r \in \mathbb{R}^{n \times r}$. We keep W_r frozen during training to preserve dominant pre-trained knowledge learned from large-scale data.

The residual component, defined as the difference between the pre-trained weights and the SVD approximation, is used to learn representations specific to fake images:

$$\Delta W = W - W_r = U_{n-r} \Sigma_{n-r} V_{n-r}^\top, \tag{3}$$

where $U_{n-r} \in \mathbb{R}^{n \times (n-r)}$, $\Sigma_{n-r} \in \mathbb{R}^{(n-r) \times (n-r)}$, $V_{n-r} \in \mathbb{R}^{n \times (n-r)}$. It is important to note that ΔW represents a learnable form associated with the remaining singular value decomposition, reflecting slight modifications or perturbations to the original weight matrix.

During training, we only optimize ΔW while keeping U_r, Σ_r, and V_r fixed. This implementation ensures that the model retains its capability to process real images via the SVD approximation and adapts to detect deepfakes through the trivial residual components of the weight matrix.

To encourage the ΔW to capture both useful and meaningful discrepancy between the real and fake, it's significant to guarantee that optimizing ΔW does not change the properties of the overall weight matrix W (*i.e.*, Minimize the impact on the real information of the pre-trained weight as much as possible). Thus, we proposed two constraints to realize this goal, as follows.

Orthogonal Constraint. We maintain the orthogonality among each singular vector to keep orthogonal subspace for learning real/fake:

$$\mathcal{L}_{\text{orth}} = \|\hat{U}^\top \hat{U} - I\|_F^2 + \|\hat{V}^\top \hat{V} - I\|_F^2, \tag{4}$$

where $\hat{U} \in \mathbb{R}^{n \times n}$ denote the concatenation of U_r and U_{n-r} along the row dimension, $\hat{V} \in \mathbb{R}^{n \times n}$ denote the concatenation of V_r and V_{n-r} along the row dimension, $\|\cdot\|_F$ denotes

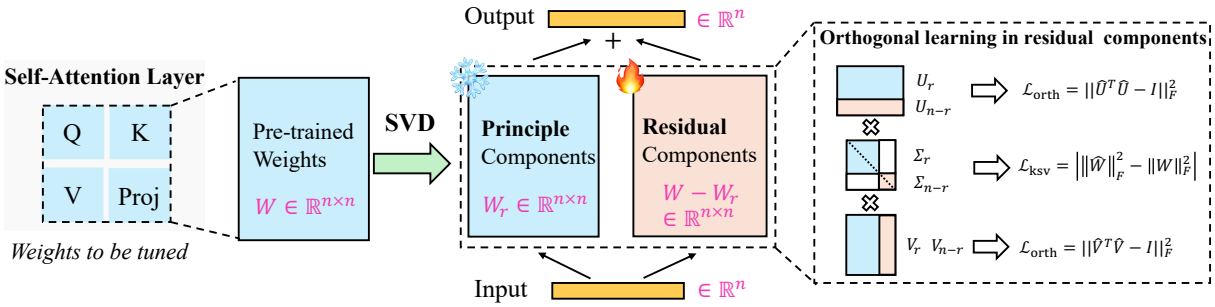

Figure 5. **The proposed approach for AIGI detection**. The left branch is the decomposition matrix of the principle components approximation using SVD, while the right residual branch enables the orthogonal learning of real/fake discriminative features.

the Frobenius norm, and I is the identity matrix of appropriate dimensions.

Singular Value Constraint. The singular values can be interpreted as a type of scaling that affects the magnitude of the corresponding singular vectors. There is a relationship between singular values and the Frobenius norm of the weight matrix being decomposed:

$$\|W\|_F = \sqrt{\sum_i \sigma_i^2}, \tag{5}$$

where σ_i denotes the i-th singular value of the corresponding weight matrix.

To maximize the reduction of the impact of real knowledge, we constrain the singular values of the optimized weight matrix $\hat{W}$ to remain consistent with those of the original weight matrix W:

$$\mathcal{L}_{\text{ksv}} = \left| \sum_{i=r+1}^{n} \hat{\sigma}_i^2 - \sum_{i=r+1}^{n} \sigma_i^2 \right| = \left| \|\hat{W}\|_F^2 - \|W\|_F^2 \right|, \tag{6}$$

where $\hat{W}$ represents the weights after the optimization of W, and $|\cdot|$ represents the absolute value. Note that this regularization will control the importance of the ΔW during optimization to prevent overfitting of learning real/fake.

Loss Function. The overall loss function for training the model combines the classification loss $\mathcal{L}_{\text{cls}}$ (e.g., cross-entropy loss for binary classification) and the orthogonality regularization loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \frac{1}{m} \sum_i^m \mathcal{L}_{\text{orth}}^i + \lambda_2 \frac{1}{m} \sum_i^m \mathcal{L}_{\text{ksv}}^i, \tag{7}$$

where λ_1, λ_2 are hyperparameters that balance the importance of the corresponding regularization term, and m represents the number of pre-trained weight matrices on which our approach is applied. In practice, we adapt our approach to the *linear layers* within the self-attention module across all transformer layers of the VFM to leverage their rich, well-learned real distributions.

Finally, we provide an algorithm illustration of the proposed approach in Alg. 1 for an overall understanding.

Algorithm 1 *Effort* Approach Algorithm

Input: Pre-trained weight matrix $W \in \mathbb{R}^{n\times n}$; Rank r; Training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$; Hyperparameters λ_1, λ_2
Output: Updated weight matrix W

▷ **Step 1: Singular Value Decomposition**
Decompose W via SVD: $W = U\Sigma V^\top$
Retain top r singular values and vectors:
  $U_r \in \mathbb{R}^{n\times r}$, $\Sigma_r \in \mathbb{R}^{r\times r}$, $V_r \in \mathbb{R}^{n\times r}$
Compute $W_r = U_r \Sigma_r V_r^\top$
Keep W_r fixed during training
Compute residual component: $\Delta W = W - W_r$
Decompose ΔW via SVD: $\Delta W = U_{n-r}\Sigma_{n-r}V_{n-r}^\top$
Initialize ΔW
Define concatenated matrices:
  $\hat{U} = [U_r, U_{n-r}] \in \mathbb{R}^{n\times n}$
  $\hat{V} = [V_r, V_{n-r}] \in \mathbb{R}^{n\times n}$
▷ **Step 2: Training Loop**
for each epoch **do**
  for each batch in $\mathcal{D}$ **do**
    ▷ Forward Pass
    Compute model output using $W = W_r + \Delta W$
    Compute classification loss $\mathcal{L}_{\text{cls}}$
    ▷ Compute Constraints
    Compute orthogonality loss:
      $\mathcal{L}_{\text{orth}} = \left\|\hat{U}^\top\hat{U} - I\right\|_F^2 + \left\|\hat{V}^\top\hat{V} - I\right\|_F^2$
    Compute singular value constraint loss:
      $\mathcal{L}_{\text{ksv}} = \left|\left\|\hat{W}\right\|_F^2 - \|W\|_F^2\right|$
    ▷ Total Loss
    $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{orth}} + \lambda_2\mathcal{L}_{\text{ksv}}$
    ▷ Backward Pass and Optimization
    Update ΔW using gradient descent to minimize $\mathcal{L}$
  end for
end for
return Updated weight matrix $W \leftarrow W_r + \Delta W$

4. Experiment

4.1. Deepfake Image Detection

Implementation Details. We utilize CLIP ViT-L/14 (Radford et al., 2021) as the default vision foundation model (VFM). We also investigate other VFMs in Tab. 5. We follow the pre-processing and training pipeline and use the

Table 1. **Benchmarking Results of Cross-dataset Evaluations (Protocol-1) and Cross-method Evaluations (Protocol-2)**. All detectors are trained on FF++_c23 (Rossler et al., 2019) and evaluated on other fake data. † indicates the results are obtained by using the model's checkpoint provided by the authors, otherwise, the results are cited from (Yan et al., 2023b; 2024a; Cheng et al., 2024).

Methods	Trainable	Cross-dataset Evaluation								Cross-method Evaluation								
	Param.	CDF-v2	DFD	DFDC	DFDCP	DFo	WDF	FFIW	Avg.	UniFace	BleFace	MobSwap	e4s	FaceDan	FSGAN	InSwap	SimSwap	Avg.
F3Net (Qian et al., 2020)	22M	0.789	0.844	0.718	0.749	0.730	0.728	0.649	0.743	0.809	0.808	0.867	0.494	0.717	0.845	0.757	0.674	0.746
SPSL (Liu et al., 2021a)	21M	0.799	0.871	0.724	0.770	0.723	0.702	0.794	0.769	0.747	0.748	0.885	0.514	0.666	0.812	0.643	0.665	0.710
SRM (Luo et al., 2021)	55M	0.840	0.885	0.695	0.728	0.722	0.702	0.794	0.767	0.749	0.704	0.779	0.704	0.659	0.772	0.793	0.694	0.732
CORE (Ni et al., 2022)	22M	0.809	0.882	0.721	0.720	0.765	0.724	0.710	0.762	0.871	0.843	0.959	0.679	0.774	0.958	0.855	0.724	0.833
RECCE (Cao et al., 2022)	48M	0.823	0.891	0.696	0.734	0.784	0.756	0.711	0.779	0.898	0.832	0.925	0.683	0.848	0.949	0.848	0.768	0.844
SLADD (Chen et al., 2022a)	21M	0.837	0.904	0.772	0.756	0.800	0.690	0.683	0.777	0.878	0.882	0.954	0.765	0.825	0.943	0.879	0.794	0.865
SBI (Shiohara & Yamasaki, 2022)	18M	0.886	0.827	0.717	0.848	0.899	0.703	0.866	0.821	0.724	0.891	0.952	0.750	0.594	0.803	0.712	0.701	0.766
UCF (Yan et al., 2023a)	47M	0.837	0.867	0.742	0.770	0.808	0.774	0.697	0.785	0.831	0.827	0.950	0.731	0.862	0.937	0.809	0.647	0.824
IID (Huang et al., 2023)	66M	0.838	0.939	0.700	0.689	0.808	0.666	0.762	0.789	0.839	0.789	0.888	0.766	0.844	0.927	0.789	0.644	0.811
LSDA† (Yan et al., 2024a)	133M	0.875	0.881	0.701	0.812	0.768	0.797	0.724	0.794	0.872	0.875	0.930	0.694	0.721	0.939	0.855	0.793	0.835
ProDet† (Cheng et al., 2024)	96M	0.926	0.901	0.707	0.828	0.879	0.781	0.751	0.828	0.908	0.929	0.975	0.771	0.747	0.928	0.837	0.844	0.867
CDFA† (Lin et al., 2024)	87M	0.938	0.954	0.830	0.881	0.973	0.796	0.777	0.878	0.762	0.756	0.823	0.631	0.803	0.942	0.772	0.757	0.781
Effort (Ours)	0.19M	0.956	0.965	0.843	0.909	0.977	0.848	0.921	0.917	0.962	0.873	0.953	0.983	0.926	0.957	0.936	0.926	0.940

Table 2. **Benchmarking Results of Cross-method Evaluations in terms of AP Performance on the UniversalFakeDetect Dataset.** † indicates that the results are obtained by using the official pre-trained model or reproduction.

