INFER: Implicit Neural Features for Exposing Realism

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deepfakes pose a significant threat to the authenticity of digital media, with current detection methods often falling short in generalizing to unseen manipulations. *INFER* is the first deepfake detection framework that leverages Implicit Neural Representations (INRs), marking a new direction in representation learning for forensic analysis. We combine high-level semantic priors from Contrastive Language–Image Pre-training (CLIP) with spatially detailed, frequency-sensitive features from INR-derived heatmaps. While CLIP captures global context grounded in natural image statistics, INR heatmaps expose subtle structural inconsistencies often overlooked by conventional detectors. Crucially, their fusion transforms the feature space in a way that enhances class separability—effectively re-encoding both spatial artifacts and semantic inconsistencies into a more discriminative representation. This complementary integration leads to more robust detection, especially under challenging distribution shifts and unseen forgery types. Extensive experiments on standard deepfake benchmarks demonstrate that our method outperforms existing approaches by a clear margin, highlighting its strong generalization, robustness, and practical utility.

1 Introduction

With the rapid progress of deep learning, it has become easier than ever to generate highly realistic synthetic media, including images, videos, and audio. One of the most widely known and debated results of this technology is deepfakes, which is artificial content that is designed to closely mimic real-world media. Today, a deepfake is typically defined as any image, video, or audio clip that has been generated or modified using deep learning methods, often to deceive viewers or mislead them into believing the content is authentic. The term deepfake comes from a combination of deep, referring to deep learning, and fake, indicating that the content is not genuine. Although early attempts to alter video content go back to the 1990s, such as the Video Rewrite system (1997), which altered a person's lip movements in video to match different audio [43]; these methods did not involve deep neural networks. The modern concept of deepfakes only became possible with the rise of powerful deep learning models. In particular, Generative Adversarial Networks (GANs) [56, 3] played a major role in creating realistic synthetic faces and videos. More recently, diffusion models [13] have made it possible to generate even more seamless and photo-realistic content that is difficult to distinguish from real media [9, 6]. As deepfake technology becomes increasingly advanced, and widely accessible [29], the creation of synthetic media is accelerating at a rapid pace. Recent estimates suggest that thousands of deepfakes are now being generated daily, with applications ranging from entertainment and satire to more harmful uses such as misinformation campaigns, identity theft, and financial fraud [22, 15, 19]. These growing risks have sparked widespread concern around media authenticity and digital trust.

In response to the growing threat of deepfakes, researchers have turned to the same technology that enabled their creation, which is deep learning, to develop effective detection methods. Broadly, deepfake detection techniques fall into two main categories: image-based and video-based approaches [24]. Image-based methods focus on analyzing individual frames to identify visual artifacts or inconsistencies, and are often simpler and faster to train [4, 50, 17]. In contrast, video-based methods aim to capture temporal inconsistencies across frames, such as unnatural facial expressions, blinking patterns, or head movements, but typically require more complex models and greater computational resources [60, 71, 27].

While a wide range of deepfake detection methods have been proposed, a persistent challenge remains: generalization to unseen manipulations and datasets. Many models perform well on specific benchmarks but struggle when faced with new deepfake generation techniques or distribution shifts in real-world data. This raises a critical question: *What types of representations can lead to better class separation and more robust detection than traditional approaches?* One promising direction involves the use of features derived from Contrastive Language–Image Pre-training (CLIP) [47]. Recent studies have shown that CLIP features, which encode high-level semantic and visual information, offer improved class separability compared to existing methodologies [44]. Building on this, further work has demonstrated that applying wavelet decomposition to CLIP-derived features can capture localized frequency components, leading to enhanced detection performance [7].

