FreqDINO: Frequency-Aware Adaptation of Vision Transformers for AI-Generated Content Detection

Anonymous ICME submission

***Abstract*—AI-generated content (AIGC) detection has become increasingly critical as modern generative models rapidly evolve, producing synthetic images that closely resemble real photographs, challenging both human perception and current forensic systems. Traditional detection methods typically rely on handcrafted features or architectures tailored to specific generative mechanisms, which often suffer from poor generalization. While recent methods leveraging pre-trained foundation models have shown promising generalization, their lack of effective fine-tuning for downstream forensic tasks limits their detection capabilities. To overcome these limitations, we propose FreqDINO, a frequency-aware adaptation framework built upon the powerful DINOv3 vision foundation model. Unlike most existing methods that only use the foundation model for feature extraction, our approach introduces two modules for model fine-tuning to better adapt to forensic tasks: DWTAdapter, which refines intermediate representations through wavelet-based attention to capture multi-scale frequency cues, and WLoRA, a wavelet-driven low-rank adaptation mechanism that enables parameter-efficient and frequency-sensitive fine-tuning. Additionally, by integrating feature representations from various stages of DINO, our method allows the model to acquire more comprehensive features for AIGC detection. Extensive experiments across diverse datasets demonstrate that FreqDINO achieves state-of-the-art performance in AIGC detection. Moreover, thorough ablation studies further validate the effectiveness and complementary roles of the proposed modules.**

***Index Terms*—AI-Generated Content Detection, Image Forensics, DINOv3**

I. Introduction

The rapid evolution of modern generative models has dramatically increased the realism and diversity of AI-generated content (AIGC), enabling synthetic images to closely resemble real photographs and thereby challenging both human perception and conventional forensic techniques. As shown in Fig. 1, the visual gap between real and generated imagery has narrowed to the point where manual inspection becomes unreliable. This growing indistinguishability makes AIGC detection an increasingly critical research problem.

To address these issues, a variety of methods have been developed to capture spatial (e.g., [1]–[5]) and frequency (e.g., [6]–[8]) artifacts introduced by AI-generated images. For example, TruFor [1] leverages a Transformer-based fusion architecture to extract both high-level and low-level traces, combining RGB images with learned noise-sensitive fingerprints to detect anomalies within the image. NPR [2] analyzes up-sampling modules in GANs and diffusion models to derive handcrafted representations of local pixel dependencies. PSCC-Net [3] extracts multi-scale features in a top-down manner by downsampling in the first half of the network. In the second half, it combines the multi-scale features through a bottom-up path and utilizes a spatio-channel correlation module to generate the prediction results. CAT-Net [6] exploits statistics preserved in JPEG DCT coefficients together with RGB cues, using a dedicated network architecture to learn compression-related patterns and fuse them with visual features for manipulation detection. FreqNet [7] focuses on high-frequency information in the frequency domain, integrating frequency features into a CNN-based classifier to improve generalization across different generators and manipulation types. Despite recent progress, many of these approaches still rely on features or architectural components closely linked to specific generative processes, such as artifacts produced by particular up-sampling modules in GAN or diffusion. Since different generative mechanisms may introduce distinct artifacts in the spatial or frequency domains of images, the generalization ability of these methods is often limited.

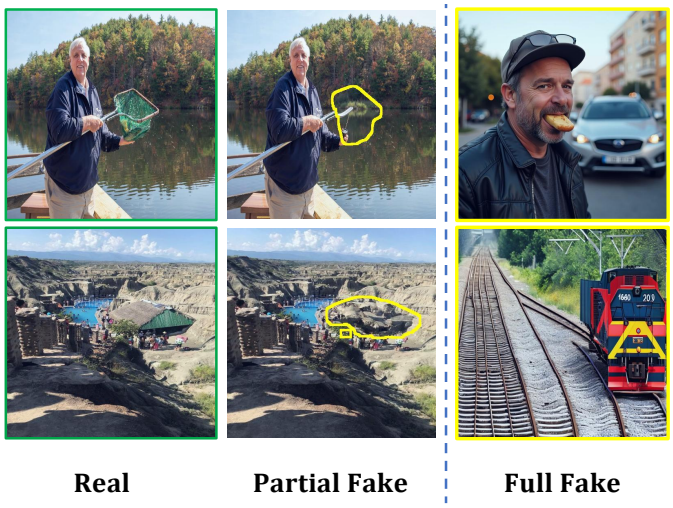

Fig. 1. Examples from the OpenSDID [9] dataset. The first column shows real images, the second column highlights manipulated regions in yellow, and the third column presents fully synthesized images.