Methods	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 (Wang et al., 2020b)	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 (Chai et al., 2020)	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-occurence (Nataraj et al., 2019)	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 (Zhang et al., 2019)	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
F3Net† (Qian et al., 2020)	99.96	84.32	69.90	99.72	56.71	100.0	78.82	52.89	46.70	63.39	64.37	70.53	73.76	81.66	74.62	89.81	91.04	90.86	71.84	76.89
UniFD (Ojha et al., 2023)	100.0	98.13	94.46	86.66	99.25	99.53	91.67	78.54	67.54	83.12	91.06	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	90.14
LGrad† (Tan et al., 2023)	100.0	93.98	90.69	99.86	79.36	99.98	67.91	59.42	51.42	63.52	69.61	87.06	99.03	99.16	99.18	93.23	95.10	94.93	97.23	86.35
FreqNet† (Tan et al., 2024b)	99.92	99.63	96.05	99.89	99.71	98.63	99.92	94.42	74.59	80.10	75.70	96.27	96.06	100.0	62.34	99.80	99.78	96.39	77.78	91.95
NPR† (Tan et al., 2024c)	100.0	99.53	94.53	99.94	88.82	100.0	84.41	97.95	99.99	50.16	50.16	98.26	99.92	99.91	99.92	99.87	99.89	99.92	99.26	92.76
FatFormer† (Liu et al., 2024)	100.0	100.0	99.98	99.75	100.0	100.0	97.99	97.94	81.21	99.84	99.93	91.99	99.81	99.09	99.87	99.13	99.41	99.20	99.82	98.16
Effort (Ours)	100.0	100.0	99.99	99.77	100.0	100.0	98.95	97.53	97.53	100.0	100.0	95.39	99.99	99.89	100.0	99.87	99.92	99.98	99.96	99.41

Table 3. **Benchmarking Results of Cross-method Evaluations in terms of Acc Performance on the UniversalFakeDetect Dataset.** † indicates that the results are obtained by using the official pre-trained model or reproduction.

Methods	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
	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 (Wang et al., 2020b)	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 (Chai et al., 2020)	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-occurence (Nataraj et al., 2019)	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 (Zhang et al., 2019)	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
F3Net† (Qian et al., 2020)	99.38	76.38	65.33	92.56	58.10	100.0	63.48	54.17	47.26	51.47	51.47	69.20	68.15	75.35	68.80	81.65	83.25	83.05	66.30	71.33
UniFD (Ojha et al., 2023)	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
LGrad† (Tan et al., 2023)	99.84	85.39	82.88	94.83	72.45	99.62	58.00	62.50	50.00	50.74	50.78	77.50	94.20	95.85	94.80	87.40	90.70	89.55	88.35	80.28
FreqNet† (Tan et al., 2024b)	97.90	95.84	90.45	97.55	90.24	93.41	97.40	88.92	59.04	71.92	67.35	86.70	84.55	99.58	65.56	85.69	97.40	88.15	59.06	85.09
NPR† (Tan et al., 2024c)	99.84	95.00	87.55	96.23	86.57	99.75	76.89	66.94	98.63	50.00	50.00	84.55	97.65	98.00	98.20	96.25	97.15	97.35	87.15	87.56
FatFormer† (Liu et al., 2024)	99.89	99.32	99.50	97.15	99.41	99.75	93.23	81.11	68.04	69.45	69.45	76.00	98.60	94.90	98.65	94.35	94.65	94.20	98.75	90.86
Effort (Ours)	100.0	99.85	99.60	95.05	99.60	100.0	87.60	92.50	81.50	98.90	98.90	69.15	99.30	96.80	99.45	97.45	97.80	97.15	98.05	95.19

codebases of DeepfakeBench (Yan et al., 2023b). Additionally, we sample 8 frames from each video for training and 32 frames for inference, following (Shiohara & Yamasaki, 2022). We use the fixed learning rate of 2e-4 for training our approach and employ the Adam (Kingma & Ba, 2014) for optimization. We set the batch size to 32 for both training and testing. We also employ several widely used data augmentations, such as Gaussian Blur and Image Compression, following other existing works (Yan et al., 2024a; Shiohara & Yamasaki, 2022; Cheng et al., 2024). For the evaluation metric, we report the widely-used video-level Area Under the Curve (AUC) to compare our approach with other works, following (Lin et al., 2024; Shiohara & Yamasaki, 2022). We compute the average model's output probabilities of each video to obtain the video-level AUC.

Evaluation Protocols and Dataset. We adopt two widely used and standard protocols for evaluation: **Protocol-1:** cross-dataset evaluation and **Protocol-2:** cross-manipulation evaluation within the FF++ data domain. For **Protocol-1**, we conduct evaluations by training the models on FaceForensics++ (FF++) (Rossler et al., 2019) and testing them on other seven deepfake detection datasets: Celeb-DF-v2 (CDF-v2) (Li et al., 2020b), DeepfakeDetection (DFD) (DFD., 2020), Deepfake Detection Challenge (DFDC) (detection challenge., 2020), the preview version of DFDC (DFDCP) (Dolhansky et al., 2019), DeeperForensics (DFo) (Jiang et al., 2020), WildDeepfake (WDF) (Zi et al., 2020), and FFIW (Zhou et al., 2021). Note that FF++ has three different compression versions and we adopt the c23 version for training all methods in our experiments, follow-

ing most existing works (Yan et al., 2024a). For **Protocol-2**, we evaluate the models on the latest deepfake dataset DF40 (Yan et al., 2024b), which contains the forgery data generated within the FF++ domain, ensuring the fake methods different while the data domains remain unchanged.

Evaluation Benchmarking. To provide a comprehensive benchmark for comparison, we introduce **13 competitive detectors**, including several classical detection methods such as F3Net (Qian et al., 2020) (ECCV'20), SPSL (Liu et al., 2021a) (CVPR'20), SRM (Liu et al., 2021a) (CVPR'21), CORE (Ni et al., 2022) (CVPRW'22), RECCE (Cao et al., 2022) (CVPR'22), and SBI (Shiohara & Yamasaki, 2022) (CVPR'22), and also several latest SOTA methods (after 2023), such as UCF (Yan et al., 2023a) (ICCV'23), IID (Huang et al., 2023) (CVPR'23), TALL (Xu et al., 2023) (ICCV'23), LSDA (Yan et al., 2024a) (CVPR'24), ProDet (Cheng et al., 2024) (NeurIPS'24), and CFDA (Lin et al., 2024) (ECCV'24). All detectors are trained on FF++ (c23) and tested on other fake data. Results in Tab. 1 demonstrate two notable advantages of our approach. **(1) generalizability**: we see that our approach consistently and largely outperforms other models across basically all tested scenarios, validating that our method is generalizable for detecting unseen fake data, even for the latest face-swapping techniques such as BleFace (Shiohara et al., 2023). **(2) efficiency**: it is worth noting that our method only needs 0.19M parameters for training to achieve superior generalization. As we can see most latest SOTA detectors such as LSDA and ProDet all use about 100M parameters for training, while we are about **1,000× smaller**.

4.2. Synthetic Image Detection

Evaluation Metrics. We follow existing works (Wang et al., 2020a; Ojha et al., 2023; Liu et al., 2024) for benchmarking and report both average precision (AP) and classification accuracy (Acc). For Acc, we set the classification threshold for each dataset to 0.5 to ensure a fair comparison.
UniversalFakeDetect Dateset. We adhere to the protocol outlined in (Wang et al., 2020a; Ojha et al., 2023) and utilize ProGAN's real and fake images as our training dataset, which includes 20 subsets of generated images. The evaluation set contains 19 subsets derived from different kinds of generative models, including ProGAN (Karras et al., 2018), CycleGAN (Zhu et al., 2017), BigGAN (Brock et al., 2018a), StyleGAN (Karras et al., 2019), GauGAN (Park et al., 2019), StarGAN (Choi et al., 2018), DeepFakes (Rössler et al., 2019), SITD (Chen et al., 2018), SAN (Dai et al., 2019), CRN (Chen & Koltun, 2017), IMLE (Li et al., 2019), Guided (guided diffusion model)(Dhariwal & Nichol, 2021), LDM (latent diffusion model) (Rombach et al., 2022a), Glide (Nichol et al., 2022), and DALLE (Ramesh et al., 2021).
Implementation Details. Similar to the setting of deepfake image detection, we adopt pre-trained CLIP ViT-L/14 as the backbone and use the Adam optimizer (Kingma & Ba, 2014) with a fixed learning rate of 2e-4. The batch size is set to 48. Other settings and details are the same with (Ojha et al., 2023).