These insights strongly suggest that combining semantic-rich embeddings with frequency-aware representations may offer a promising path toward more generalizable deepfake detection. Motivated by this, we seek an alternative representation, that can be combined with CLIP embeddings, which not only captures frequency characteristics but also retains spatial context, enabling the model to reason about where and how manipulations occur within an image. While many decomposition methods exist, we observe that Implicit Neural Representations (INRs) [57] offer a unique formulation. They model images as continuous functions over spatial coordinates, implicitly encoding both fine-grained structure and frequency content within their network parameters. In doing so, the layer-wise activations of INRs naturally act as a form of spectral decomposition [8], revealing localized frequency responses across the image. Unlike traditional CNNs that operate on fixed grids, INRs provide a flexible and expressive representation that has recently shown promise across various signal domains, including images, audio, and video [57, 49, 53]. This makes them particularly well-suited for capturing the subtle artifacts introduced by generative manipulations. By leveraging the representational power of INRs, we aim to build a more robust and manipulation-sensitive feature space that complements high-level semantic cues and improves generalization to unseen deepfake types. To the best of our knowledge, *this work is the first to explore the use of INRs for deepfake detection, leveraging their spatial-frequency sensitivity to identify manipulation artifacts.*

2 Related works

2.1 Deepfakes

Deepfake detection has become a widely studied domain due to the rise of powerful generative models. Early methods [1, 58, 34] employ a feature encoder followed by a binary classifier to predict manipulated content. XceptionNet [12] is based on depthwise separable convolutions with residual connections. Similarly, CapsuleNet [41] better captures spatial hierarchies in manipulated media. However, these approaches were prone to overfitting and exhibited poor generalization to unseen data. The current deepfake detection landscape can be categorized along two major axes: frame-level vs. video-level detection methods and spatial domain vs. frequency domain methods. Frame-level methods [54, 25, 30] analyze individual frames for manipulation without considering temporal consistency. Video-level methods [65, 68, 21] leverage temporal information across frames to enhance robustness. When it comes to spatial domain approaches [42, 75], they detect inconsistencies at the pixel level. On the other hand, frequency domain approaches [32, 61, 26] focus on spectral artifacts introduced during manipulation. Recently, several works such as LSDA [69] and SBI [55] have proposed dataset augmentation strategies to increase dataset size with high-quality synthetic samples, which has been shown to improve model performance. In contrast, we deliberately avoid using any augmentations in order to highlight the efficacy of INRs in implicitly capturing subtle manipulation artifacts in spatial-spectral domains. Consequently, for a fair comparison, we exclude baselines that employ dataset augmentation. [44] shows the advantage of using semantic CLIP features for deepfake detection. Wavelet-CLIP [7] appends it with additional frequency features obtained using wavelet transform to further improve performance. In our approach, we leverage the superior spatial-spectral decomposition capability of INRs, combined with the semantic richness of CLIP features. Our work falls under the frame-level detection category and utilizes spatial-spectral information derived from INRs to improve deepfake detection performance without the aid of data augmentations.

2.2 Implicit neural representations

INRs are neural networks that model continuous signals, such as images, audio, or video, by learning mappings from input coordinates (e.g., spatial or temporal) to signal values (e.g., RGB intensities or waveform amplitudes) [57]. Unlike traditional discrete representations, INRs encode the signal directly within network parameters, enabling smooth interpolation, compact storage, and high-resolution reconstruction [53]. This continuous formulation makes them especially well-suited for capturing fine-grained structure and spectral properties.

A critical factor in the expressiveness of INRs is their activation functions. Standard activations like ReLU, Sigmoid, and Tanh are proven to be inadequate, as they fail to preserve high-frequency components of the signal when encoded those to INR. To address this, positional embeddings (PEs) were introduced to inject high-frequency information into the input space [63]; however, it has been noted that INRs with PE scheme often fail to generalize well for unseen coordinates. Subsequent work [57] proposed Sinusoidal activations with carefully chosen weight initialization to directly represent high-frequency content. More recent efforts have introduced spatial-spectral compact activations, improving generalization while relaxing initialization constraints [49, 53].

The most prominent application of INRs has been in Neural Radiance Fields (NeRFs) [18], where they model 3D scenes as continuous volumetric functions for photorealistic view synthesis. Beyond NeRFs, INRs have found use in a wide range of tasks, including image and video super-resolution [2], denoising [53, 66, 28], deblurring [31], inpainting [67], and compression of images, videos, and 3D shapes [59]. INRs have also been applied in medical imaging for reconstruction from sparse data [40], audio processing for waveform modeling [57], and hyperspectral imaging [11, 74]. These diverse applications highlight the versatility of INRs as a compact and expressive alternative to traditional discrete models. Despite this broad adoption, none of these works have explored the use of spatial–spectral INR features for deepfake detection. Our work is the first to investigate this direction, revealing INR-derived activations as a powerful and discriminative modality for detecting subtle manipulations in visual media.