Recent studies have shown that applying pretrained foundation models to downstream forensic tasks, such as deep fake detection [10]–[12] and AIGC detection [9], [13], [14], can achieve impressive performance. This is due to the foundation model's ability to learn robust features during pretraining, which can then be adapted to identify subtle patterns and artifacts associated with image manipulation or generation. For instance, CLIP (Contrastive Language-Image Pretraining) [15] associates images with textual descriptions using a con-

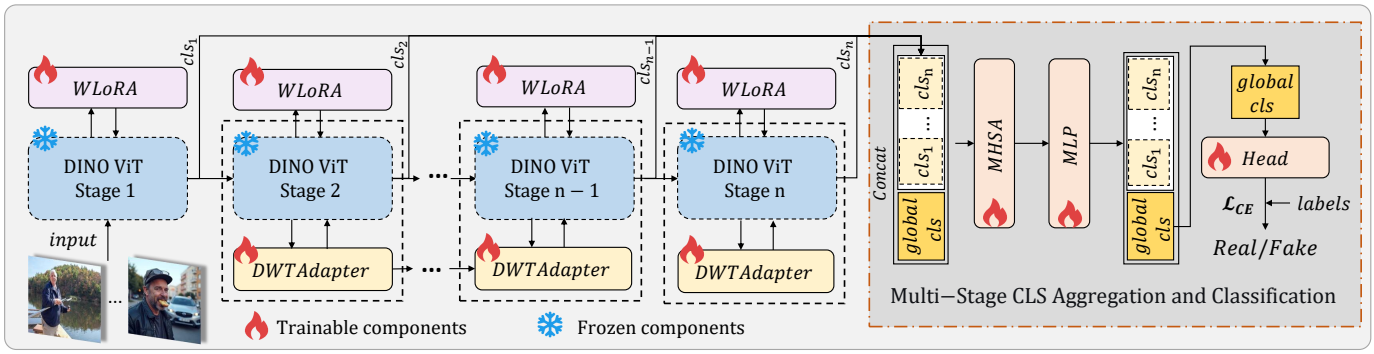

Fig. 2. The framework of the proposed model. The CLS tokens extracted from each ViT stage are concatenated with a global CLS token, followed by a multi-head self-attention layer, an MLP layer, and a classification head. The final global CLS token is used as the output representation for classification.

trastive learning framework. CLIP excels at zero-shot learning, enabling tasks like image classification and retrieval without task-specific training. Several CLIP-based AIGC detection methods have emerged to date. For instance, UniFD [13] and RINE [14] utilize frozen CLIP encoders while introducing lightweight modules, such as nearest-neighbor matching, linear classifiers, and projection and weighting modules, to generate representations for synthetic image detection. MaskCLIP [9] leverages a frozen CLIP encoder and learns a set of continuous prompt vectors for distinguishing real/fake concepts through prompt tuning. It then performs fake image detection by calculating the cosine similarity between image-text pairs. The methods mentioned above primarily use foundation models for feature extraction without effective fine-tuning. Since these foundation models are not specifically designed for forensic tasks, their representations often include irrelevant information, making it more challenging for classifiers to accurately distinguish between real and fake images.

To address this challenge, we propose FreqDINO, a method based on fine-tuned DINOv3 [16] for AI-generated content detection. Unlike previous frameworks that rely on the CLIP model, we adopt DINOv3 as the backbone network, which has shown superior performance in several image-related tasks [16]. To better adapt DINOv3 for forensic tasks, we introduce two key components for effective fine-tuning. The first component, DWTAdapter, is a plug-in module that uses Discrete Wavelet Transform (DWT) to decompose feature maps into subbands. It includes a Frequency Squeeze-and-Excitation Attention (FreqSEAttn) mechanism that adaptively adjusts the frequency bands to better capture artifacts generated by AIGC. The second component, WLoRA, is a wavelet-based low-rank adaptation module that applies DWT before low-rank projection. It efficiently adjusts a small number of parameters to focus on frequency-specific features that the pre-trained foundation model may have overlooked, enhancing the model's ability to detect spectral signals related to tampering. Furthermore, by incorporating intermediate representations from various stages of DINOv3, we further enhance the detection performance of the proposed model. Overall, our principal contributions are threefold:

- We introduce DWTAdapter, a plug-in module that injects frequency-domain information into the backbone network. This enables the model to capture subtle, visually elusive manipulation cues that are often missed in spatial or semantic features.
- We propose WLoRA, a wavelet-based parameter-efficient tuning mechanism. It enhances the model's adaptability to manipulation-relevant spectral patterns, which are typically overlooked by fixed, pre-trained representations.
- Experimental results demonstrate the effectiveness of the proposed FreqDINO through experiments on the OpenSDID [9] dataset. Extensive ablation studies further highlight the essential roles of both components, supporting the design principle of a deeply integrated, frequency-aware modeling framework.

II. Methodology

The network structure of the proposed FreqDINO is shown in Fig. 2. The method utilizes DINOv3 as the backbone. To address the challenges in AIGC detection, we introduce two key components for model fine-tuning: 1) a DWTAdapter component for intra-backbone feature refinement through frequency-aware attention; 2) a Wavelet-based Low-Rank Adaptation component for parameter-efficient fine-tuning; and 3) a component for integrating feature representations from various stages of DINO for final classification. In the following, we will detail these three components.

A. DWTAdapter

The DWTAdapter is a plug-in module designed to refine feature maps within the backbone by incorporating frequency-domain awareness. As shown in Fig. 3, it operates through three sequential steps:

Frequency Decomposition: Given an input feature $X \in \mathbb{R}^{L_1 \times E_1}$, where L_1 representing the token sequence length and E_1 representing the feature dimension, we first apply a linear transformation to obtain X'. Subsequently, we reshape it and apply a 2D DWT using a selected wavelet kernel (e.g., Haar, Symlets). This non-learnable transformation decomposes X' into four subbands: the low-frequency approximation (LL) and

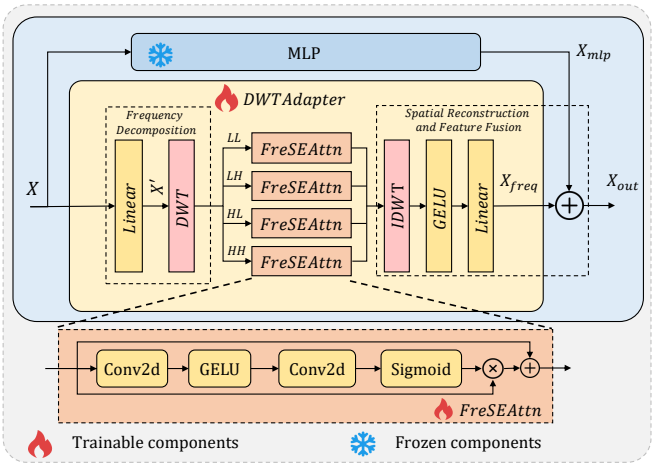

Fig. 3. The structure of the DWTAdapter and an illustration of its application in fine-tuning the MLP layers of the ViT within the DINO framework.

three high-frequency detail coefficients capturing horizontal (LH), vertical (HL), and diagonal (HH) orientations. This decomposition offers an explicit and structured representation of the feature's spectral content, providing the model with direct access to information that is often entangled and hidden in the spatial domain.

Frequency-SE Attention (FreqSEAttn): The wavelet coefficients LL, LH, HL, and HH are processed separately by the FreqSEAttn mechanism, which dynamically prioritizes the most informative features within each subband. For each subband, a lightweight convolutional squeeze-and-excitation (SE) block computes attention scores. This block consists of a 3×3 convolution, followed by a GELU activation, another 3×3 convolution, and a sigmoid activation. The resulting attention map is then multiplied by the input subband and integrated through a residual connection. This creates a learnable spectral gating mechanism that enhances critical spectral components (e.g., high-frequency artifacts in counterfeit detection) while suppressing irrelevant semantic information.

Spatial Reconstruction and Feature Fusion: The attended frequency features are reconstructed into the spatial domain using an Inverse DWT (IDWT). This refined spatial output, X_{freq}, is then combined with the original MLP-path feature, X_{mlp}, through direct element-wise summation: $X_{\text{out}} = X_{\text{mlp}} + X_{\text{freq}}$. This design ensures training stability by preserving the original gradient flow while maintaining the foundational spatial semantics from the frozen backbone.