Table 4. **Ablation studies regarding the proposed SVD, singular value constraint ($\mathcal{L}_{\text{ksv}}$), and orthogonal constraint ($\mathcal{L}_{\text{orth}}$).** All models are trained on FF++ (c23) and tested on other datasets.

Ours			CDF-v2	SimSwap	Avg.
SVD	$\mathcal{L}_{\text{ksv}}$	$\mathcal{L}_{\text{orth}}$			
×	×	×	0.857	0.860	0.859
✓	×	×	0.940	0.910	0.925
✓	✓	×	0.944	0.927	0.936
✓	×	✓	0.945	0.914	0.930
✓	✓	✓	0.956	0.926	0.941

Table 5. **Ablation studies regarding different vision foundation models (VFMs) were used**. All models are trained on FF++ (c23) and tested on CDF-v2 and SimSwap.

VFMs	#Params	#ImgSize	CDF-v2	SimSwap	Avg.
BEIT-v2 (Peng et al., 2022)	303M	224	0.855	0.821	0.838
+ Ours	0.14M	224	0.894	0.850	0.872
SigLIP (Zhai et al., 2023)	316M	256	0.877	0.713	0.795
+ Ours	0.19M	256	0.895	0.778	0.867
CLIP (Radford et al., 2021)	307M	224	0.857	0.860	0.859
+ Ours	0.19M	224	0.956	0.926	0.941

Evaluation Analysis. The AP and Acc results are presented in Tab. 2 and Tab. 3, respectively. Our method attains impressive detection results, achieving 95.19% mAcc and 99.41% mAP across the 19 test subsets. One similar approach to ours is UniFD, which also preserves the original pre-trained knowledge of CLIP and fine-tunes only the FC layer for discrimination. In contrast to UniFD which directly performs discrimination in the pre-trained knowledge space, our approach utilizes SVD to create an orthogonal low-ranked subspace for learning forgeries while preserving the essential high-ranked representational space, achieving the discrimination by leveraging both representational and forgery subspaces, achieving a significant improvement of 9.27% in mAcc and 13.81% in mAP over UniFD. Besides, when compared to the SOTA method, FatFormer, we achieve 4.33% mAcc improvement without relying on the extra text encoder of CLIP. This further demonstrates the superiority of our approach.

4.3. Ablation Study and Analysis

Incremental improvement of the proposed designs. Ablation studies in Tab. 4 demonstrate incremental gains from the proposed SVD method and loss constraints: the baseline (no modules) achieves an average AUC of 0.859, while adding SVD alone boosts performance to 0.925 (+6.6%), underscoring its efficacy in isolating forgery artifacts via orthogonal feature decomposition. Further integrating the singular value constraint ($\mathcal{L}_{\text{ksv}}$) and orthogonal constraint ($\mathcal{L}_{\text{orth}}$) refines performance, yielding 0.936 and 0.930 AUC, respectively, with their combined synergy (SVD + $\mathcal{L}_{\text{ksv}}$ +

Table 6. **Ablation studies on synthetic image detection regarding the tunable $n - r$ values in SVD (Ours) and r values in LoRA**. All models are trained on ProGAN's images and tested on 19 different generative models' images. "FFT" indicates the full fine-tuning. "Linear-Prob" indicates fine-tuning FC layer only, where we reproduce the results from UniFD (Ojha et al., 2023).

Archs.		$n-r$	r	mAcc
UniFD (Linear-Prob)		–	–	81.02
Baseline (FFT)		–	–	86.22
LoRA	Variant-1	–	256	91.42
	Variant-2	–	64	91.06
	Variant-3	–	16	91.89
	Variant-4	–	4	93.53
	Variant-5	–	1	93.03
Ours	Variant-1	256	–	92.13
	Variant-2	64	–	93.68
	Variant-3	16	–	94.45
	Variant-4	4	–	94.37
	Variant-5	1	–	95.19

Table 7. **Cross-dataset generalization evaluations with existing *adapter-based* deepfake detectors**. The results are cited from their original papers. The metric is **frame-level AUC.**

Methods	CDF-v2	DFD	DFDC	Avg.
LoRA (Kong et al., 2023)	0.838	0.834	0.717	0.796
MoE-LoRA (Kong et al., 2024)	0.867	0.904	–	–
Dual-Adapter (Shao et al., 2023)	0.717	–	0.727	–
Ours	0.901	0.923	0.798	0.874

$\mathcal{L}_{\text{orth}}$) achieving peak performance (0.941 AUC). These results highlight the significant contribution of SVD as the primary driver of performance gains while illustrating the complementary benefits of $\mathcal{L}_{\text{ksv}}$ and $\mathcal{L}_{\text{orth}}$ in further refining the generalization performance.

Compatibility with other vision foundation models. By default, we choose CLIP as the vision foundation model (VFM) in our experiments. To validate the generality and versatility of our approach, we conduct an ablation study to apply our method to other VFMs, including BEIT-v2 (Peng et al., 2022) and SigLIP (Zhai et al., 2023). Results in Tab. 5 show that our approach can be seamlessly applied to other VFMs to improve the model's generalization performance.

Comparison with Existing Adapter-Based Detectors. Parameter-efficient fine-tuning (PEFT) has become a popular technique for adapting pre-trained large models to downstream tasks (Ding et al., 2023). Low-ranked adaptation (LoRA) (Hu et al., 2021) is a widely used approach for PEFT. Previous works (Kong et al., 2023; Liu et al., 2024) employing LoRA in the VFMs have achieved good empirical generalization results for detection. Additionally, (Kong et al., 2024) and (Shao et al., 2023) introduce MoE-LoRA techniques and dual adapters into the deepfake detection fields. However, the existing adapter-based methods do **not explicitly ensure this orthogonality**, still having the potential to distort the pre-existing pre-trained knowledge and result in unexpected generalization results. In contrast, our

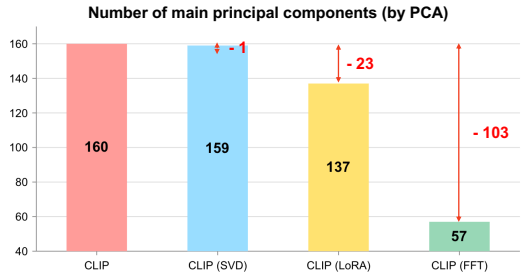

Figure 6. **PCA quantifying feature space information retention, measured by the minimum number of components required to explain $\geq 90\%$ variance.** Our SVD-based fine-tuning preserves 159/160 principal components (retaining 99.3% original variance), while LoRA and FFT exhibit notable degradation, with 14.4% and 64.4% variance loss, respectively. See Tab. 8 for more results.

Table 8. **Comparison of different fine-tuning methods.** We compute both the effective rank and mean accuracy for evaluation. Our SVD retains most principal components (notably higher effective rank), thus achieving the best detection results.

Fine-tuning Methods	Face Domain		General Domain	
	Effective Rank	Accuracy	Effective Rank	Accuracy
Baseline (No FT)	160	-	479	-
SVD (ours)	159	95.60	316	95.19
LoRA	137	89.40	304	93.03
FFT	57	85.70	238	86.22

method **explicitly** constructs two **orthognoal subspaces** for pre-trained knowledge and forgery using SVD, ensuring the pre-existing pre-trained knowledge will not be distorted, thereby achieving better generalization performance. To verify this, we provide several empirical results in Tab. 7 and Tab. 6. From these results, we can see that our proposed SVD-based method achieves clearly higher generalization results than adapter-based methods in both deepfake detection and synthetic image detection fields, as our approach **explicitly preserves the pre-trained knowledge while learning the forgery patterns.** We also use PCA to compute the rank of the feature space, similar to Fig. 3. Results in Fig. 6 highlight the superiority of our method, where full-parameters fine-tuning and LoRA can lead to a notable reduction (-103 and -23, respectively) while our method best retains the pre-trained knowledge, maintaining a higher-ranked feature space for better generalization.

The naively trained models leverage very limited forgery patterns for discrimination. In our motivation, we argue that conventional training paradigms cause models to over-rely on very limited forgery patterns for discrimination, thereby causing the highly low-ranked and constrained feature space, limiting their expressivity and generalization (Fig. 2 and Fig. 3). To further validate this argument, we analyze the **discrimination behavior** of baseline models (*e.g.,* Xception and CLIP) through decision boundary visualization using the model's output logits of real and fake classes (Fig. 7). The linear alignment of predictions along $y = -x + b$ reveals that **Xception collapses real/fake**

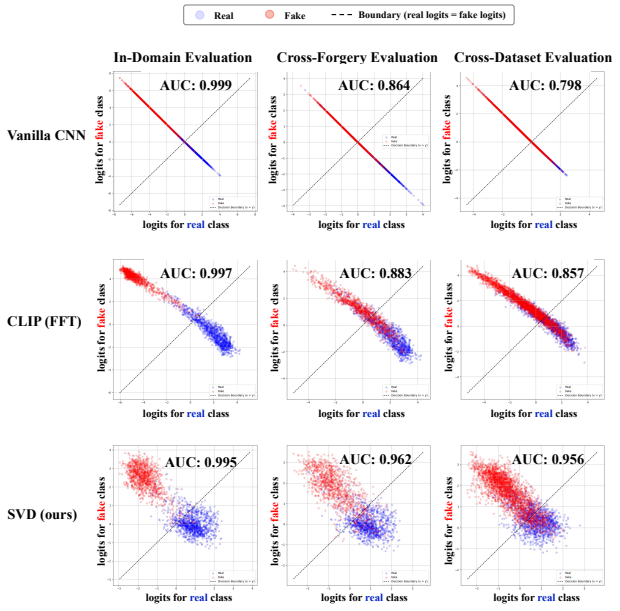

Figure 7. **Evidence for validating the discrimination dimension from the *logits space*. The first row** shows that the vanilla CNN (*i.e.,* Xception) overfits to the seen fakes and relies only on the forgery pattern for discrimination. **The second row** shows that fully fine-tuning the VFM (*i.e.,* CLIP) involves pre-trained knowledge but also distorting part of it during learning forgery patterns. **The third row** shows that our method uses SVD to achieve the orthogonality, thereby best retaining the pre-trained knowledge.

discrimination into a single discriminative dimension, confirming its exclusive dependence on forgery cues. Notably, while CLIP's original higher-ranked representations initially preserve informative structures, full fine-tuning catastrophically degrades this structure, forcing decisions into a similarly collapsed subspace ($y = -x + b$). Our method addresses this concern by learning forgery-related features in a novel orthogonal complement subspace relative to CLIP's original representational embedding space, thereby achieving optimal generalization performance.