3 Methodology

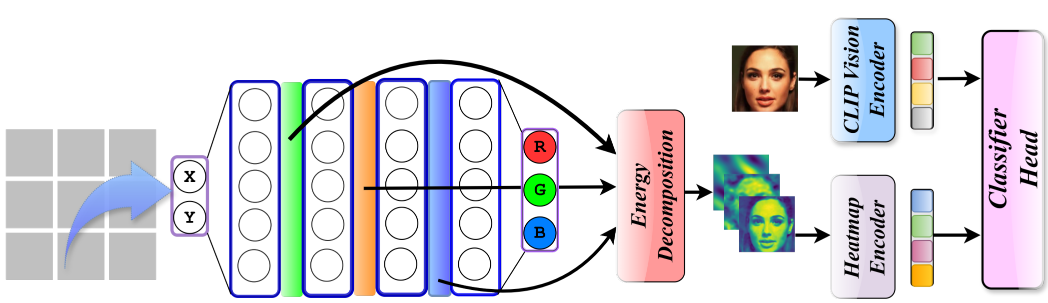

Figure 1: **Overview of the *INFER* Pipeline:** *INFER* begins by associating a spatial coordinate grid with each input image, which is then overfitted using a carefully designed INR. Internal activations from each INR layer are extracted and decomposed using PCA to isolate dominant energy directions. The resulting PCA-based heatmaps are stacked along the batch dimension and processed through a dedicated Heatmap Encoder. In parallel, the RGB image is passed through a CLIP ViT-L/14 encoder to obtain a global semantic embedding. Finally, the INR-derived and CLIP-derived features are concatenated and fed into a classification head for deepfake detection.

3.1 Dataset preparation

To build a robust dataset for training and evaluation, we follow a systematic preprocessing pipeline comprising frame extraction, face detection, and alignment. We begin by extracting frames from each video, followed by face detection using the RetinaFace [16] detector. Detected faces are then cropped based on the bounding boxes and aligned using five facial landmark keypoints. The alignment is performed via a *warp and affine* transformation to standardize the facial geometry across samples. All faces are resized following this alignment process. *INFER* is trained on c23 version of the FaceForensics++ (FF++) dataset [52], which simulates realistic video compression artifacts.When it comes to the number of frames, we extract 10 frames per fake video and 40 frames per real video to curate the training set. This sampling strategy ensures a balanced real-to-fake ratio, which helps minimize class bias during training. A critical goal in deepfake detection is to ensure generalization across unseen forgery types. To assess this, we evaluate the trained model on four out-of-distribution (OOD) benchmarks: **Celeb-DF v1** ($\mathrm{CDF_{v1}}$) [36], **Celeb-DF v2** ($\mathrm{CDF_{v2}}$) [35], **FaceShifter** (FSh) [52], and the **Deep Fake Detection** (DFD) [52] dataset.

3.2 Improving deepfake detection via modality fusion

CLIP embeddings have already shown strong performance in deepfake detection [7] as it excels in capturing high-level semantic cues such as identity, expression consistency, and scene realism [5]. Using a pretrained ViT-L/14 encoder, we extract a global semantic embedding $\mathbf{c} \in \mathbb{R}^{768}$ by feeding in the input image I. These features provide robust scene-wide context; however, they may lack explicit spatial and spectral structure.