B. WLoRA

LoRA (Low-Rank Adaptation) [22] is an efficient fine-tuning method for foundation models, adjusting low-rank matrices to reduce computation and storage while maintaining performance. Based on our experiments (see Table II), directly applying LoRA with DINO for forensic tasks does not yield satisfactory results. Therefore, we propose WLoRA, which integrates Discrete Wavelet Transform with LoRA to enable

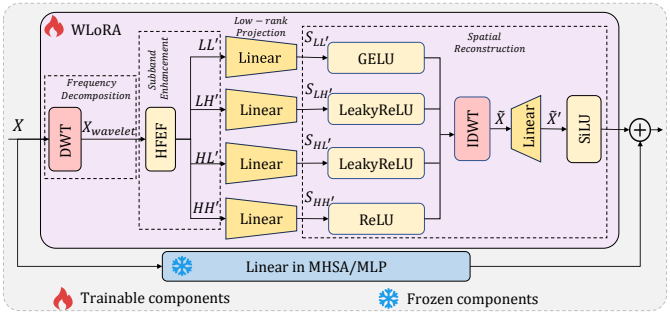

Fig. 4. The structure of the WLoRA and an illustration of its application in fine-tuning the linear layers within the DINO framework.

parameter-efficient fine-tuning. As shown in Fig. 4, it consists of four steps: frequency decomposition, subband enhancement, low-rank projection, and spatial reconstruction.

Frequency Decomposition: Given an input feature map $X \in \mathbb{R}^{L_2\times E_2}$ from the transformer layer, where L_2 representing the token sequence length and E_2 representing the feature dimension, we reshape it and apply a 2D DWT to decompose it into four subbands:

$$LL, LH, HL, HH = \mathcal{DWT}(X') \tag{1}$$

where LL subband represents the approximation (low-frequency) coefficients, while LH, HL, and HH correspond to the detail (high-frequency) coefficients along the horizontal, vertical, and diagonal directions, respectively. This decomposition enables the model to leverage information from different frequency branches.

Subband Enhancement: To adaptively modulate spectral components, a High-Frequency Enhancement Filter (HFEF) applies learnable scaling parameters λ_i to each subband:

$$\begin{cases} LL' = \lambda_1 \cdot LL, \\ LH' = \lambda_2 \cdot LH, \\ HL' = \lambda_3 \cdot HL, \\ HH' = \lambda_4 \cdot HH. \end{cases} \tag{2}$$

This enables the model to amplify or attenuate specific frequency bands according to their discriminative relevance for the AIGC detection task.

Low-rank Projection: Following enhancement, each subband $S_i \in \mathbb{R}^{L_2\times D_2\times D_2}$, where $i \in \{LL', LH', HL', HH'\}$, is flattened and projected to a lower dimension using a trainable matrix $P_{\text{down}} \in \mathbb{R}^{d_{\text{high}}\times d_{\text{low}}}$, where $d_{\text{high}} = D_2 \times D_2$ and we introduce a parameter $ratio$ to control the dimension of the low-rank projection by $ratio = d_{\text{low}}/d_{\text{high}}$:

$$S_i' = flatten(S_i) \cdot P_{down} \tag{3}$$

which yields a compressed representation $S_i' \in \mathbb{R}^{L_2\times d_{\text{low}}}$. This low-rank projection reduces the number of parameters while adjusting the relevant frequency components. An ablation study on the parameter $ratio$ is presented in Table III.

TABLE I

DETECTION PERFORMANCE ON THE OPENSDID BENCHMARK. ALL MODELS ARE TRAINED ON THE SD1.5 TRAINING SET AND TESTED ON SD1.5 AND OTHER DATASETS. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY.