How to determine the value of rank in tuning SVD? Our choice of using a lower rank (specifically, "n-r"=1) for fine-tuning DFD is primarily motivated by two critical factors: **First, the nature of the real-fake classification task itself makes it relatively straightforward**. Specifically, fake samples in existing training sets tend to exhibit a limited number of distinctive forgery patterns (FF++ contains only four forgery types), each with relatively simple and consistent characteristics. Due to this simplicity and limited diversity, a low-rank adaptation with a small rank (e.g., "n-r"=1, 4, or 16) is sufficient for the model to effectively learn these forgery patterns. As demonstrated by Table 5 in our paper, choosing ranks of 1, 4, or 16 yields very similar performance results. Given this observation, we prioritize efficiency and parameter economy, making rank 1 the optimal choice. **Second, the inherent characteristics of binary classification further justify selecting a smaller rank.** Binary classification tasks typically do not require the model to learn extensive and nuanced patterns, but rather to identify just enough distinctive features to separate the two classes, making the learned feature space inherently constrained. Thus, binary classification inherently simplifies the complexity of the learning problem, meaning that employing a higher rank would not provide significant additional benefit.

5. Conclusion

In this paper, we start our research from a new perspective to excavate the failure reason of the generalization in AIGI detection, namely the asymmetry phenomena, where a naively trained detector very quickly shortcuts to the seen fake patterns, collapsing the feature space into a low-ranked structure that limits expressivity and generalization. To address this, we propose integrating higher-ranked pre-trained knowledge from vision foundation models to expand the feature space. Simultaneously, we decompose the feature space into two orthogonal subspaces, for preserving pre-trained knowledge while learning forgery. Beyond LoRA and full-parameters tuning, we explicitly ensure the orthogonality, maintaining the higher rank of the whole feature space for better generalization. Furthermore, we reveal a very important prior for generalizable AIGI detection that fake data actually originates from real data in a hierarchical structure, not independently. Our method leverages this prior by preserving pre-trained semantic components while adapting to fake detection, enabling discrimination in semantic-aligned subspaces with reduced model complexity and improved generalization. Extensive experiments with deep analysis on both deepfake and synthetic image detection benchmarks have demonstrated the superior advantages of the proposed method in AIGI detection.

Impact Statement

This research advances application-driven machine learning by proposing a novel method for detecting AI-generated images. By effectively identifying deepfakes and curbing malicious uses of generative models, it holds significant potential for positive social impact. However, a possible negative outcome is the misuse of our method to enhance the realism of deepfake generators. To address this concern, we plan to implement measures like access control. Overall, we urge the research community to minimize negative impacts while leveraging the positive contributions of this work.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China (No. 62202014, 62332002, 62425101).

References

Wukong, 2022. 5. Inhttps://xihe.mindspore.cn/modelzoo/wukong, 2022. 5.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018a.

Brock, A. et al. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018b.

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., and Yang, X. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pp. 4113–4122, 2022.

Chai, L., Bau, D., Lim, S.-N., and Isola, P. What makes fake images detectable? understanding properties that generalize. In *ECCV*, pp. 103–120. Springer, 2020.

Chen, B., Zeng, J., Yang, J., and Yang, R. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, 2024.

Chen, C., Chen, Q., Xu, J., and Koltun, V. Learning to see in the dark. In *CVPR*, 2018.

Chen, L., Zhang, Y., Song, Y., Liu, L., and Wang, J. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *CVPR*, pp. 18710–18719, 2022a.

Chen, L., Zhang, Y., Song, Y., Wang, J., and Liu, L. Ost: Improving generalization of deepfake detection via one-shot test-time training. In *NeurIPS*, 2022b.

Chen, Q. and Koltun, V. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017.

Cheng, J., Yan, Z., Zhang, Y., Luo, Y., Wang, Z., and Li, C. Can we leave deepfake data behind in training deepfake detector? *NeurIPS*, 2024.

Choi, J., Kim, T., Jeong, Y., Baek, S., and Choi, J. Exploiting style latent flows for generalizing deepfake video detection. In *CVPR*, pp. 1133–1143, 2024.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

Dai, T., Cai, J., Zhang, Y., Xia, S.-T., and Lei, Z. Second-order attention network for single image super-resolution. In *CVPR*, 2019.

detection challenge., D., 2020. https://www.kaggle.com/c/deepfake-detection-challenge Accessed 2021-04-24.

DFD., 2020. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html Accessed 2021-04-24.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Dhariwal, P. et al. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., and Ge, Z. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *CVPR*, pp. 3994–4004, 2023.

Fu, X., Yan, Z., Yao, T., Chen, S., and Li, X. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *AAAI*, 2025.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pp. 10696–10706, 2022a.

Gu, Z., Yao, T., Chen, Y., Ding, S., and Ma, L. Hierarchical contrastive inconsistency learning for deepfake video detection. In *ECCV*, pp. 596–613. Springer, 2022b.

Guan, J., Zhou, H., Hong, Z., Ding, E., Wang, J., Quan, C., and Zhao, Y. Delving into sequential patches for deepfake detection. *NeurIPS*, 35:4517–4530, 2022.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. *NeurIPS*, 30, 2017.

Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, 2021.

Haliassos, A., Mira, R., Petridis, S., and Pantic, M. Leveraging real talking faces via self-supervision for robust forgery detection. In *CVPR*, pp. 14950–14962, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., and Ye, D. Implicit identity driven deepfake face swapping detection. In *CVPR*, pp. 4490–4499, 2023.

Jeong, Y. et al. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *WACV*, pp. 48–57, 2022.

Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Khalid, H. and Woo, S. S. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *CVPRW*, pp. 656–657, 2020.

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

Kong, C., Li, H., and Wang, S. Enhancing general face forgery detection via vision transformer with low-rank adaptation. In *ICCV*, pp. 102–107. IEEE, 2023.

Kong, C., Luo, A., Bao, P., Yu, Y., Li, H., Zheng, Z., Wang, S., and Kot, A. C. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*, 2024.

Korshunov, P. and Marcel, S. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

Larue, N., Vu, N.-S., Struc, V., Peer, P., and Christophides, V. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *ICCV*, pp. 21011–21021, 2023.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32, 2018.

Li, J., Xie, H., Li, J., Wang, Z., and Zhang, Y. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, 2021.

Li, K., Zhang, T., and Malik, J. Diverse image synthesis from semantic layouts via conditional imle. In *ICCV*, 2019.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B. Face x-ray for more general face forgery detection. In *CVPR*, 2020a.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, 2020b.

Lin, Y., Song, W., Li, B., Li, Y., Ni, J., Chen, H., and Li, Q. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. *arXiv preprint arXiv:2409.14444*, 2024.

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., and Yu, N. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *CVPR*, 2021a.

Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., and Zhao, Y. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pp. 10770–10780, 2024.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pp. 10012–10022, 2021b.

Liu, Z. et al. Global texture enhancement for fake face detection in the wild. In *CVPR*, pp. 8060–8069, 2020.

Luo, A., Cai, R., Kong, C., Kang, X., Huang, J., and Kot, A. C. Forgery-aware adaptive vision transformer for face forgery detection. *arXiv preprint arXiv:2309.11092*, 2023a.

Luo, A., Kong, C., Huang, J., Hu, Y., Kang, X., and Kot, A. C. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE TIFS*, 19:1168–1182, 2023b.

Luo, Y., Zhang, Y., Yan, J., and Liu, W. Generalizing face forgery detection with high-frequency features. In *CVPR*, 2021.

Luo, Y., Du, J., Yan, K., and Ding, S. Lare$\hat{2}$: Latent reconstruction error based method for diffusion-generated image detection. In *CVPR*, pp. 17006–17015, 2024.

Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., and Yu, N. F 2 trans: High-frequency fine-grained transformer for face forgery detection. *IEEE TIFS*, 18:1039–1051, 2023.

MidJourney. https://www.midjourney.com/home.

Mohri, M. and Rostamizadeh, A. Rademacher complexity bounds for non-iid processes. *Advances in neural information processing systems*, 21, 2008.

Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., and Manjunath, B. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019.

Ni, Y., Meng, D., Yu, C., Quan, C., Ren, D., and Zhao, Y. Core: Consistent representation learning for face forgery detection. In *CVPRW*, pp. 12–21, 2022.

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

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022.

Ojha, U. et al. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pp. 24480–24489, 2023.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pp. 86–103. Springer, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *ICML*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022a.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022b.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pp. 1–11, 2019.

Shao, R., Wu, T., Nie, L., and Liu, Z. Deepfake-adapter: Dual-level adapter for deepfake detection. *arXiv preprint arXiv:2306.00863*, 2023.

Shiohara, K. and Yamasaki, T. Detecting deepfakes with self-blended images. In *CVPR*, pp. 18720–18729, 2022.

Shiohara, K., Yang, X., and Taketomi, T. Blendface: Redesigning identity encoders for face-swapping. In *ICCV*, pp. 7634–7644, 2023.

Sun, K., Liu, H., Ye, Q., Gao, Y., Liu, J., Shao, L., and Ji, R. Domain general face forgery detection by learning to weight. In *AAAI*, volume 35, pp. 2638–2646, 2021.

Sun, K., Yao, T., Chen, S., Ding, S., Li, J., and Ji, R. Dual contrastive learning for general face forgery detection. In *AAAI*, volume 36, pp. 2316–2324, 2022.

