LAMM-ViT: AI Face Detection via Layer-Aware Modulation of Region-Guided Attention

Jiangling Zhang[a], **Weijie Zhu**[a], **Jirui Huang**[a] **and Yaxiong Chen**[b,c,*]

[a]Wuhan University of Technology
[b]Sanya Science and Education Innovation Park, Wuhan University of Technology, Sanya 572000, China
[c]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China

Abstract. Detecting AI-synthetic faces presents a critical challenge: it is hard to capture consistent structural relationships between facial regions across diverse generation techniques. Current methods, which focus on specific artifacts rather than fundamental inconsistencies, often fail when confronted with novel generative models. To address this limitation, we introduce Layer-aware Mask Modulation Vision Transformer (LAMM-ViT), a Vision Transformer designed for robust facial forgery detection. This model integrates distinct Region-Guided Multi-Head Attention (RG-MHA) and Layer-aware Mask Modulation (LAMM) components within each layer. RG-MHA utilizes facial landmarks to create regional attention masks, guiding the model to scrutinize architectural inconsistencies across different facial areas. Crucially, the separate LAMM module dynamically generates layer-specific parameters, including mask weights and gating values, based on network context. These parameters then modulate the behavior of RG-MHA, enabling adaptive adjustment of regional focus across network depths. This architecture facilitates the capture of subtle, hierarchical forgery cues ubiquitous among diverse generation techniques, such as GANs and Diffusion Models. In cross-model generalization tests, LAMM-ViT demonstrates superior performance, achieving 94.09% mean ACC (a +5.45% improvement over SoTA) and 98.62% mean AP (a +3.09% improvement). These results demonstrate LAMM-ViT's exceptional ability to generalize and its potential for reliable deployment against evolving synthetic media threats.The code is available at https://github.com/WHUT-ZJL/LAMM-ViT.

1 Introduction

Recent advancements in generative models, particularly Generative Adversarial Networks (GANs) [9, 14] and Diffusion Models (DMs) [4, 24], have revolutionized the creation of synthetic facial images. These models now generate faces that are virtually indistinguishable from authentic photographs, achieving unprecedented levels of photorealism. While this technology offers legitimate applications in entertainment and privacy protection, it also raises significant concerns about potential abuse to create fake profiles, spread misinformation and undermine public trust in visual media[2, 32]. As highlighted by Liu et al. [19], the realistic nature of synthetic human face images generated by models like StyleGAN and diffusion-based methods poses serious social trust concerns due to their potential exploitation for malicious purposes.

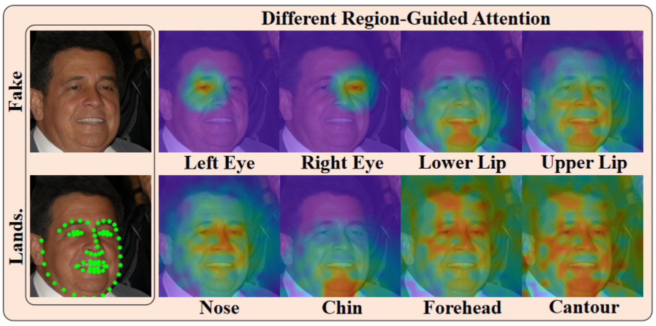

Figure 1: Example visualization of region-guided attention patterns generated by our LAMM-ViT model. The visualization demonstrates how different our region-guided attention heads focus on distinct facial regions with minimal overlap when analyzing AI-synthetic faces. Our method captures diverse forgery clues across various generative techniques, including subtle inconsistencies in texture patterns, unnatural symmetry, blending artifacts, and structural irregularities that persist across different generation methods.

Despite significant research efforts in synthetic image detection, most existing approaches face a critical limitation: poor generalization to new generation techniques not seen during training [32]. This challenge stems from the fact that different generative models introduce different artifacts and patterns in their output. As observed by Wang et al. [34], the generation processes between different models (e.g., GANs, VAEs, diffusion models) are entirely different, rendering previously developed detectors ineffective when confronted with images from novel generation methods. Furthermore, Jeong et al. [13] note that owing to extraordinary advancements in synthesis technology, an increasing array of distinctive frequency-level artifact representations have emerged, further complicating detection.

Current detection methods fall into two categories: space-based methods analyzing pixel-level patterns and frequency-based methods examining spectral properties. Spatial-based approaches often employ CNN classifiers with various data pre-processing or augmentation strategies [19, 32], while others target specific fingerprints left by the generation techniques [22]. However, Wang et al. [34] note that classifiers trained on certain generators (like ProGAN) struggle with fake images from unfamiliar sources (such as diffusion models). Frequency-domain methods [7, 32] exploit abnormalities in the spectrum of synthetic images, particularly those caused by upsampling operations in generation pipelines. These approaches show promise

[*] Corresponding Author. Email: chenyaxiong@whut.edu.cn

but often fail against newer generation techniques, producing fewer detectable artifacts [30].

In this paper, we propose a novel detection approach that exploits a common vulnerability across diverse generation models: their inability to maintain consistent relationships between facial structures. Our main insight is that while modern generative models are good at creating globally coherent faces, they often introduce subtle inconsistencies in the relationships between facial regions that can be detected by paying careful attention to these regions. This view is consistent with the observation by He et al. that self-supervised models that examine global structure provide a more comprehensive perspective for detecting synthetic content.