Type	Method	SD1.5		SD2.1		SDXL		SD3		Flux.1		Avg	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Non-Foundation Model Based	CNNDet [17]	0.8460	0.8504	0.7156	0.7594	0.5970	0.6872	0.5627	0.6708	0.3572	0.5757	0.6157	0.7087
	GramNet [18]	0.8051	0.8035	0.7401	0.7666	0.6528	0.7076	0.6435	0.7029	0.5200	0.6337	0.6723	0.7229
	MVSS-Net [19]	0.9347	0.9365	0.7927	0.8233	0.5985	0.7042	0.6280	0.7213	0.2759	0.5678	0.646	0.7506
	PSCC-Net [3]	0.9607	0.9614	0.7685	0.8094	0.5570	0.6881	0.5978	0.7089	0.5177	0.6704	0.6803	0.7676
	TruFor [1]	0.9012	**0.9773**	0.3593	0.5562	0.5804	0.6641	0.5973	0.6751	0.4912	0.6162	0.5859	0.6978
	IML-ViT [20]	0.9447	0.7573	0.697	0.6119	0.4098	0.4995	0.4469	0.5125	0.182	0.4362	0.5361	0.5635
	NPR [2]	0.7941	0.7928	0.8167	0.8184	0.7212	0.7428	0.7343	0.7547	0.6762	0.7136	0.7485	0.7645
	CAT-Net [6]	**0.9615**	0.9615	0.7932	0.8246	0.6476	0.7334	0.6526	0.7361	0.2266	0.5526	0.6563	0.7616
	ObjectFormer [8]	0.7172	0.7522	0.6679	0.7255	0.4919	0.6292	0.4832	0.6254	0.3792	0.5805	0.5479	0.6626
	FreqNet [7]	0.7588	0.777	0.6097	0.6837	0.5315	0.6402	0.535	0.6437	0.3847	0.5708	0.5639	0.6631
Foundation Model Based	UniFD [13]	0.7745	0.7760	0.8062	0.8192	0.7074	0.7483	0.7109	0.7517	0.6110	0.6906	0.7220	0.7572
	RINE [14]	0.9108	0.9098	0.8747	0.8812	0.7343	0.7876	0.7205	0.7678	0.5586	0.6702	0.7598	0.8033
	DeCLIP [21]	0.8068	0.7831	0.8402	0.8277	0.7069	0.7055	0.6993	0.684	0.5177	0.6561	0.7142	0.7313
	MaskCLIP [9]	0.9264	0.9272	0.8871	0.8945	0.7802	0.8122	0.7307	0.7801	0.5649	0.6850	0.7779	0.8198
	Proposed	0.9392	0.9380	**0.9590**	**0.9590**	**0.9366**	**0.9379**	**0.9225**	**0.9254**	**0.6994**	**0.7596**	**0.8913**	**0.9040**

Spatial Reconstruction: An IDWT then reconstructs a spatial feature map from the compressed subbands:

$$\tilde{X} = \mathcal{IDWT}(Activation(S_i')) \tag{4}$$

where $Activation$ refers to the activation layer and $\tilde{X} \in \mathbb{R}^{L_2 \times D_2' \times D_2'}$. Finally, a dimensionality increase projection via a learnable matrix $P_{\text{up}} \in \mathbb{R}^{d_{\text{low}} \times d_{\text{high}}}$ restores the feature for integration, following a activation layer $Activation$:

$$\tilde{X}' = Activation(flatten(\tilde{X}) \cdot P_{up}) \tag{5}$$

Please note that in the aforementioned process, only the low-rank matrices P_{down}, P_{up}, and the scaling parameters λ_i are tuned during training. This ensures parameter efficiency while making the adaptation process inherently sensitive to spectral forensic traces.

C. Multi-Stage CLS Aggregation and Classification

In ViT [23], the [CLS] token is employed to learn a holistic representation of the input sequence. Recent studies [14] have further demonstrated that incorporating intermediate representations from different network stages—ranging from fine-grained spatial details to high-level semantic features—can significantly enhance detection performance. Building on this insight, we propose concatenating the $[CLS_i]$ tokens from each ViT stage with a global $[CLS]$ token, where i denotes the stage index as shown in Fig. 2. Here, the global $[CLS_{\text{global}}]$ token serves as a summary of the entire input sequence, aggregating key features from all stages. This composite representation is expressed as:

$$C = [CLS_{\text{global}}; CLS_1; \ldots; CLS_n] \tag{6}$$

This aggregated sequence C is then passed through a multi-head self-attention layer and an MLP block, producing a transformed sequence. The resulting global token is then extracted and fed into the final classification head to produce the prediction. Our experimental results show that the proposed multi-stage CLS aggregation and classification approach improves our average F1-score by more than 4%.