Tan, C., Zhao, Y., Wei, S., Gu, G., and Wei, Y. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR*, pp. 12105–12114, June 2023.

Tan, C., Liu, P., Tao, R., Liu, H., Zhao, Y., Wu, B., and Wei, Y. Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection. *arXiv preprint arXiv:2403.06803*, 2024a.

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., and Wei, Y. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI*, volume 38, pp. 5052–5060, 2024b.

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., and Wei, Y. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, pp. 28130–28139, 2024c.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pp. 6105–6114. PMLR, 2019.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *ICML*, pp. 10347–10357. PMLR, 2021.

Trinh, L. and Liu, Y. An examination of fairness of ai models for deepfake detection. *arXiv*, 2021.

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., and Li, S.-N. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *ICMR*, pp. 615–623, 2022.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020a.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pp. 8695–8704, 2020b.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection. In *ICCV*, pp. 22445–22455, 2023a.

Wang, Z., Bao, J., Zhou, W., Wang, W., and Li, H. Altfreezing for more general video face forgery detection. In *CVPR*, pp. 4129–4138, 2023b.

Wu, H., Zhou, J., and Zhang, S. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023.

Xu, Y., Liang, J., Jia, G., Yang, Z., Zhang, Y., and He, R. Tall: Thumbnail layout for deepfake video detection. In *ICCV*, pp. 22658–22668, 2023.

Yan, Z., Zhang, Y., Fan, Y., and Wu, B. Ucf: Uncovering common features for generalizable deepfake detection. *arXiv preprint arXiv:2304.13949*, 2023a.

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., and Wu, B. Deepfakebench: A comprehensive benchmark of deepfake detection. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *NeurIPS*, volume 36, pp. 4534–4565. Curran Associates, Inc., 2023b.

Yan, Z., Luo, Y., Lyu, S., Liu, Q., and Wu, B. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, pp. 8984–8994, 2024a.

Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Yuan, L., Wang, C., Ding, S., et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024b.

Yan, Z., Ye, J., Li, W., Huang, Z., Yuan, S., He, X., Lin, K., He, J., He, C., and Yuan, L. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023.

Zhang, D., Xiao, Z., Li, S., Lin, F., Li, J., and Ge, S. Learning natural consistency representation for face forgery video detection. *arXiv preprint arXiv:2407.10550*, 2024.

Zhang, X., Karaman, S., and Chang, S.-F. Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515*, 2019.

Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., and Xia, W. Learning self-consistency for deepfake detection. In *ICCV*, 2021.

Zheng, Y., Bao, J., Chen, D., Zeng, M., and Wen, F. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, pp. 15044–15054, 2021.

Zhou, T., Wang, W., Liang, Z., and Shen, J. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5778–5788, 2021.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

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

Zhu, X., Wang, H., Fei, H., Lei, Z., and Li, S. Z. Face forgery detection by 3d decomposition. In *CVPR*, 2021.

Zi, B., Chang, M., Chen, J., Ma, X., and Jiang, Y.-G. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pp. 2382–2390, 2020.

A. Additional Results and Ablations

In this section, we provide additional experimental results and ablation studies of our proposed approach.

A.1. Results on GenImage Benchmark

In our manuscript, we present the benchmarking outcomes of the UniversalFakeDetect Dataset. Additionally, we report the results obtained from another widely utilized benchmark known as GenImage (Zhu et al., 2024). This GenImage dataset predominantly utilizes the Diffusion model for image generation, incorporating models such as Midjourney (MidJourney.), SDv1.4 (Rombach et al., 2022b), SDv1.5 (Rombach et al., 2022b), ADM (Dhariwal et al., 2021), GLIDE (Nichol et al., 2021), Wukong (Wuk, 2022. 5), VQDM (Gu et al., 2022a), and BigGAN (Brock et al., 2018b). Following the settings defined for GenImage, we designate SDv1.4 as the training set and the remaining models as the test set. Given the diverse image sizes within the GenImage dataset, images with a size smaller than 224 pixels are duplicated and subsequently cropped to 224 pixels, following (Tan et al., 2024c). We employ the same setting to re-implement FreqNet, FatFormer, and NPR, and also report the results of UnivFD and DRCT from (Chen et al., 2024).

The results on the GenImage dataset are presented in Table 9. When SDv1.4 is employed as the training set, our method attains an overall accuracy rate of 91.1% across the entire test set. Compared to similar methods that utilize CLIP as the backbone, such as UnivFD and FatFormer, our approach improves accuracy by 11.6% and 2.2%, respectively. Moreover, when contrasted with the latest state-of-the-art (SOTA) method DRCT (ICML 2024), the proposed method achieves a 1.6% enhancement in accuracy. This clearly indicates that our method demonstrates superior generalization capabilities and achieves SOTA performance on the GenImage benchmark.

A.2. Comparison with Existing Video Detectors

In the **manuscript**, we mainly compare our method with image detectors. Here, we provide an individual result to compare our approach with existing SOTA video detectors. Following (Zhang et al., 2024; Xu et al., 2023), we conduct evaluations on the widely-used CDF-v2 (Li et al., 2020b) and DFDC (detection challenge., 2020) using the video-level AUC metric. We have considered both the classical detectors such as LipForensics and the latest SOTA detectors such as NACO (ECCV'24) for a comprehensive comparison. Results in Tab. 10 demonstrate that our image-based approach achieves higher generalization performance in both CDF-v2 and DFDC, improving 4.5% and 3.9% points than the second-best video-based models. This further validates the superior generalization performance of our approach.

A.3. Robustness Evaluation

To evaluate our model's robustness to random perturbations, we adopt the methodology used in previous studies (Haliassos et al., 2021; Zheng et al., 2021), which involves examining three distinct types of degradation: Block-wise distortion, Change contrast, and JPEG compression. We apply each of these perturbations at five different levels to assess the model's robustness under varying degrees of distortion, following (Chen et al., 2022b; Yan et al., 2024a). The video-level AUC results for these unseen perturbations, using the model trained on FF++ (c23), are depicted in Fig. 8. Generally, our approach shows higher results than other methods, demonstrating the better robustness of our approach than other models.

B. Additional Cross-Method Evaluation

The capability of face forgery detectors to generalize to new manipulation methods is crucial in practical, real-world applications. In our manuscript, we present cross-method evaluations using the DF40 dataset (Yan et al., 2024b). Specifically, we train the models with four manipulation methods from FF++ (c23) and then test them on the other eight manipulation techniques provided in DF40. Furthermore, we conduct an additional cross-method evaluation following the protocol introduced in (Sun et al., 2022; Miao et al., 2023; Luo et al., 2023a). This protocol involves training the model on diverse manipulation types of samples and subsequently testing it on unknown manipulation methods. The results of this evaluation are reported in Tab. 11. It is evident that our proposed method attains remarkable performance in cross-manipulation evaluation. In terms of accuracy (ACC), it outperforms the latest SOTA detector FA-ViT by 2.85% on GID-DF and 3.14% on GID-F2F, respectively.

B.1. Additional Ablation Studies

Impact of Different Vision Foundation Models We initialize the ViT backbone with several widely used pre-trained weights from different vision foundation models, including BEIT-v2 (Peng et al., 2022), CLIP (Radford et al., 2021), and SigLIP (Zhai et al., 2023). The results are shown in Tab. 12 and Tab. 13. It is evident that our proposed approach improves the generalization performance of different pre-trained ViTs. On the other hand, we note that different initialization significantly impacts generalization performance, indicating the importance of choosing a suitable pre-trained initialization. Through empirical results, we discover that the ViT pre-trained on CLIP exhibits the highest performance in both deepfake detection and synthetic image detection tasks. Therefore, we choose CLIP as the default setting for our approach.

Table 9. **Benchmarking results of cross-method evaluations in terms of Acc performance on the Genimage dataset.** We follow (Zhu et al., 2024) and use the SDv1.4 as the training set while others as the testing sets. We directly cite the results of ResNet-50, DeiT-S, Swin-T, CNNSpot, Spec, F3Net, and GramNet from (Zhu et al., 2024). We obtain the results of UnivFD and DRCT from (Chen et al., 2024), and FreqNet, NPR, and FatFormer by using the official checkpoints for reproduction. We report the Accuracy metric for comparison following (Chen et al., 2024).

Methods	Venues	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	mAcc
ResNet-50 (He et al., 2016)	CVPR 2016	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S (Touvron et al., 2021)	ICML 2021	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T (Liu et al., 2021b)	ICCV 2021	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot (Wang et al., 2020b)	CVPR 2020	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec (Zhang et al., 2019)	WIFS 2019	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net (Qian et al., 2020)	ECCV 2020	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
GramNet (Liu et al., 2020)	CVPR 2020	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
UnivFD (Ojha et al., 2023)	CVPR 2023	91.5	96.4	96.1	58.1	73.4	94.5	67.8	57.7	79.5
NPR (Tan et al., 2024c)	CVPR 2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FreqNet (Tan et al., 2024b)	AAAI 2024	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.8
FatFormer (Liu et al., 2024)	CVPR 2024	92.7	100.0	99.9	75.9	88.0	99.9	98.8	55.8	88.9
DRCT (Chen et al., 2024)	ICML 2024	91.5	95.0	94.4	79.4	89.2	94.7	90.0	81.7	89.5
Ours	–	82.4	99.8	99.8	78.7	93.3	97.4	91.7	77.6	91.1

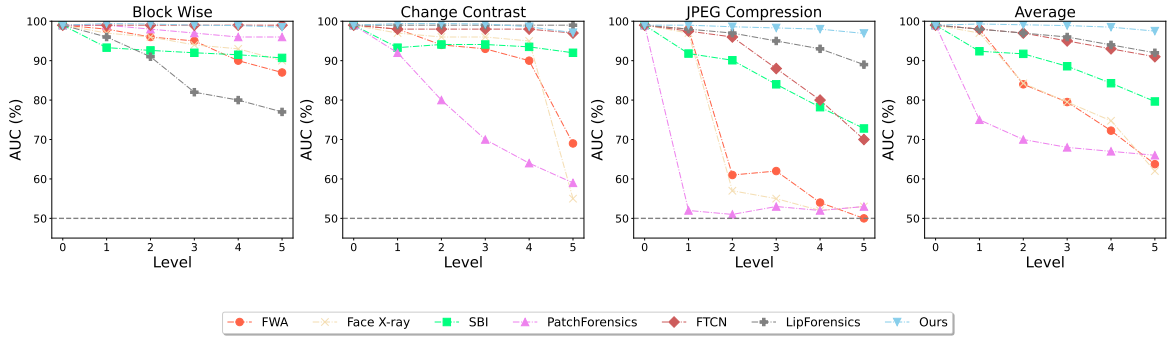

Figure 8. **Robustness to unseen perturbations.** We present video-level AUC for five distinct degradation levels across three types of perturbations in (Jiang et al., 2020).