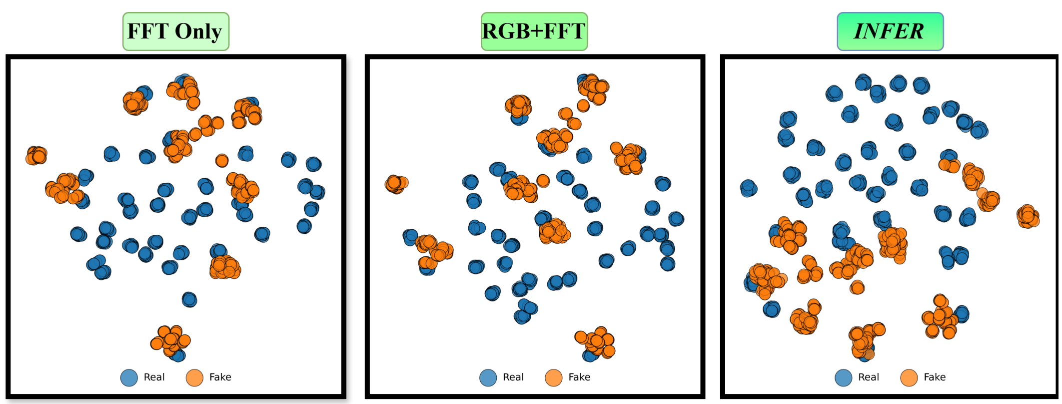

Figure 2: **t-SNE visualization of feature embeddings from the $\mathbf{CDF_{v1}}$ dataset using different input modalities**: A clear progression in class separability is observed: FFT-based features show moderate entanglement between real and fake samples, while combining RGB+FFT yields modest improvement by integrating spatial cues. In contrast, *INFER*-derived features exhibit well-defined, compact clusters with a pronounced margin between classes. This suggests that the spatial–spectral decomposition provided by INR heatmaps restructures the feature space in a way that enhances the separability making analogies to the effect of a kernel transformation in classical machine learning

To address this limitation, we explored whether fusing CLIP embeddings with additional modalities could yield improved separability. Specifically, we combined CLIP features with the RGB image and its FFT-based frequency representation [23] to inject complementary spatial or spectral information (see Section 4.2 for detailed explanation). However, as seen in both Figure 2 (see the first two figures) and Table 2, even though these conventional representations offer some separation in feature space, greater class separability can be achieved through a further transformation on the feature space. Specifically, the first figure of Figure 2 shows that a degree of separation exists when using FFT. However, the second figure further suggests that combining both FFT and RGB transforms the feature space in a way that enhances class separation even more. This behavior is also reflected in the AUC values reported in Table 2. These observations motivate the idea that modality fusion along with CLIP embeddings can improve class separability, but they also raise the question: which modality can further transform the data to enhance this separation? This motivates the need for a new representation that should ideally include both spatial and spectral features while encoding the required discriminative features. To this end, we explore the possibility of using INRs to derive such features in a multiscale and interpretable manner. The following sections demonstrate on how INRs can be leveraged alongside CLIP embeddings to improve deepfake detection through enhanced class separability.

3.3 Formulation of an INR

An INR defines a continuous function that maps spatial coordinates $\mathbf{x} \in \mathbb{R}^2$ to RGB values $s(\mathbf{x}) \in \mathbb{R}^3$. This function is typically implemented as a fully connected neural network $f_\theta : \mathbb{R}^2 \to \mathbb{R}^3$, where θ represents the learnable parameters. Unlike conventional representations [46] that store an image as a discrete grid of pixels, the INR encodes the image in its weights, allowing continuous evaluation at any spatial location. Given a 2D spatial coordinate $\mathbf{x} \in \Omega \subset \mathbb{R}^2$, the network predicts RGB values $\hat{s}(\mathbf{x}) \in \mathbb{R}^3$ through the following layer-wise activations

$$\mathbf{h}_0 = \mathbf{x}, \quad \mathbf{h}_\ell = \phi(\mathbf{W}_\ell \mathbf{h}_{\ell-1} + \mathbf{b}_\ell), \quad \ell = 1, \dots, L-1, \quad \hat{s}(\mathbf{x}) = \mathbf{W}_L \mathbf{h}_{L-1} + \mathbf{b}_L$$

where $\phi(\cdot)$ is a nonlinear activation (e.g., Sinusoid, Gaussian), and $\mathbf{W}_\ell$, $\mathbf{b}_\ell$ are learnable weights and biases respectively. The network is trained to minimize the MSE loss given by $\mathcal{L}_{\text{recon}} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \|f_\theta(\mathbf{x}) - s(\mathbf{x})\|_2^2$, where Ω denotes the set of spatial coordinates in the image domain and $|\Omega| = H \times W$, the H and W represent height and width of the image respectively.