III. EXPERIMENTS

A. Experiments Setting

We conducted our experiments on the OpenSDID [9] dataset, a comprehensive benchmark designed for detecting diffusion-generated images in open-world environments. OpenSDID simulates realistic human-like editing behaviors by incorporating a diverse set of image manipulations and generative models. It includes outputs from several advanced vision-language models (VLMs) such as LLaMA 3 Vision [24], LLaVA [25], InternVL 2 [26], and Qwen 2 VL [27], which generate varied textual prompts reflecting different user intents. The dataset features images generated by leading-edge diffusion models, including Stable Diffusion versions 1.5 [28], 2.1 [28], SDXL [29], SD 3 [30], and Flux.1 [31], showcasing the latest advancements in text-to-image generation. OpenSDID consists of 300,000 images of varying sizes, with an approximately equal distribution of real and synthetic images. Of these, 200,000 images (100,000 real from Megalith-10M [32] and 100,000 synthetic from SD1.5) are used for training, while the remaining 100,000 images are reserved for testing, including 20,000 images each from SD 1.5, SD 2.1, SDXL, SD 3, and Flux.1. Some image examples are shown in Fig. 1.

For training FreqDINO, we use the Adam optimizer and a learning rate of 2×10^{-5} for 10 epochs. The standard data augmentation pipeline from OpenSDI [9] is applied, which includes Gaussian blur, JPEG compression, random scaling, horizontal/vertical flipping, random brightness/contrast adjustments, random rotation, and random cropping to 512×512. We employ a DINOv3-ViTL/16 backbone pre-trained on LVD-1689M. The adapter dimension ratio to 1/4, and the symlets4 wavelet kernel is used. The evaluation metrics are F1-score and accuracy (ACC). All code associated with this paper will be made publicly available upon its acceptance.

B. Comparison with Related Works

The experimental results evaluated on SD 1.5 (consistent with the training set source) and the other four datasets (i.e.,

TABLE II
ABLATION STUDY ON THE IMPACT OF DIFFERENT CONFIGURATIONS FOR DWTADAPTER AND WLORA IN THE PROPOSED FREQDINO MODEL.

Fine-tuning Setup		SD1.5		SD2.1		SDXL		SD3		Flux.1		Avg	
DWTAdapter	WLoRA	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
×	×	0.8150	0.8136	0.8640	0.8689	0.8433	0.8524	0.8052	0.8217	**0.7212**	**0.7611**	0.8097	0.8235
×	w/LoRA	0.9438	0.9456	0.9504	0.9523	0.8906	0.9004	0.8165	0.8435	0.3894	0.6180	0.7981	0.8520
×	✓	0.9450	0.9430	0.9422	0.9425	0.9020	0.9062	0.8920	0.8975	0.5653	0.6811	0.8493	0.8741
✓	×	**0.9702**	**0.9698**	**0.9691**	**0.9695**	0.9210	0.9254	0.9027	0.9097	0.4570	0.6418	0.8440	0.8832
✓	w/LoRA	0.9319	0.9294	0.9530	0.9525	0.8929	0.8980	0.8892	0.8948	0.5910	0.6936	0.8516	0.8737
✓	✓	0.9392	0.9380	0.9590	0.9590	**0.9366**	**0.9379**	**0.9225**	**0.9254**	0.6994	0.7596	**0.8913**	**0.9040**

TABLE III
ABLATION STUDY ON THE IMPACT OF DIFFERENT RATIO SETTINGS IN THE LOW-RANK PROJECTION OF WLORA IN THE PROPOSED FREQDINO.

Ratio	SD1.5		SD2.1		SDXL		SD3		Flux.1		Avg	
Setting	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
1/2	0.9409	0.9391	0.9593	0.9587	0.9266	0.9279	**0.9301**	**0.9313**	0.6654	0.7367	0.8845	0.8987
1/4	0.9392	0.9380	0.9590	0.9590	**0.9366**	**0.9379**	0.9225	0.9254	**0.6994**	**0.7596**	**0.8913**	**0.9040**
1/6	**0.9570**	**0.9566**	**0.9688**	**0.9691**	0.9296	0.9329	0.8859	0.8958	0.4632	0.6438	0.8409	0.8796

TABLE IV
ABLATION STUDY ON THE IMPACT OF DIFFERENT WAVELET KERNEL IN BOTH DWTADAPTER AND WLORA IN THE PROPOSED FREQDINO MODEL.