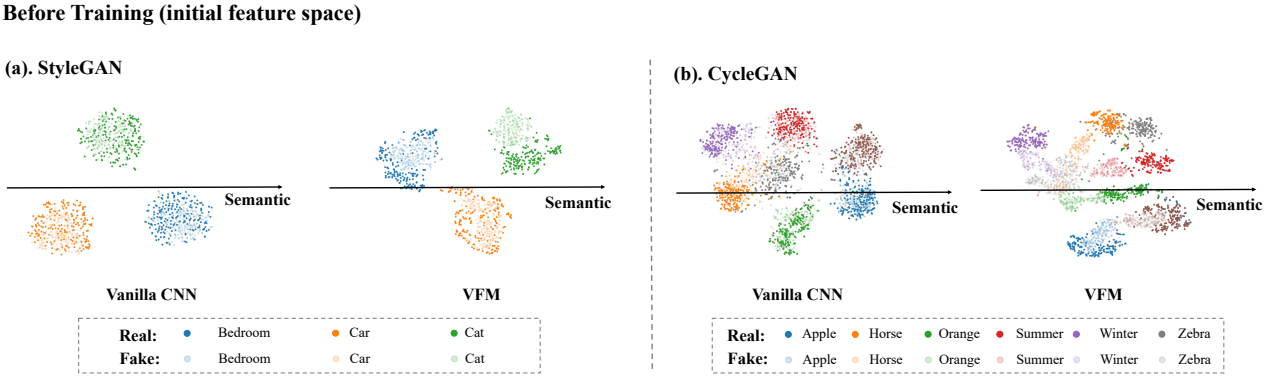

Figure 9. **t-SNE visualizations of the *initial* latent feature spaces between vanilla CNN (Wang et al., 2020b) and CLIP (Radford et al., 2021).** We show that both the pre-trained CNN and CLIP can identify different semantic objects.

Impact of Different ViT Backbones Here, we investigate the effects of different ViT architectures. Specifically, we consider two backbones that were implemented in the original paper of CLIP: ViT-Base-16 and ViT-Large-14. We conduct evaluations on both deepfake detection and synthetic image detection benchmarks, as shown in Tab. 14 and Tab. 15. Compared to fully fine-tuning the CLIP model, our proposed approach consistently demonstrates substantial enhancements in generalization performance across these ViT backbones. It is worth noting that CLIP-Large performs better than CLIP-Base by a notable margin. Based on this ablation experiment, we ultimately choose ViT-Large as our

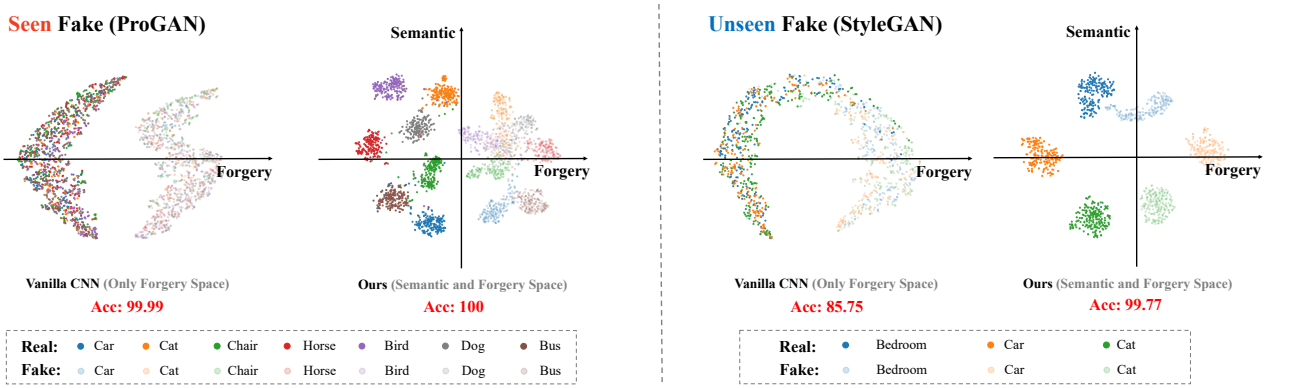

Figure 10. **t-SNE visualizations of the latent feature spaces between vanilla CNN (Wang et al., 2020b) and ours.** We use the testing set of ProGAN and StyleGAN within UniversalFakeDetect Dataset (Wang et al., 2020b) for visualization. We see that after fine-tuning on AIGI data, the baseline quickly shortcuts to the fake and "forgets" the pre-existing semantic knowledge, thereby resulting in a highly constrained feature space.

Table 10. **Cross-dataset generalization evaluations with existing SOTA video detectors**. The results of other detectors are directly cited from their original papers. The metric is video-level AUC.

Methods	Venues	CDF-v2	DFDC	Avg.
LipForensics (Haliassos et al., 2021)	CVPR 2021	0.824	0.735	0.780
FTCN (Zheng et al., 2021)	ICCV 2021	0.869	0.740	0.805
HCIL (Gu et al., 2022b)	ECCV 2022	0.790	0.692	0.741
RealForensics (Haliassos et al., 2022)	CVPR 2022	0.857	0.759	0.808
LTTD (Guan et al., 2022)	NeurIPS 2022	0.893	0.804	0.849
AltFreezing (Wang et al., 2023b)	CVPR 2023	0.895	–	–
TALL-Swin (Xu et al., 2023)	ICCV 2023	0.908	0.768	0.838
StyleDFD (Choi et al., 2024)	CVPR 2024	0.890	–	–
NACO (Zhang et al., 2024)	ECCV 2024	0.895	0.767	0.831
Ours	–	0.956	0.843	0.900

Table 11. **Additional results of cross-manipulation evaluation on FF++ (c23).** Following (Miao et al., 2023; Luo et al., 2023a), we conduct evaluations by training on the other three manipulated methods while testing on the remaining one. Specifically, GID-DF means training on the other three manipulated methods (FF-F2F, FF-FS, FF-NT) while testing on the FF-DF. Results of other methods are cited from (Miao et al., 2023; Luo et al., 2023a).

Methods	GID-DF		GID-F2F	
	Acc	AUC	Acc	AUC
EfficientNet (Tan & Le, 2019)	82.40	91.11	63.32	80.10
MLGD (Li et al., 2018)	84.21	91.82	63.46	77.10
LTW (Sun et al., 2021)	85.60	92.70	65.60	80.20
DCL (Sun et al., 2022)	87.70	94.90	68.40	82.93
M2TR (Wang et al., 2022)	81.07	94.91	55.71	76.99
F3Net (Qian et al., 2020)	83.57	94.95	61.07	81.20
F2Trans (Miao et al., 2023)	89.64	97.47	81.43	90.55
CFM (Luo et al., 2023b)	85.00	92.74	76.07	84.55
FA-ViT (Luo et al., 2023a)	92.86	98.10	82.57	91.20
Ours	95.71	99.26	85.71	93.83

Table 12. **Ablation studies on deepfake image detection (Cross-dataset) regarding different vision foundation models (VFMs) were used**. All models are trained on FF++ (c23) and tested on CDF-v2 and SimSwap.

VFMs	#Params	#ImgSize	CDF-v2	SimSwap	Avg.
BEIT-v2 (Peng et al., 2022)	303M	224	0.855	0.821	0.838
+ Ours	0.14M	224	0.894	0.850	0.872
SigLIP (Zhai et al., 2023)	316M	256	0.877	0.713	0.795
+ Ours	0.19M	256	0.895	0.778	0.867
CLIP (Radford et al., 2021)	307M	224	0.857	0.860	0.859
+ Ours	0.19M	224	0.956	0.926	0.941

Table 13. **Ablation studies on synthetic image detection (UniversalFakeDetect Dataset) regarding different vision foundation models (VFMs) were used**. All models are trained on ProGAN's images and tested on 19 different generative models' images.

VFMs	#Params	#ImgSize	mAP	mAcc
BEIT-v2 (Peng et al., 2022)	303M	224	93.50	79.11
+ Ours	0.14M	224	97.39	83.66
SigLIP (Zhai et al., 2023)	316M	256	94.30	81.23
+ Ours	0.19M	256	96.24	90.46
CLIP (Radford et al., 2021)	307M	224	97.95	86.22
+ Ours	0.19M	224	99.41	95.19

default backbone.

C. Additional Analysis and Visualizations

C.1. Additional t-SNE Visualizations

We further visualize the t-SNE of the seen fake ProGAN, unseen fake StyleGAN and useen CycleGAN for the comparison of vanilla CNN (Res-50 (Wang et al., 2020b)) and ours (see Fig. 9, Fig. 10). As we can see from both two figures, our approach maximizes and preserves the pre-trained knowledge while fitting the forgery patterns during training, whereas the vanilla CNN overfits the seen fake

Table 14. **Ablation studies on deepfake image detection (Cross-dataset) regarding different ViT architectures were used**. We employ the two architectures implemented in the original paper of CLIP (Radford et al., 2021) for experiments. All models are trained on FF++ (c23).