3.4 How can we deploy INRs for deepfake detection?

3.4.1 Limitations of naïve usage

A natural and compelling question is how INRs can effectively be leveraged for the task of deepfake detection. By design, an INR defines a continuous mapping from spatial coordinates to signal values, serving as a compact and differentiable representation of the underlying content [57]. At first glance, this architectural structure appears to offer no more than a mechanism for image reconstruction, ultimately feeding the reconstructed signal into a downstream classifier. This approach is functionally equivalent to using the original image itself and therefore fails to leverage any of the internal representations or structural advantages uniquely offered by INRs. A more promising direction is to utilize the weights of the trained INR as discriminative features directly [39]. However, this approach comes with significant computational overhead. Consider an INR composed of L fully connected layers, each with hidden dimension d_h. The total number of trainable parameters is approximately $(d_h^2+d_h)(L-2)+5d_h+3$, accounting for one input layer, $(L-2)$ hidden layers, and one output layer. Empirically, we find that faithful reconstruction of face images with low reconstruction error typically requires at least three hidden layers and a hidden width of at least 64 neurons *(see Supplementary Material)*, leading to thousands of parameters. Directly feeding these weights into a classifier is therefore computationally expensive and potentially impractical for large-scale deployment.

3.4.2 Spectral bias and representation dynamics

The challenges noted above motivate the need for more efficient and informative INR representations, especially those unique to INRs yet compact and suitable for downstream tasks. One such direction is to explore structural patterns or emergent behaviors within the weight space. A key insight from the INR literature is *spectral bias* [48, 72], where lower-frequency components of the signal are learned earlier during optimization, while higher frequencies emerge later. Despite its empirical support, there is no definitive theory specifying the number of epochs required to learn each frequency band. Furthermore, as each image, whether real or manipulated, follows its own optimization trajectory, designing a universal schedule or analytical tool to probe weight space remains a challenging open problem.

3.4.3 The pathway of a coordinate through the INR

This challenge can be approached by analyzing how an individual spatial coordinate propagates through the layers of an INR, in conjunction with the known phenomenon of spectral bias. Once an INR is trained to reconstruct an image with a minimum L_2 error, the image is no longer stored directly as pixel values. Instead, it is implicitly encoded in the network parameters θ of a function $f_\theta : \mathbb{R}^2 \to \mathbb{R}^3$. As this function takes spatial coordinates $\mathbf{x} = (x, y)$ as input and outputs RGB values

$s(\mathbf{x})$, it effectively captures both the spatial layout and frequency characteristics of the image through the network's parameters in an implicit manner [51]. For any input location $\mathbf{x}$, the network processes it through a series of transformations across L layers, producing a sequence of internal activations $\{\mathbf{h}_\ell(\mathbf{x})\}_{\ell=1}^{L-1}$. This trajectory can be viewed as a coordinate-conditioned representation path, which describes how the INR internally responds to that specific point. Each transformation can be written as $\mathbf{h}_\ell = T_\ell(\mathbf{h}_{\ell-1})$, where T_ℓ denotes the learned mapping at layer ℓ that incrementally refines the previous layer's representation until the final output recovers the original signal.

This layered refinement process is reminiscent of classical signal decomposition methods, such as wavelet transforms [73] or multiresolution pyramids [20], which also emphasize hierarchical encoding. However, unlike handcrafted bases that isolate spatial or frequency information, INRs inherently couple both due to their continuous, coordinate-based formulation. As a consequence, the early layers tend to capture coarse, global features (typically corresponding to low frequencies), while the deeper layers progressively encode finer, localized variations (high frequencies). This behavior closely resembles with with the notion of spectral bias in neural networks.

3.4.4 Extracting interpretable features from INR layers.

We begin by examining the internal activations $\mathbf{h}_\ell(\mathbf{x}) \in \mathbb{R}^{d_\ell}$ at each layer $\ell \in \{1, \ldots, L-1\}$ and spatial coordinate $\mathbf{x} \in \Omega \subset \mathbb{R}^2$. These activations form tensors of size $H \times W \times d_\ell$. While these feature maps encode rich information, they are high-dimensional, difficult to interpret, and infeasible to directly use in downstream classification due to memory constraints.

To obtain a compact yet informative representation, we seek a transformation that reduces each activation vector to a scalar, while preserving the most structurally meaningful content for deepfake detection. From a signal processing perspective, this corresponds to emphasizing high-energy components—regions where the network's response is most active and discriminative. As an initial step, we explored the L_2 norm of the activation vectors. Although smooth and easy to compute, these maps were often dominated by magnitude rather than structure, leading to limited interpretability and poor spatial localization *(See Supplementary material)*.