Wavelet	SD1.5		SD2.1		SDXL		SD3		Flux.1		Avg	
Kernel	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
Haar	**0.9509**	**0.9505**	**0.9636**	**0.9638**	0.9296	0.9323	0.8855	0.8942	0.5020	0.6577	0.8463	0.8797
Db2	0.9253	0.9263	0.9509	0.9520	0.9068	0.9127	0.8422	0.8606	0.4982	0.6577	0.8247	0.8619
Coif4	0.9297	0.9280	0.9532	0.9526	**0.9373**	0.9374	0.9168	0.9186	0.6335	0.7156	0.8741	0.8904
Sym4	0.9392	0.9380	0.9590	0.9590	0.9366	**0.9379**	**0.9225**	**0.9254**	**0.6994**	**0.7596**	**0.8913**	**0.9040**

SD2.1, SDXL, SD3, and Flux.1) are shown in Table I. From the table, we can make the following three observations:

- For SD1.5, spatial-based methods (e.g., TruFor) and frequency domain methods (e.g., CAT-Net) achieve relatively higher performance compared to foundation-model-based methods. Our method represents the current state-of-the-art among foundation-model-based approaches, with both F1 and ACC scores exceeding 0.93.
- For the other four test datasets, the detection performance of most methods significantly decreases. For instance, on the SD3 dataset, the F1-score of CAT-Net drop from 0.9615 to 0.6526, while our method shows a performance decline of less than 2%. Overall, our method achieves state-of-the-art performance across all four cross-dataset evaluations. It is worth noting that for the Flux.1 dataset, the performance of all methods significantly drops, indicating that the statistical characteristics of the Flux.1 test set differ greatly from those of the training set (i.e., SD1.5), which increases the detection difficulty.
- On average, our method achieves the highest performance among all methods, with an F1-score of 0.8913 and accuracy of 0.9040, surpassing the second-best method, MaskCLip, by 11.34% and 8.42%, respectively. The improvement is significant.

C. Ablation Study

About DWTAdapter and WLoRA: In this experiment, we compare different configurations for DWTAdapter and WLoRA, including scenarios with or without these components, and replacing WLoRA with standard LoRA. The results in Table II show that, in most cases, the proposed model, which incorporates both DWTAdapter and WLoRA, achieves the best or second-best performance compared to the other configurations. On average, our model delivers the best performance across both F1 and ACC. Removing either the DWTAdapter or WLoRA leads to significant performance declines. For example, in terms of F1, both result in a drop of over 4.20%. Similarly, replacing WLoRA with standard LoRA also causes a noticeable decrease in performance, with F1-score dropping by 3.97%. These experiments highlight the effectiveness of the two components introduced in this paper.

About the Low-rank Projection: In this experiment, we compare settings with different parameter of $ratio$ values in the Low-rank Projection. The experimental results in Table III show that adjusting this $ratio$ significantly impacts performance. For instance, when the ratio is set to 1/4 (the proposed setting), the model achieves an average F1-score of 0.8913, the highest among all configurations. In contrast, using a ratio of 1/6 leads to a performance decrease, with the average F1-score dropping to 0.8409.

About Wevelet kernel: In this experiment, we compare settings with different wavelet kernels used in both DWTAdapter and WLoRA. The experimental results in Table IV show that the Haar wavelet kernel achieves the best performance on the SD1.5 and SD2.1 datasets. However, for the Flux.1 dataset, it performs poorly compared to other kernels, with an F1-score of around 0.5000. Among all kernels tested, the sym4 kernel consistently delivers the best or second-best results across all datasets, achieving the best overall

average performance. These results highlight the importance of selecting the appropriate wavelet kernel, as it significantly impacts the model's performance across different datasets.

IV. Conclusion

In this paper, we propose FreqDINO, a frequency-aware adaptation framework for AI-generated content detection, built upon the powerful DINOv3 model. By introducing the DWTAdapter and WLoRA modules, FreqDINO effectively fine-tunes the foundation model for forensic tasks, capturing subtle frequency-domain artifacts that are often missed by traditional methods. Our extensive experiments demonstrate that FreqDINO outperforms existing approaches in AIGC detection, offering a robust and generalizable solution. The proposed modules significantly enhance model performance, validating the importance of frequency-aware fine-tuning in modern detection systems.