To leverage this insight, we present a Mask-Guided Vision Transformer architecture with Layer-aware Mask Modulation (LAMM), which dynamically focuses on critical facial regions and their interrelationships at various feature abstraction levels. Unlike previous approaches that rely solely on spatial or frequency domain analysis, our method uses facial landmarks to create region-specific attention masks that guide the model toward discriminative facial features across spatial dimensions. This approach is partially inspired by Chen et al. [2], who demonstrated that specific facial regions contain important detection cues, but we enhance this concept through our dynamic masking approach.

The LAMM module adaptively recalibrates the attention mask at different network depths, allowing the detector to capture forgery clues across multiple abstraction levels. Similar to how FreqNet [30] incorporates frequency learning within CNNs, our approach integrates facial region awareness within a transformer architecture. We further introduce a region-gated multi-head attention mechanism that selectively modulates attention based on facial regions. This enhances the model's ability to detect subtle inconsistencies that persist across different generation methods.

The main contributions of our work are as follows:

- We introduce a region-gated multi-head attention mechanism that selectively modulates attention to key facial areas, enabling the detection of subtle artifacts across different generation methods.
- We propose a novel facial landmark-guided Vision Transformer architecture with Layer-aware Mask Modulation (LAMM) that dynamically focuses on discriminative facial regions for improved detection of AI-generated face images.
- We conduct extensive experiments on different datasets generated by various diffusion models and GANs and show that our method significantly outperforms state-of-the-art methods in cross-dataset generalization scenarios.

2 Related Work

In this section, we present a comprehensive overview of existing research on AI-generated face detection. We categorize current techniques into three main categories: image-based detection approaches, frequency-domain detection methods, and attention-guided detection mechanisms.

2.1 Image-based AI-Generated Face Detection

Early methods for detecting AI-generated faces primarily focused on exploiting spatial artifacts in the pixel domain. Rossler et al.[25] utilized the Xception architecture for deepfake detection, demonstrating effective performance on high-quality synthetic media. Several approaches have targeted specific facial regions - Li et al.[17] focused on eye region inconsistencies, while Haliassos et al.[11] examined mouth movement irregularities. Face X-ray[16] identified blending boundaries between forged faces and backgrounds, while SBIs [28] expanded on blending-based forgery detection.

To address generalization challenges, Wang et al.[32] showed that a detector trained with careful data augmentation on a single specific CNN generator could generalize to unseen architectures. As synthetic media technology advances, approaches like FakeSpotter [31] utilize layer-wise neuron behavior for classification, while GramNet [20] leverages the Gram matrix to extract global texture as a robust representation. ICT [5] models identity differences in inner and outer facial regions to better identify inconsistencies across various generation techniques.

2.2 Frequency-based Detection Methods

A significant body of research has demonstrated that frequency domain analysis can reveal artifacts invisibly embedded in AI-generated images. Frank et al.[8] and Durall et al.[7] observed that CNN-generated images consistently fail to reproduce realistic spectral distributions, suggesting fundamental limitations in current generative models. F3-Net [23] explores frequency statistics differences between real and fake images, employing both frequency-aware decomposed components and local frequency statistics to capture forgery patterns.

More specialized frequency-based approaches include FrePGAN [13], which converts RGB images to frequency maps to highlight generation artifacts, and FDFL [15], which proposes adaptive frequency feature learning to mine subtle artifacts. LOG [21] integrates information from both color and frequency domains through a two-branch recurrent network, while BiHPF [12] emphasizes amplifying artifact magnitudes through dual high-pass filters. Wang et al. [33] introduces dynamic graph learning to exploit relation-aware features across spatial and frequency domains.

2.3 Attention-Guided Detection Approaches

Attention mechanisms have proven highly effective for deepfake detection by focusing on discriminative facial regions. MAT [35] pioneered attentional mechanisms to highlight suspicious regions, while Stehouwer et al. [3] introduced attention guided by ground truth manipulation masks. Several approaches have employed multi-headed attention modules to correlate low-level textural features with high-level semantics at different facial regions, though fixed attention paradigms limit adaptation to diverse forgery types.

Recent Vision Transformer (ViT) based approaches show particular promise due to their inherent attention mechanisms. FTCN [37] extracts temporal information using specialized attention, while PCL [36] employs region-specific attention to extract distinct source features. However, Chen et al. [1] note that most existing methods use fixed attention weights across network layers, limiting their ability to detect hierarchical facial forgery artifacts. Our LAMM-ViT differs by dynamically adjusting attention at different network depths using facial landmarks and layer-specific parameters, enabling detection of structural inconsistencies across diverse generation techniques at multiple feature abstraction levels.