To address this, we adopt a simple, non-learnable alternative that extracts the dominant energy component of each layer's response. Specifically, we use Principal Component Analysis (PCA) to identify the most expressive direction in the activation space. Projecting each feature vector $\mathbf{h}_\ell(\mathbf{x})$ onto this direction yields a scalar heatmap that summarizes the layer's internal representation at each location. The sequence of PCA-derived scalar maps $\{A_\ell(\mathbf{x})\}_{\ell=1}^{L-1}$ forms a structured representation that captures how an INR distributes signal content across layers. We interpret this set as an approximate multiscale decomposition: $I(x, y) \mapsto \mathbf{a}(x, y) := [A_1(x, y), \ldots, A_{L-1}(x, y)] \in \mathbb{R}^{L-1}$.

3.4.5 Discriminative nature of the multiscale decomposition

Figure 3 presents two examples from the CDF_{v2} dataset: the top row corresponds to a real image, and the bottom row to a deepfake. Each row visualizes the spatial–spectral multiscale decomposition obtained from the INR's internal activations across layers. The final column shows the image reconstructed by the INR, which appears visually similar in both cases despite notable differences in their internal representations. While the quantitative results demonstrate that *INFER* significantly improves deepfake detection across datasets (See Section 4), the proposed decomposition also reveals subtle structural discrepancies, particularly mid-to-deep layers—that are not easily observable in the RGB image or FFT maps. These visual differences provide a glimpse into the discriminative nature of INR-derived representations, though additional non-visible cues encoded in the internal activations may also contribute to the classifier's decision-making process.

In **Layer 1**, both real and fake activations exhibit wave-like patterns with visually high-frequency textures, which may arise due to the deployed sinusoidal activation function in the INR. Despite their appearance, these early activations primarily capture low-level spatial variations and lack semantic distinction, making them visually similar across real and fake images.

By **Layer 2**, the activations begin to reflect mid-level facial structure. For the real image (top), the representation becomes more coherent where it highlights eyes, nose, and mouth regions with smoother transitions. In contrast, the fake image (bottom) shows irregular, noisy responses lacking semantic consistency. This instability suggests the INR struggles to encode manipulated features cleanly at mid-to-deep levels.

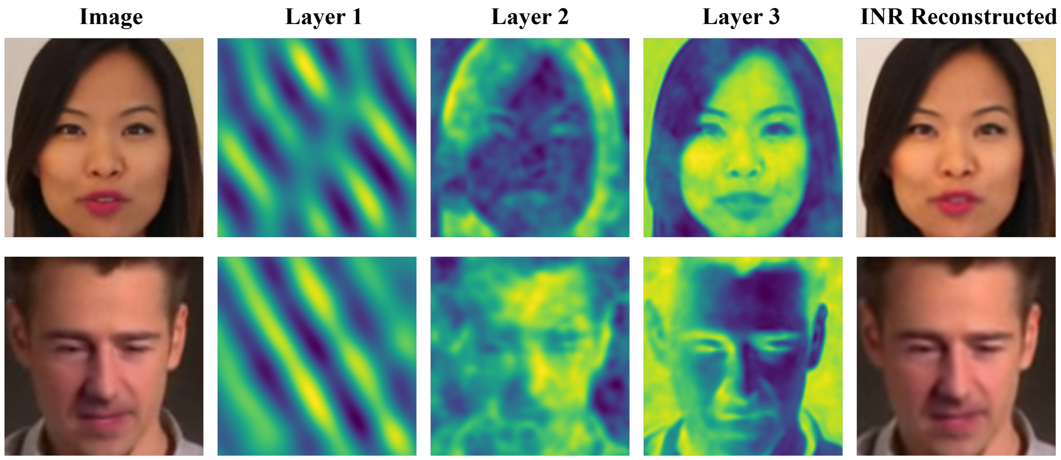

Figure 3: **Despite producing visually faithful reconstructions for both real and fake images (last column), the INR exhibits markedly different internal dynamics across layers**: This visualization underscores a key insight about implicit representations: models can reproduce perceptually accurate outputs while encoding fundamentally different internal pathways. By projecting layer activations via PCA, we expose these hidden trajectories—revealing that while the output may conceal manipulation, the network's internal structure does not.