However, a limitation of our work is that, unlike some existing methods that also offer image detection and localization (e.g., [9]), FreqDINO primarily focuses on improving detection accuracy. We view this as a crucial first step, as enhancing detection performance forms the foundation for subsequent localization tasks. In future work, we aim to extend our framework by incorporating segmentation models, enabling joint detection and localization capabilities, and further advancing the field of AIGC forensics.

References

[1] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, et al., "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20606–20615.

[2] Chuangchuang Tan, Huan Liu, Yao Zhao, et al., "Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28130–28139.

[3] Xiaohong Liu, Yaojie Liu, Jun Chen, et al., "Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7505–7517, 2022.

[4] Ziyi Xi, Wenmin Huang, Kangkang Wei, et al., "Ai-generated image detection using a cross-attention enhanced dual-stream network," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. 2023, pp. 1463–1470, IEEE.

[5] Yuanhang Huang, Weiqi Luo, Xiaochun Cao, et al., "A forensic framework with diverse data generation for generalizable forgery localization," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 9732–9745, 2025.

[6] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, et al., "Learning JPEG compression artifacts for image manipulation detection and localization," *International Journal of Computer Vision*, vol. 130, no. 8, pp. 1875–1895, 2022.

[7] Chuangchuang Tan, Yao Zhao, Shikui Wei, et al., "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *Association for the Advancement of Artificial Intelligence*, 2024, pp. 5052–5060.

[8] Junke Wang, Zuxuan Wu, Jingjing Chen, et al., "Objectformer for image manipulation detection and localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2354–2363.

[9] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong, "Opensdi: Spotting diffusion-generated images in the open world," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 4291–4301.

[10] Zhiyuan Wang, Yanxiang Chen, Yuanzhi Yao, et al., "Idcnet: Image decomposition and cross-view distillation for generalizable deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 8373–8386, 2025.

[11] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, et al., "Transcending forgery specificity with latent space augmentation for generalizable deepfake detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8984–8994.

[12] Xinjie Cui, Yuezun Li, Ao Luo, et al., "Forensics adapter: Adapting CLIP for generalizable face forgery detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 19207–19217.

[13] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee, "Towards universal fake image detectors that generalize across generative models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24480–24489.

[14] Christos Koutlis and Symeon Papadopoulos, "Leveraging representations from intermediate encoder-blocks for synthetic image detection," in *Computer Vision-European Conference on Computer Vision*, 2024, vol. 15130, pp. 394–411.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, vol. 139, pp. 8748–8763.

[16] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al., "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[17] Sheng-Yu Wang, Oliver Wang, Richard Zhang, et al., "Cnn-generated images are surprisingly easy to spot... for now," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8692–8701.

[18] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8057–8066.

[19] Xinru Chen, Chengbo Dong, Jiaqi Ji, et al., "Image manipulation detection by multi-view multi-scale supervision," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14165–14173.

[20] Xiaochen Ma, Bo Du, Xianggen Liu, et al., "Iml-vit: Image manipulation localization by vision transformer," *arXiv preprint arXiv:2307.14863*, 2023.

[21] Stefan Smeu, Elisabeta Oneata, and Dan Oneata, "Declip: Decoding CLIP representations for deepfake localization," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 149–159.

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, et al., "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations*, 2022.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[24] Meta AI, "Llama 3.2: Revolutionizing edge ai and vision with open, customizable models," https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, 2024.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, et al., "Visual instruction tuning," in *Advances in Neural Information Processing Systems 36*, 2023.

[26] Zhe Chen, Weiyun Wang, Hao Tian, et al., "How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites," *Science China Information Sciences*, vol. 67, no. 12, 2024.

[27] Peng Wang, Shuai Bai, Sinan Tan, et al., "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al., "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10674–10685.

[29] Dustin Podell, Zion English, Kyle Lacey, et al., "SDXL: improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[30] Patrick Esser, Sumith Kulal, Andreas Blattmann, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.

[31] Black Forest Labs, Stephen Batifol, Andreas Blattmann, et al., "FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space," *arXiv preprint arXiv:2506.15742*, 2025.

[32] Ollin Boer Bohan, "Megalith-10m: A dataset of public domain photographs," https://huggingface.co/datasets/madebyollin/megalith-10m, 2024.