3 Methodology

To address the challenge of detecting AI-generated faces with high generalization capability, we propose a novel framework, the Mask-Guided Vision Transformer with Layer-aware Mask Modulation

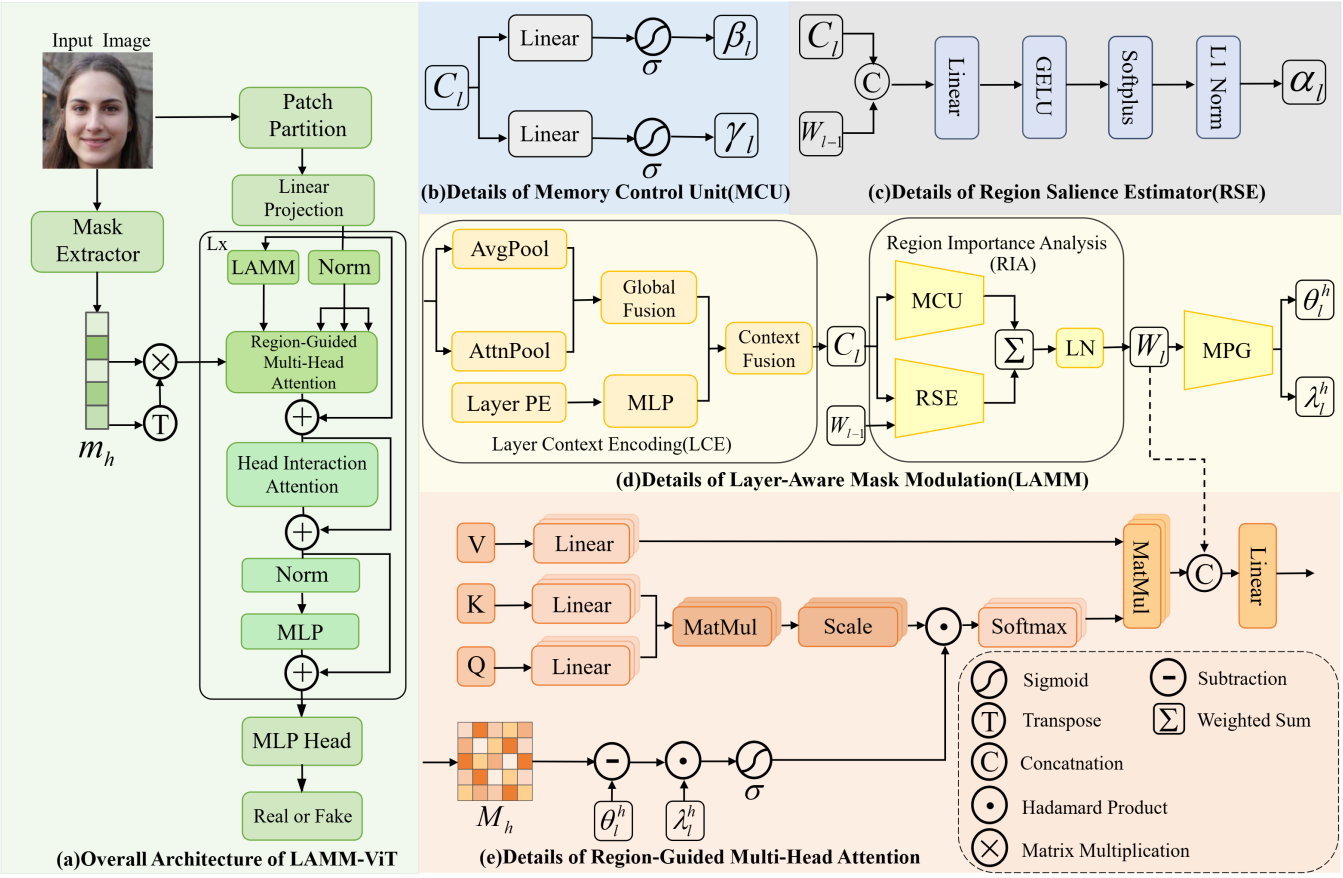

Figure 2: Overall architecture of the proposed LAMM-ViT. (a) Main pipeline showing the integration of Mask Extractor, Region-Guided Multi-Head Attention, Head Interaction Attention (which is a simple self-attention mechanism), and LAMM within the ViT blocks. (b) Details of the Memory Control Unit (MCU) used in LAMM. (c) Details of the Region Salience Estimator (RSE) used within LAMM's RIA component. (d) Detailed breakdown of the Layer-Aware Mask Modulation (LAMM) module, including Layer Context Encoding (LCE) and Region Importance Analysis (RIA) which generates layer-specific mask weights W_l and gating parameters $(\theta_l^h, \lambda_l^h)$ via the Mask Parameter Generator (MPG, detailed in Section 3.3.3). (e) Details of the Region-Guided Multi-Head Attention (RG-MHA) mechanism, showcasing the region gating process.

(LAMM-ViT). Our approach enhances the standard Vision Transformer (ViT) [6] architecture by incorporating explicit facial region guidance and dynamically adapting this guidance across different network layers, indexed by l. The overall architecture is depicted in Figure 2(a).

3.1 Input Processing and Mask Generation

Given an input face image $I \in \mathbb{R}^{H\times W\times C}$, where H, W are height and width, and C is the number of channels, we first extract facial landmarks using an off-the-shelf detector. We then generate continuous Gaussian masks for K key facial regions (eyes, nose, mouth, etc.), resulting in a multi-channel region mask tensor $R \in \mathbb{R}^{K\times H\times W}$.

Concurrently, the input image I is processed into patch embeddings $X_p \in \mathbb{R}^{N_p\times D}$, where $N_p = HW/P^2$ is the number of patches, P is the patch size, and D is the embedding dimension. A learnable class token x_{cls} is prepended, and positional embeddings E_{pos} are added to form the initial sequence $X_0 \in \mathbb{R}^{(N_p+1)\times D}$.