In **Layer 3**, the differences become more pronounced. The real image produces well-aligned, semantically interpretable activations that faithfully reconstruct identity features, whereas the fake image exhibits distorted contours and exaggerated edge responses—visual evidence of manipulation artifacts that become amplified through the INR's encoding process.

Even though the final INR reconstructions (rightmost column) appear visually similar, the internal activations reveal a clear distinction in representation quality.

3.5 Fusing semantic and multiscale representations

To extract robust and discriminative features from the PCA-projected INR heatmaps, we design a compact convolutional encoder tailored to the spatial–spectral nature of these representations. INR-derived heatmaps encode multiscale structural information across layers but can also exhibit smooth gradients and locally diffuse patterns due to the continuity and frequency sensitivity inherent in the INR formulation. Capturing useful cues from such signals requires an architecture that is both spatially aware and resistant to low-frequency redundancy.

We employ stacked 3×3 convolutional layers to effectively capture local spatial correlations while preserving translational structure. Each convolution is followed by Batch Normalization to stabilize learning and reduce internal covariate shift, and a GELU activation to introduce smooth, non-linear transformations that preserve gradient flow while enhancing expressive capacity. To reduce spatial resolution while retaining global context, we apply an `AdaptiveAvgPool2d` operation that maps the feature maps to a fixed 4×4 resolution, independent of the input size. This is followed by a fully connected projection and Layer Normalization to produce a compact, fixed-dimensional feature embedding.

The heatmap encoder serves as an effective counterpart to the CLIP encoder by transforming localized INR-derived activations into a structured, learnable form. The final CLIP feature and heatmap encoder output are concatenated and passed through a classifier head composed of three fully connected layers with a hidden dimension of 256. This classification module is trained end-to-end using cross-entropy loss to discriminate between real and fake inputs. A visual summary of the entire *INFER* pipeline is shown in Figure 1.

4 Experiments

4.1 Experimental setup

To validate the effectiveness of *INFER*, we conduct extensive experiments across multiple deepfake datasets. The training set consists of videos generated using four popular face manipulation techniques: **Deepfakes**, **Face2Face**, **FaceSwap**, and **NeuralTextures**. These methods span a range of manipulation styles, providing a diverse training distribution. The utilized evaluation datasets, which are already discussed in Section 3.1, are distinct from the training data in both manipulation technique and visual domain, allowing us to rigorously test the generalizability of the learned modules. The performance of the proposed method is measured using the Area Under the Curve (AUC) metric. Further, all the reported values for state-of-the-art (SOTA) methods are either obtained from their respective papers or from [7].

Table 1 summarizes the performance of the proposed *INFER* compared to existing SOTA methods across four widely-used OOD deepfake detection benchmarks ("–" indicates results not reported in prior works). As evident from the results, *INFER* consistently achieves superior AUC scores, demonstrating strong generalization capability even under distribution shift. For the Celeb-DF family of datasets, $\mathrm{CDF_{v1}}$ and $\mathrm{CDF_{v2}}$, *INFER* attains AUC scores of 0.863 and 0.819, respectively. On $\mathrm{CDF_{v1}}$, it outperforms the best prior method, SRM (0.792), by a relative margin of **8.22%**. On $\mathrm{CDF_{v2}}$, it surpasses the best-performing CLIP-based method, which is Wavelet-CLIP (0.759), by **7.32%**. Notably, when compared against plain CLIP (0.743), the improvement is over **9.28%**, validating the complementary nature of the INR-derived modality. On the FSh dataset, *INFER* achieves an AUC of 0.747, outperforming Wavelet-CLIP (0.732) by a relative margin of **2.00%**. For the DFD dataset, both the F-G method and the proposed *INFER* achieve the same AUC score. It can be stated that, *INFER* delivers consistently strong performance across all benchmarks without requiring dataset-specific tuning or modality customization.