VFMs	#Params	#ImgSize	CDF-v2	SimSwap	Avg.
CLIP-Base/16	86M	224	0.854	0.833	0.844
+ Ours	0.07M	224	0.915	0.919	0.917
CLIP-Large/14	307M	224	0.857	0.860	0.859
+ Ours	0.19M	224	0.956	0.926	0.941

Table 15. **Ablation studies on synthetic image detection (UniversalFakeDetect Dataset) regarding different architectures were used**. All models are trained on ProGAN's images and tested on 19 different kinds of generative models' images.

VFMs	#Params	#ImgSize	mAP	mAcc
CLIP-Base/16	86M	224	96.25	82.52
+ Ours	0.07M	224	98.47	88.46
CLIP-Large/14	307M	224	97.95	86.22
+ Ours	0.19M	224	99.41	95.19

method, learning forgery patterns only, thereby resulting in a highly low-ranked feature space (see Fig. 6 and Fig. 7 of the manuscript for details) and causing the overfitting to seen forgery patterns in the training set. Additionally, we see that the logit distribution of the vanilla CNN has a larger overlapping between fake and real, while ours is highly smaller, suggesting that our approach achieves a better generalization performance.

C.2. Self-Attention Map Visualizations

Here, we perform the self-attention maps visualization of the original CLIP-ViT model (Original), the fully fine-tuned CLIP-ViT model (FFT), the LoRA-trained CLIP-ViT model (LoRA), and our proposed orthogonal trained CLIP-ViT model (Ours) on the UniversalFakeDetect dataset (see Fig. 11, Fig. 12, Fig. 13 and Fig. 14). Specifically, for each block of the ViT, the self-attention map denotes the self-attention coefficient matrix calculated between the [CLS] token and the patch tokens. In the case of the LoRA component, the self-attention maps are generated from left to right using the original + LoRA weights, the original weights, and the LoRA weights, respectively. In the case of the Ours component, the self-attention maps are generated from left to right using the principal + residual weights, the principal weights, and the residual weights, respectively. Surprisingly, we observe that the semantic information is primarily concentrated in the earlier blocks, and our proposed approach establishes orthogonality between the semantic subspace and the learned forgery subspace at the level of the self-attention map. It further explains that our proposed approach can better preserve the pre-trained knowledge while learning fake patterns.

D. Limitation and Future Work

The core idea of this paper is to decompose the original feature space into two orthogonal subspaces for preserving pre-trained knowledge while learning the forgery. In our manuscript and supplementary, we have conducted extensive experiments and in-depth analysis on both deepfake and synthetic image detection benchmarks, showing the superior advantages in both generalization and efficiency. One limitation of our work is that our approach regards all fake methods in one class during training real/fake classifiers, potentially ignoring the specificity and generality of different fake methods.

In the future, we plan to expand our approach into a *incremental learning* framework, where each fake method will be regarded as "one SVD branch", ensuring the orthogonality between different fake methods, thereby avoiding the severe forgetting of previous learned fake methods. This extension design will help our approach better address the future deepfake types in the real-world scenario. Additionally, although our work's scope mainly focuses on deepfake and synthetic image detection, our approach also has the potential to be applied to other similar fields such as face anti-spoofing, anomaly detection, etc. Furthermore, we hope our proposed approach can inspire future research in developing better orthogonal modeling strategies.

Ethics & Reproducibility. All of the facial images that are utilized are sourced from publicly available datasets and are accompanied by appropriate citations. This guarantees that there is no infringement upon personal privacy. We will make all codes and checkpoints available for public access upon acceptance.

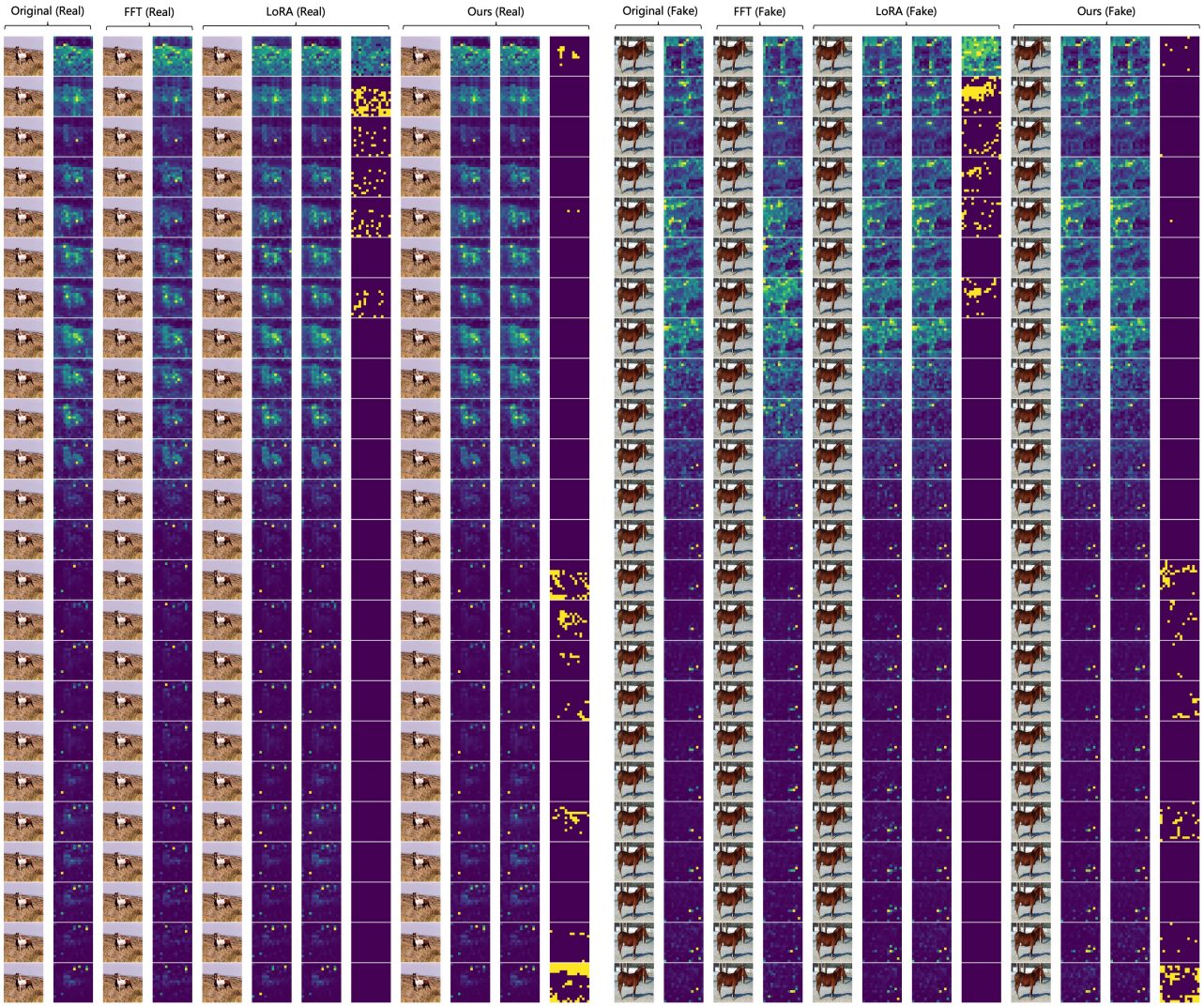

Figure 11. **Self-attention map visualizations of CycleGAN part in UniversalFakeDetect Dataset (Wang et al., 2020b).** We visualize the fake image of CycleGAN part and the corresponding real image for each block of the CLIP-ViT model (there are a total of 24 blocks, with IDs gradually increasing from top to bottom).