The region map is then processed by the **Mask Processor**. This component projects each region mask from size $H \times W$ to a patch-level vector of dimension N_p, resulting in K initial mask vectors. To capture relationships between regions, additional combined mask vectors are created via learnable weighted sums of these initial vectors, targeting specific facial groupings. This results in a total of H mask vectors (where H equals the number of attention heads), comprising both individual regions and their combinations.

Finally, a zero vector is prepended to represent the mask for the CLS token, forming the final mask tensor $\mathcal{M} \in \mathbb{R}^{H\times(N_p+1)}$. Each mask vector $m^h \in \mathbb{R}^{N_p+1}$ in $\mathcal{M}$ corresponds to a specific region or combination for potential guidance.

3.2 Region-Guided Transformer Block

Each Transformer block in our framework enhances standard attention mechanisms by incorporating region-aware processing. This design guides the model to focus on specific facial regions and their interactions, enabling more effective detection of inconsistencies in AI-generated faces.

The RG-MHA mechanism, shown in Figure 2(e), adapts the standard multi-head self-attention to focus on facial region inconsistencies. From the input sequence X_{l-1} of layer l, queries Q_l, keys K_l, and values V_l are computed linearly.

To enable region-aware attention, we first construct an attention gating mask $M^h \in \mathbb{R}^{(N_p+1)\times(N_p+1)}$ for each attention head h. This mask is derived from the corresponding mask vector $m^h \in \mathbb{R}^{N_p+1}$ from the mask tensor $\mathcal{M}$ via an outer product:

$$M^h = m^h(m^h)^T, \tag{1}$$

where m^h is the mask vector for head h, and M^h is the resulting attention gating mask.

Using this mask, we compute a region gate $G_l^h \in \mathbb{R}^{(N_p+1)\times(N_p+1)}$ that selectively emphasizes attention to specific facial regions and their interactions:

$$G_l^h = \sigma(\lambda_l^h \cdot (M^h - \theta_l^h)), \tag{2}$$

where σ is the sigmoid function, and λ_l^h and θ_l^h are layer-specific gating parameters dynamically generated by the Mask Parameter Generator (MPG) in Section 3.3.3.

The region gate modulates the attention mechanism by element-wise multiplication with the attention scores before softmax normalization:

$$\text{Attention}(Q_l^h, K_l^h, V_l^h) = \text{softmax}\left(\frac{Q_l^h(K_l^h)^T}{\sqrt{d}} \odot G_l^h\right) V_l^h, \tag{3}$$

where $\odot$ is element-wise multiplication, $d = D/H$ is the dimension per head for a Transformer with H attention heads, and Q_l^h, K_l^h, and V_l^h are the query, key, and value matrices for head h in layer l.

Finally, the outputs from all heads are combined using a weighted concatenation strategy. Each head's output is weighted by its corresponding layer-specific weight W_l^h, which is an element of the mask weight vector $W_l = [W_l^1, W_l^2, ..., W_l^H]$ generated by the LAMM module for layer l, where $h \in \{1, 2, ..., H\}$ represents the index of each attention head out of a total of H attention heads.

After processing through the attention mechanism and feed-forward network, the output of each layer forms the sequence X_l, which serves as input to the subsequent layer.

3.3 Layer-Aware Mask Modulation

The LAMM module dynamically adjusts how different facial regions influence attention at each network depth. This enables the model to progressively refine its focus on discriminative facial features across different abstraction levels. As illustrated in Figure 2(d), LAMM generates layer-specific parameters that control the RG-MHA mechanism: mask weights W_l for weighting head outputs, and gating parameters that control the strength and threshold of regional attention.

3.3.1 Layer Context Encoding

The Layer Context Encoding component captures the network's state at each layer l, providing essential context for adaptive parameter generation. LCE computes a context vector C_l specific to encoder layer l by combining information from two sources.

First, layer position information PE_l is encoded for each layer index l using a learned embedding approach. This embedding captures the depth-specific characteristics of the network.

Second, global features g_l are extracted from the input sequence through pooling methods and then combined:

$$g_l = \text{LayerNorm}(\text{Linear}(\text{Concat}(g_l^{avg}, g_l^{att}))), \tag{4}$$

where g_l^{avg} and g_l^{att} are the average-pooled and attention-pooled features.

Finally, the context fusion mechanism integrates the layer position information with the global feature state:

$$C_l = \text{LayerNorm}(\text{MLP}(\text{Concat}(g_l, PE_l))), \tag{5}$$

3.3.2 Region Importance Analysis

The Region Importance Analysis component determines which facial regions should receive more attention at each layer. RIA dynamically updates the mask weights W_l used in RG-MHA by leveraging both current layer information and accumulated knowledge from previous layers. It consists of a Region Salience Estimator (RSE) that assesses the current importance of each region, and a Memory Control Unit (MCU) that balances new information with previously learned patterns.

These components work together to update the mask weights through a recurrent-like mechanism:

$$W_l = \gamma_l \odot \alpha_l + \beta_l \odot W_{l-1}, \tag{6}$$

where α_l represents the new region importance scores computed by the RSE based on layer context C_l, γ_l and β_l are adaptive coefficients produced by the MCU that control the balance between new information and historical knowledge.

3.3.3 Mask Parameter Generator (MPG)

The Mask Parameter Generator produces the specific parameters that control regional gating in the attention mechanism. Taking the layer context C_l and the mask weights W_l as input, it generates two sets of parameters.

The gating strength parameters λ_l control how strongly each head emphasizes its assigned regions:

$$\lambda_l = \text{FC}(C_l) \odot \text{MLP}(\text{Concat}(C_l, W_l)), \tag{7}$$

where FC is a fully connected layer and MLP is a multi-layer perceptron. The parameter λ_l is then used to derive head-specific gating strength parameters λ_l^h for each attention head.

The gating threshold parameters θ_l determine the sensitivity of each head to its assigned regions:

$$\theta_l = \theta_{base} + \text{MLP}(\text{Concat}(C_l, W_l)), \tag{8}$$

where θ_{base} is a base threshold value. Similarly, θ_l is used to derive head-specific threshold parameters θ_l^h. These adaptively generated parameters enable each attention head to dynamically adjust its regional focus based on both network depth and learned feature representations.

3.4 Loss Function

Our training objective combines classification accuracy with a novel diversity-promoting mechanism specifically designed to enhance generalization across various generation techniques.

For the primary task of distinguishing between real and AI-generated faces, we employ the standard Cross-Entropy loss ($\mathcal{L}_{\text{ce}}$):

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)], \tag{9}$$

where y_i is the ground truth label, $\hat{y}_i$ is the predicted probability for each sample, and N is the number of samples.

To address the challenge of detecting diverse forgery patterns across different generation techniques, we introduce a novel Mask Diversity Loss ($\mathcal{L}_{\text{div}}$). This component leverages our LAMM-ViT's region-guided attention mechanism to encourage the model to utilize different facial region combinations when analyzing different samples. The key insight is that various generative techniques produce artifacts in different facial regions, requiring multiple detection strategies tailored to different artifact patterns.

For this purpose, we utilize the layer mask weights W_l for each input sample i at layer l. These weights reflect how the model assigns region importance during processing.

We first define the cosine similarity between mask weight vectors for two samples:

$$\cos(W_{l,i}, W_{l,j}) = \frac{W_{l,i} \cdot W_{l,j}}{||W_{l,i}|| \cdot ||W_{l,j}||}, \tag{10}$$

where $W_{l,i} \cdot W_{l,j}$ is the dot product between vectors, and $||W||$ represents the Euclidean norm.

The Mask Diversity Loss measures the average similarity between all pairs of sample mask weights across all network layers:

$$\mathcal{L}_{\text{div}} = \frac{1}{L} \sum_{l=1}^{L} \frac{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} \cos(W_{l,i}, W_{l,j})}{N(N-1)}, \tag{11}$$

where L is the total number of layers in the network. Higher pairwise similarity indicates lower diversity in attention strategies, which is penalized by this loss term.

Our total loss function combines these components:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \eta \mathcal{L}_{\text{div}}, \tag{12}$$

where $\eta = 0.2$ was found to yield the best results in our experiments. This balanced approach ensures accurate classification while simultaneously promoting diverse and adaptive attention mechanisms across samples, enhancing generalization capabilities for detecting various types of generation artifacts.

4 Experiments

In this section, we first introduce the overall experimental setup, and then present extensive experimental results to demonstrate the superiority of our method.

4.1 Experimental Setup

Datasets. We conduct experiments on a subset of the AI-Face-FairnessBench [18] dataset, which includes both real and AI-generated face images. The real images are sourced from IMDB-WIKI datasets. For AI-generated images, we include a diverse collection from multiple generative models, categorized into GANs and Diffusion Models (DMs). The GAN-based models include AttGAN, MMDGAN, StarGAN, MSGGAN, STGAN, StyleGAN, StyleGAN2, StyleGAN3, VQGAN and ProGAN. The diffusion-based models include DALLE2, IF, Midjourney, DCFACe, Latent Diffusion, Palette, SD v1.5, and SD Inpainting.

For training our model, we use real images from IMDB-WIKI datasets, along with fake images generated by StyleGAN3, Latent Diffusion, and SD V1.5. We deliberately include multiple generative models in our training set to enable our region-based method to learn the relationship between face regions between different generative techniques which is essential for generalizable detection.

Implementation Details. All images are resized to 224×224 resolution for consistency. We extract facial landmarks from each image using DLIB's[26] 68-point facial landmark detector and generate region-specific masks for the eight facial areas as described in Section 3. We implement our Region-Gated Vision Transformer with 12 transformer layers, 12 attention heads, and embedding dimension of 768. We train using the AdamW optimizer with a learning rate of 1×10^{-4}, weight decay of 0.05, and batch size of 64. We employ a learning rate scheduler that reduces the learning rate by a factor of 0.5 when validation performance plateaus for 3 consecutive epochs. All models are trained for a maximum of 100 epochs with early stopping based on validation accuracy. We implement our approach with PyTorch and use mixed-precision training to improve efficiency.

Evaluation Metrics. Following previous works [8, 22, 32], we evaluate performance using Average Precision (AP) and Accuracy (ACC). For computing ACC, we use a classification threshold of 0.5 following standard practice.

Baselines. We compare our method with several state-of-the-art approaches: (i) Wang et al. [32], which demonstrated that a standard classifier trained on a single CNN generator with careful data augmentation can generalize surprisingly well to unseen architectures; (ii) F3Net [23], which proposed mining frequency-aware clues using decomposed image components and local frequency statistics within a two-stream framework; (iii) Gragnaniello et al. [10], which analyzed the generalization ability of detectors across different GAN architectures and challenging scenarios, studying the impact of augmentation and training strategies; (iv) LGrad [29], which introduced using gradients from a pretrained model as a generalized representation for GAN-generated artifacts; (v) Ojha et al. [22], which proposed using features from large pretrained models (like CLIP), not explicitly trained for fake detection, to achieve better generalization across diverse generative model families; and (vi) FreqNet [30], which leverages frequency space domain learning for improved generalizability.