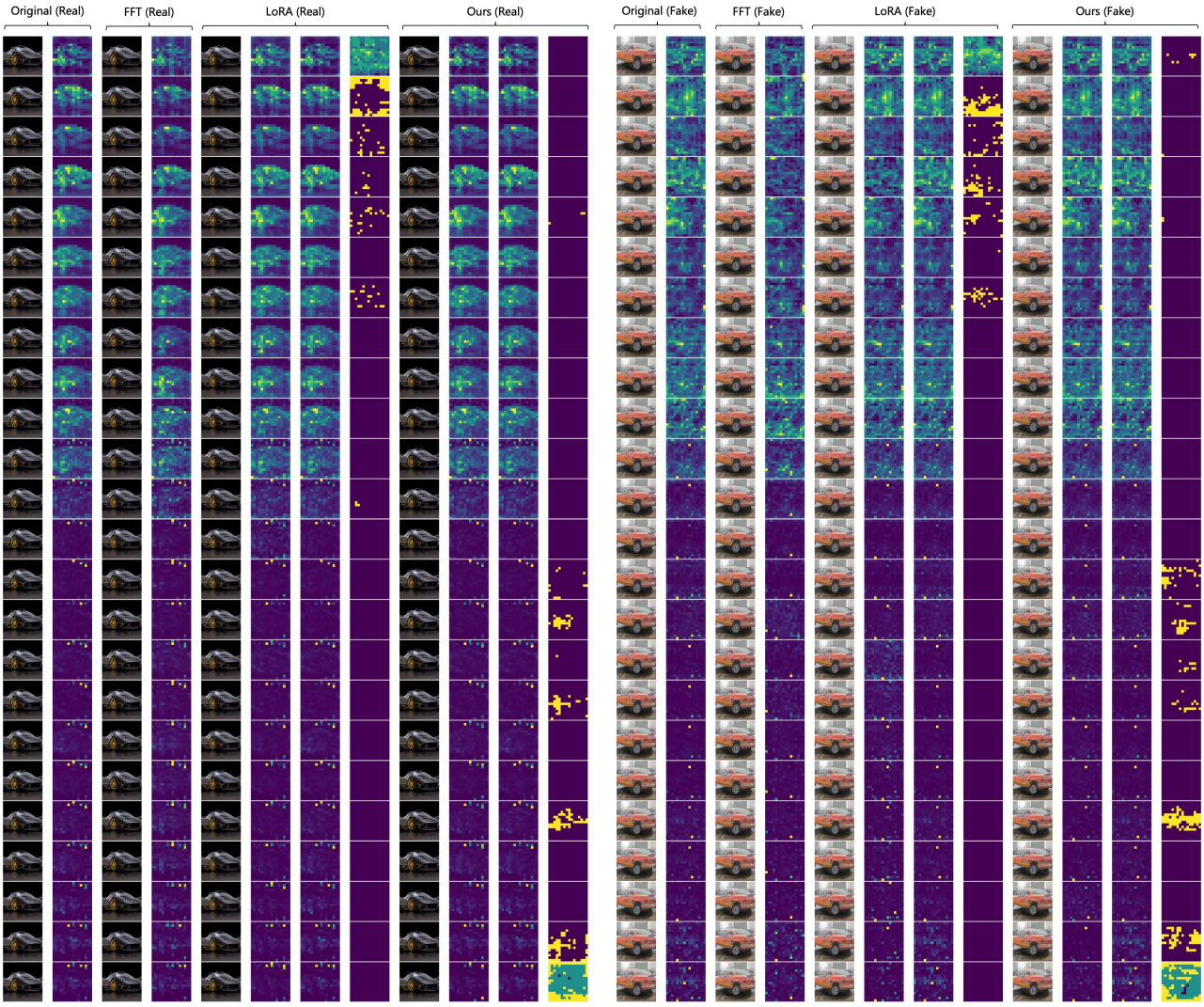

Figure 12. **Self-attention map visualizations of ProGAN part in UniversalFakeDetect Dataset (Wang et al., 2020b).** We visualize the fake image of ProGAN part and the corresponding real image for each block of the CLIP-ViT model (there are a total of 24 blocks, with IDs gradually increasing from top to bottom).

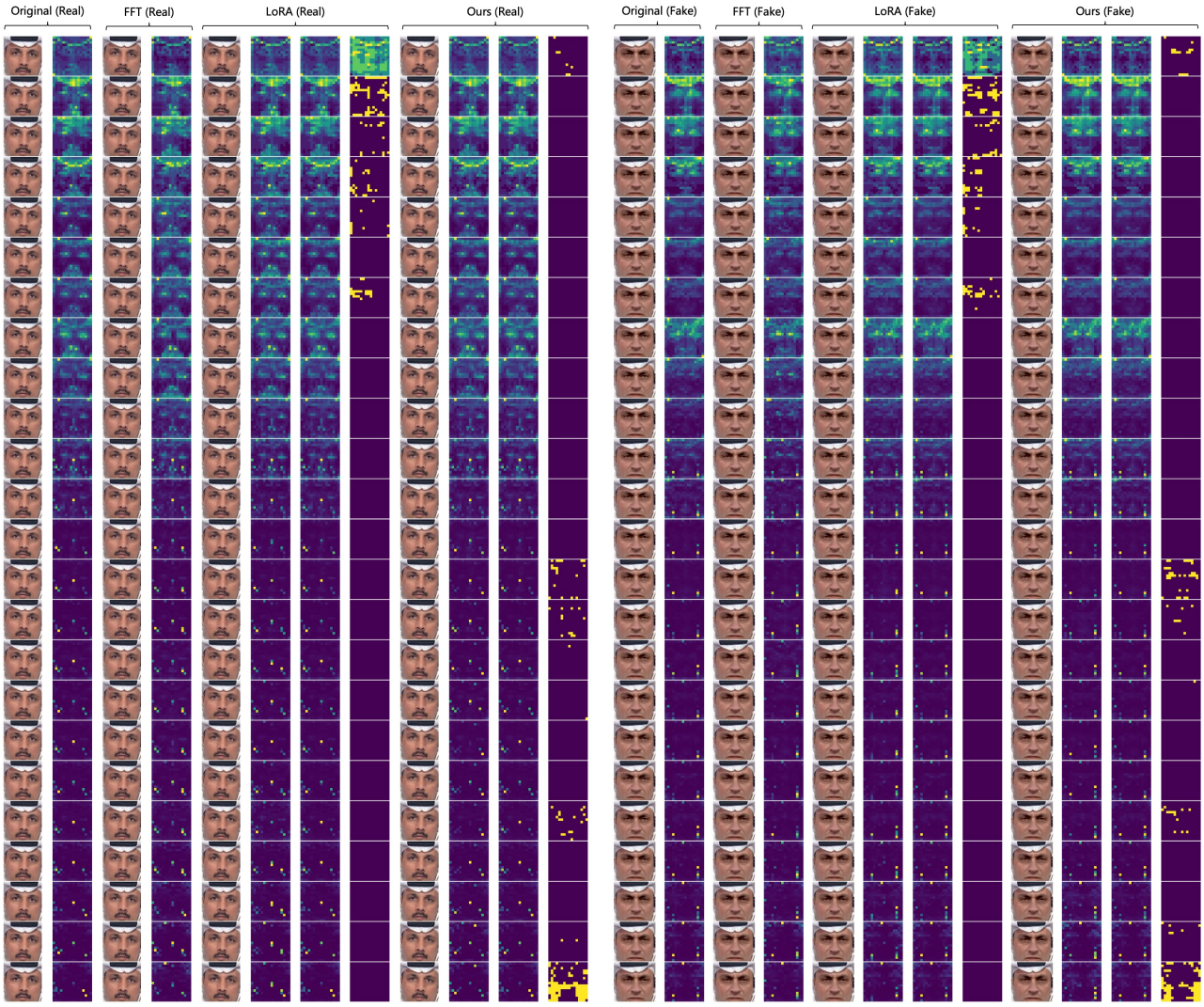

Figure 13. **Self-attention map visualizations of DeepFake part in UniversalFakeDetect Dataset (Wang et al., 2020b).** We visualize the fake image of DeepFake part and the corresponding real image for each block of the CLIP-ViT model (there are a total of 24 blocks, with IDs gradually increasing from top to bottom).

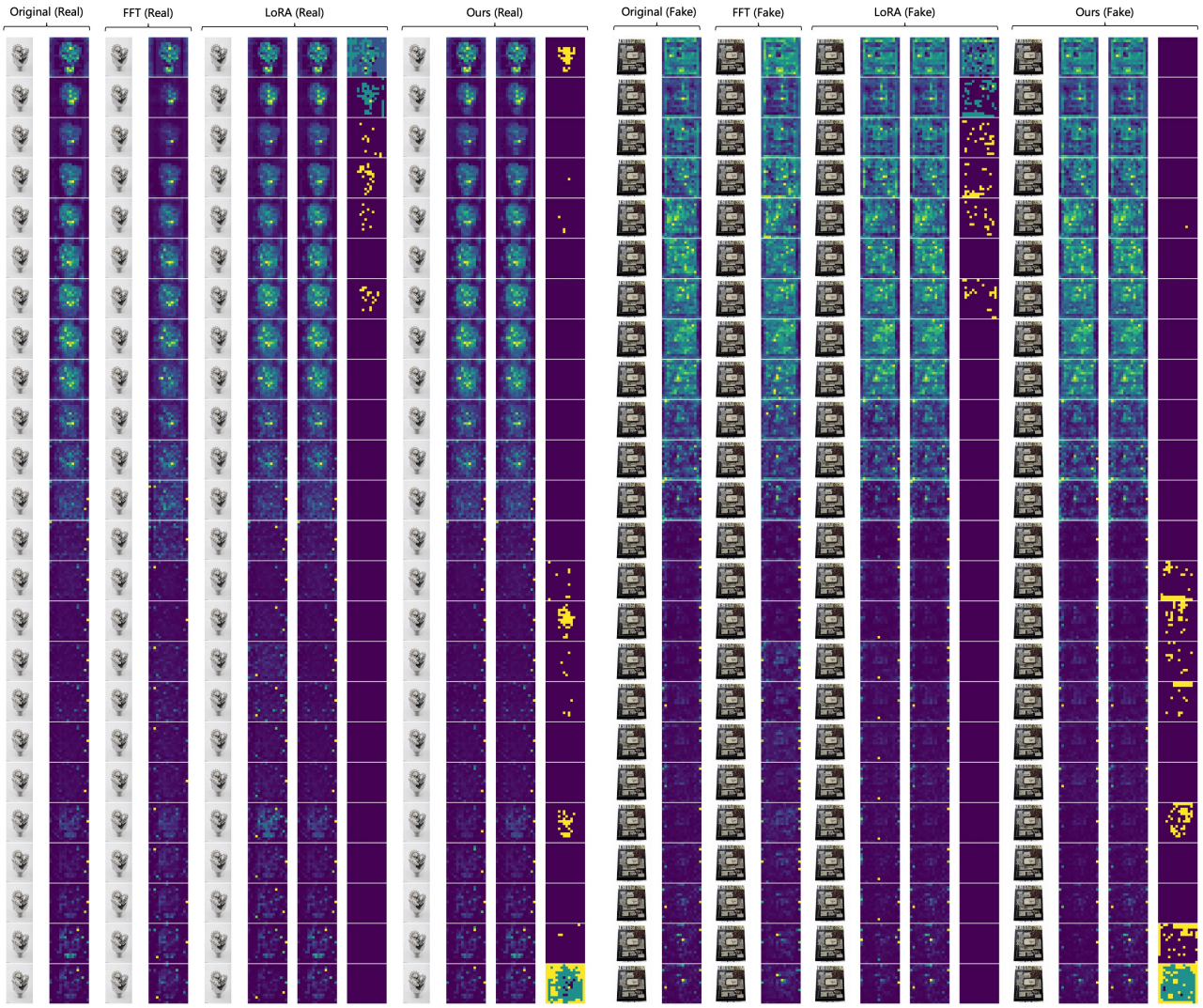

Figure 14. **Self-attention map visualizations of LDM part in UniversalFakeDetect Dataset (Wang et al., 2020b).**We visualize the fake image of LDM part and the corresponding real image for each block of the CLIP-ViT model (there are a total of 24 blocks, with IDs gradually increasing from top to bottom).