4.2 Comparison with State-of-the-Art Methods

To evaluate the effectiveness and generalization capabilities of LAMM-ViT, we perform extensive comparisons with state-of-the-art detection methods across 18 diverse generative models. The quantitative results are summarized in Table 1.

Cross-dataset Performance and Model Generalizability. Our model shows exceptional cross-model generalization performance, achieving **94.09%** mean ACC and **98.62%** mean AP across all tested generators, significantly outperforming the strongest baseline (Wang et al. [32] with 88.64% ACC and 95.53% AP) by **+5.45%** in ACC and **+3.09%** in AP.

The key strength of LAMM-ViT lies in its consistent performance across diverse generator types. While competing methods exhibit extreme performance variations, our approach maintains robust accuracy with no catastrophic failures. For instance, F3Net [23] achieves perfect accuracy on several generators (MMDGAN, MSGGAN) but drops to chance level on others (VQGAN, DCFACE), and FreqNet [30] similarly shows inconsistent results across different models.

LAMM-ViT particularly excels on challenging generators where baselines struggle. On StyleGAN and StyleGAN2 where methods like Gragnaniello et al. [10] achieve only ∼50% accuracy, our approach maintains excellent performance (97.40% and 97.14% re-

Table 1: Performance comparison (ACC/AP %) with state-of-the-art methods across 18 diverse AI-generated face models (GANs and Diffusion). Our method (LAMM-ViT) demonstrates superior generalization, achieving the highest mean ACC (94.09%) and AP (98.62%). Best results per generator are highlighted in bold.

Generator	Wang [32]	F3Net [23]	Grag [10]	LGrad [29]	Ojha [22]	FreqNet [30]	Ours (LAMM)
AttGAN	81.81/98.98	84.72/97.93	99.63/99.99	**99.92/100.0**	84.26/94.33	99.54/99.97	82.76/95.70
MMDGAN	92.50/98.32	99.50/**100.0**	99.50/99.99	**100.0/100.0**	87.75/95.69	**100.0/100.0**	99.75/**100.0**
MSGGAN	92.50/99.90	**100.0/100.0**	99.50/**100.0**	**100.0/100.0**	77.50/90.83	99.25/**100.0**	97.00/99.96
StarGAN	72.83/82.94	56.11/83.24	99.82/**100.0**	**99.87**/99.99	95.44/98.96	99.25/99.98	79.12/94.63
STGAN	95.75/99.32	**100.0/100.0**	99.25/99.87	**100.0/100.0**	83.75/93.71	99.75/**100.0**	97.75/99.78
StyleGAN	92.76/98.60	87.58/94.99	49.96/77.99	50.74/88.87	88.46/95.88	50.11/55.22	**97.40/99.73**
StyleGAN2	92.55/98.43	84.72/94.98	49.89/63.27	51.54/82.79	85.38/94.33	50.05/51.02	**97.14/99.65**
StyleGAN3	94.06/99.00	**99.99/100.0**	99.70/**100.0**	99.95/**100.0**	96.17/**100.0**	99.85/**100.0**	97.19/99.54
VQGAN	86.76/94.29	50.30/83.45	49.77/55.80	50.72/77.73	81.60/91.97	52.95/79.20	**94.56/98.68**
ProGAN	93.50/99.11	99.74/**100.0**	99.66/**100.0**	**99.95/100.0**	93.06/97.88	99.46/**100.0**	96.62/99.51
Midjourney	92.50/98.13	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	95.00/98.68
IF	90.59/98.66	**100.0/100.0**	99.01/**100.0**	**100.0/100.0**	96.53/**100.0**	99.01/**100.0**	97.03/99.47
DALLE2	80.49/91.36	**100.0/100.0**	**100.0/100.0**	**100.0/100.0**	97.56/**100.0**	**100.0/100.0**	91.46/97.83
DCFACE	81.43/90.29	50.00/81.05	53.34/76.54	49.93/40.57	72.05/85.54	49.97/46.00	**97.25/99.33**
Latent Diffusion	93.95/99.47	**100.0/100.0**	99.79/**100.0**	99.94/**100.0**	96.13/**100.0**	99.88/**100.0**	97.38/99.69
Palette	84.67/92.76	50.00/76.66	49.79/38.61	50.88/77.96	51.88/56.31	53.75/79.75	**93.29/98.28**
SD Inpainting	84.83/92.59	99.54/**100.0**	99.70/**100.0**	**99.92/100.0**	95.58/99.51	98.98/**100.0**	88.98/96.52
SD v1.5	92.05/97.45	**99.99/100.0**	99.65/**100.0**	99.89/**100.0**	95.74/99.53	98.62/**100.0**	93.89/98.20
Mean	88.64/95.53	86.79/95.12	86.00/89.56	86.29/92.66	87.71/94.13	86.13/89.50	**94.09/98.62**

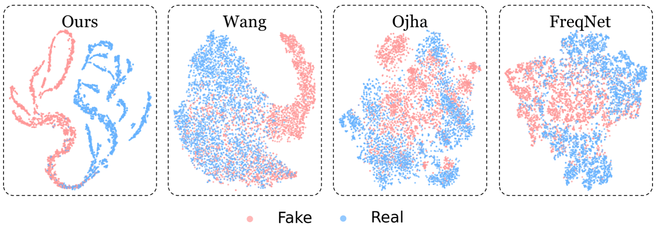

Figure 3: The t-SNE visualization of features extracted from our model and some state-of-the-art models [32, 22, 30] on all test datasets. Our model demonstrates clearer separation between real and synthetic clusters compared to competing approaches.

spectively). For difficult diffusion models like DCFACE and Palette, LAMM-ViT achieves 97.25% and 93.29% accuracy where most competitors perform poorly.

Most notably, LAMM-ViT demonstrates balanced effectiveness across both GAN-based and diffusion-based models without favoring the generative family it was trained on—a crucial advantage for real-world deployment where source generators are typically unknown. This consistent performance suggests our method captures fundamental structural inconsistencies common across generation techniques rather than overfitting to specific artifacts.

Feature Space Analysis. The t-SNE visualization in Figure 3 reveals a distinct separation between real images (blue cluster) and various synthetic image clusters. Unlike previous methods [32, 22, 30] where real and fake clusters often significantly overlap, our feature space maintains clear decision boundaries with logical positioning of different generator families. This structured representation confirms that LAMM-ViT learns the discriminative features of generalization across generative techniques rather than merely detecting model-specific artifacts that frequency-based methods typically target.

Table 2: Robustness evaluation against common image perturbations.

Perturbed	GAN-based		Diffusion-based		Mean	
	ACC	AP	ACC	AP	ACC	AP
No	93.93	98.72	94.29	98.50	94.09	98.62
noise	94.48	98.77	96.25	99.14	95.27	98.93
jpeg	94.33	98.77	94.59	99.17	94.45	98.95
blur	94.62	98.91	94.51	98.57	94.57	98.76
cropping	89.14	92.36	92.71	95.84	90.73	93.91
combined	89.67	93.71	91.95	95.87	90.68	94.67

4.3 Robustness to Image Perturbations

We evaluated LAMM-ViT's resilience against common image manipulations that typically occur in real-world scenarios. Following Frank et al. [8], we applied perturbations to test images with a probability of 50%, including gaussian noise, jpeg compression, blurring, cropping, and a challenging combined scenario. As shown in Table 2,Our model shows significant stability across most perturbations without retraining. The performance under gaussian noise, jpeg compression and blurring is always high, and high accuracy and precision indicators are maintained in these common distortions. Cropping causes a modest decline since it removes important spatial context, yet performance remains robust. Even under the demanding combined perturbation scenario, LAMM-ViT maintains strong performance with only a manageable drop compared to standard conditions. This consistency across different perturbations emphasizes the advantage of LAMM-ViT in focusing on structural relationships between facial regions rather than low-level textures or frequency artifacts that are easily degraded, highlighting its applicability to powerful real-world deployments of increasingly complex synthetic media.

4.4 Ablation Study

Ablation Study on Components. We ablate key components—region Mask guidance, Region-Guided Multi-Head Attention (RG-MHA), and Layer-aware Mask Modulation (LAMM)—to

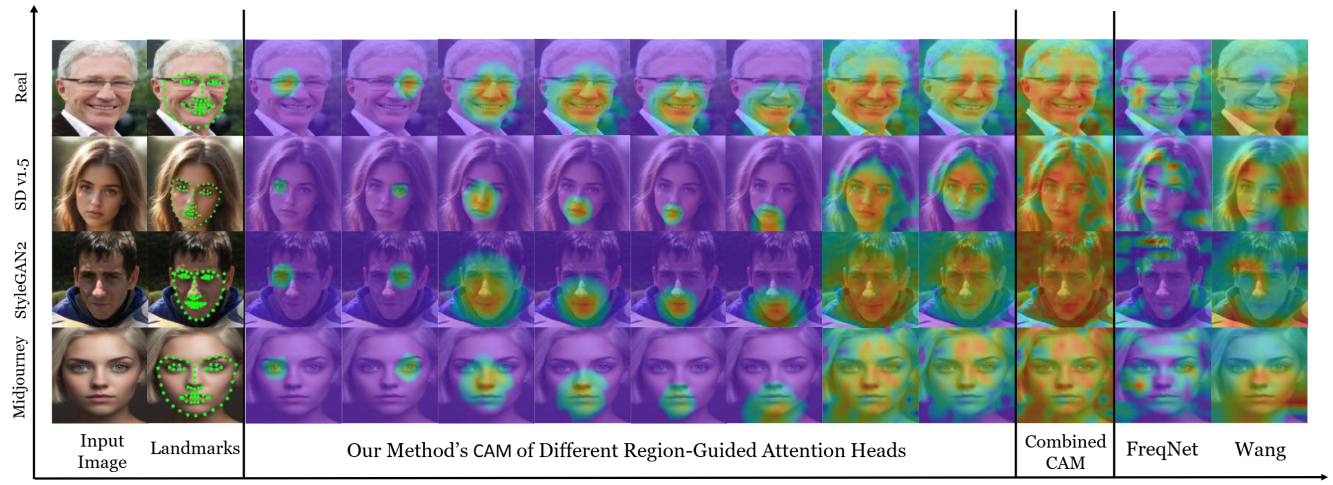

Figure 4: CAM visualizations of LAMM-ViT on various AI-generated faces from Midjourney, StyleGAN2, and SDv1.5. Comparison between our regional head-specific CAM, combined CAM, and baseline methods (FreqNet and Wang).

Table 3: Ablation study on LAMM-ViT components. Each experiment follows the same configuration as our main experiments, varying only the inclusion of specific architectural components.

Component				Mean	
ViT	Mask	RG-MHA	LAMM	ACC	AP
✓	-	-	-	82.37	89.54
✓	✓	-	-	52.57	53.83
✓	✓	✓	-	56.62	61.91
✓	✓	-	✓	43.82	45.75
✓	✓	✓	✓	**94.09**	**98.62**

evaluate their individual contributions. As shown in Table 3, the standard Vision Transformer baseline achieves reasonable performance, but simply adding static facial region masks without the corresponding guidance mechanisms leads to significant performance degradation. This suggests that static masks alone inappropriately restrict the ViT's attention patterns. Similarly, including RG-MHA or LAMM alone yields suboptimal results, indicating the synergistic nature of these components rather than independence. When all components are integrated, our full model demonstrates substantially superior performance, confirming that the dynamic relationship between region-directed attention and layer-specific modulation is critical for effectively capturing forgery patterns.

Ablation Study on Loss Functions. Table 4 presents the comparison between training with only Cross-Entropy loss (CE), only our proposed Diversity loss, and the combined loss function. The results demonstrate that while Cross-Entropy loss alone provides reasonable classification performance, it cannot match the combined approach. Training with only the Diversity loss predictably fails to provide sufficient classification guidance. However, when both losses are combined, we observe substantial improvement, confirming that our proposed Diversity loss successfully encourages the model to learn multiple detection strategies targeting different facial regions, enhancing generalization across diverse generation techniques.

4.5 *Visualization of Region-Gated Attention*

To evaluate our mechanism's interpretability, we visualize attention patterns using Grad-CAM [27]. As shown in Figure 4, our approach offers several noteworthy insights. Firstly, it becomes evident that our method excels in extracting diverse spatial attention cues through its specialized attention heads. Regional CAM visualizations show that the different attention in the LAMM-ViT is focused on different facial regions with minimal overlap which demonstrates the effectiveness of our region-guided design. This provides strong evidence of the orthogonality within extracted regional representations. Our method also captures transition zones between different parts of the face, revealing important spatial relationships. In contrast, baseline methods (FreqNet and Wang et al.) demonstrate more scattered focus, often with attention concentrated on limited or less semantically meaningful areas. These visualizations confirm that our region-gated attention mechanism effectively guides the model to recognize multiple face regions independently thereby contributing to LAMM-ViT robust cross-model generation performance.

Table 4: Ablation study on loss function configurations. All experiments use identical settings to our main experiments, only changing the loss function components.

Loss		Mean	
Lce	Ldiv	ACC	AP
✓	✓	**94.09**	**98.62**
✓	-	89.97	95.73
-	✓	49.95	51.21

5 Conclusion

In this paper, we presented LAMM-ViT, a novel Vision Transformer architecture for detecting AI-generated faces with robust cross-model generalization. By integrating Region-Guided Multi-Head Attention with Layer-aware Mask Modulation, our approach focuses on structural inconsistencies between facial regions—a common weakness across generation techniques. Experiments demonstrated LAMM-ViT's superior performance, achieving 94.09% mean accuracy and 98.62% mean AP across 18 different generative models, significantly outperforming state-of-the-art methods. The model maintains consistent performance on both GAN-based and diffusion-based generators without catastrophic failure, highlighting its practical utility against evolving synthetic media threats. Our approach shows that region-focused, hierarchical attention mechanisms offer a promising direction for developing generalizable forgery detection systems capable of addressing increasingly photorealistic AI-generated content.

Acknowledgements

This work was supported in part by Hainan Province “Nanhai New Star” Technology Innovation Talent Platform Project under Grant NHXXRCXM202361, Sanya Science and Technology Special Fund under Grant 2024KJFX034, the General Program the Hubei Natural Science Foundation (No.2025AFB615). Key Research and Development and Achievement Transformation Program of Inner Mongolia Autonomous Region under Grant 2025YFHH0078.

References

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

[2] L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems*, 35:24597–24610, 2022.

[3] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

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

[5] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9468–9478, 2022.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] R. Durall, M. Keuper, and J. Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.

[8] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020.

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

[10] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2021.

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

[12] Y. Jeong, D. Kim, S. Min, S. Joe, Y. Gwon, and J. Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022.

[13] Y. Jeong, D. Kim, Y. Ro, and J. Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1060–1068, 2022.

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

[15] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6458–6467, 2021.

[16] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[17] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee, 2018.

[18] L. Lin, X. Wang, S. Hu, et al. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. *arXiv preprint arXiv:2406.00783*, 2024.

[19] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 772–781, 2021.

[20] Z. Liu, X. Qi, and P. H. Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020.

[21] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part VII 16*, pages 667–684. Springer, 2020.

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

[23] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

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

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

[26] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[28] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022.

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

[30] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[31] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*, 2019.

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

[33] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7278–7287, 2023.

[34] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

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

[36] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[37] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021.