Model	**Venue**	**$\mathrm{CDF_{v1}}$**	**$\mathrm{CDF_{v2}}$**	**FSh**	**DFD**	**Avg.**
		General SOTA Methods				
MesoNet [1]	WIFS-18	0.735	0.609	0.566	0.548	0.615
MesoInception [1]	WIFS-18	0.736	0.696	0.643	0.607	0.671
EfficientNet [62]	ICML-19	0.790	0.748	0.616	0.815	0.742
Xception [12]	ICCV-19	0.779	0.736	0.624	0.816	0.739
Capsule [41]	ICASSP-19	0.790	0.747	0.646	0.684	0.717
DSP-FWA [34]	CVPR-19	0.789	0.668	0.555	0.740	0.688
CNN-Aug [64]	CVPR-20	0.742	0.702	0.598	0.646	0.672
FaceX-ray [33]	CVPR-20	0.709	0.678	0.655	0.766	0.702
FFD [14]	CVPR-20	0.784	0.744	0.605	0.802	0.734
$\mathrm{F^3}$-Net [45]	ECCV-20	0.776	0.735	0.591	0.798	0.725
SRM [38]	CVPR-21	<u>0.792</u>	0.755	0.601	0.812	0.740
CORE [42]	CVPR-22	0.779	0.743	0.603	0.802	0.732
RECCE [10]	CVPR-22	0.767	0.731	0.609	0.812	0.730
UCF [70]	ICCV-23	0.779	0.752	0.646	0.807	0.746
F-G [37]	CVPR-24	0.744	–	–	**<u>0.848</u>**	0.796
		CLIP-Based Methods				
CLIP [44]	CVPR-23	0.743	0.750	0.730	–	0.741
Wavelet-CLIP [7]	WACV-25	0.756	<u>0.759</u>	<u>0.732</u>	–	0.749
INFER (Ours)	**–**	**0.863**	**0.819**	**0.747**	**0.848**	**0.819**

Table 1: **AUC performance across cross-dataset evaluations**. The top section lists general SOTA methods, while the bottom focuses on CLIP-based approaches, including the proposed *INFER*.

4.2 Ablation studies

An ablation study was conducted to evaluate which modality provides the most discriminative information when combined with CLIP embeddings for the task of deepfake detection. The comparison includes the standard CLIP module, as well as additional fusion configurations as described below. In the setting labeled FFT, the Fourier transform of the input image is processed through a shallow CNN and its embeddings are concatenated with CLIP features. In the RGB+FFT configuration, both RGB and FFT representations are passed through separate shallow CNNs, and their respective embeddings are fused with CLIP embeddings.

Method	**CDF$_{v1}$**	**CDF$_{v2}$**	**Avg.**
CLIP [44]	0.743	0.750	0.7465
FFT	0.759	0.760	0.7595
RGB+FFT	0.786	0.794	0.7900
INFER	**0.863**	**0.819**	**0.8410**

Table 2: **AUC scores and average performance across CDF datasets**.

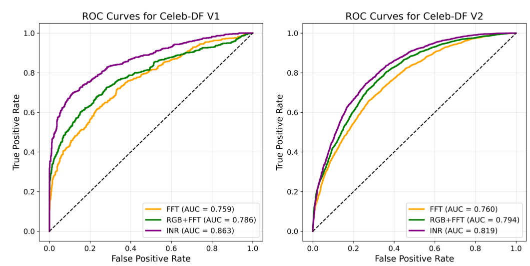

Figure 4: **ROC curves for CDF$_{v1}$ and CDF$_{v2}$**

As can be seen from Table 2, adding FFT features to CLIP embeddings yields a noticeable performance gain, improving the average AUC from 0.7465 to 0.7595 (**+1.71%**), and this performance gain is closer to Wavelet CLIP. Incorporating both RGB and FFT features further improves performance to 0.7900 (**+5.50%** over CLIP), confirming that spatial and spectral cues complement CLIP's semantic information. However, our INR-based method (*INFER*) significantly outperforms all other variants, achieving an average AUC of 0.8410. This represents a **+6.06%** gain over RGB+FFT, and a **+11.24%** improvement over CLIP alone. The corresponding ROC curves for these ablations are provided in Figure 4. These results highlight the strong discriminative power of INR-derived features, which provide a unified spatial–spectral representation that is more expressive than separately extracted RGB or FFT features, even though those are derived directly from the same RGB image. By revealing subtle manipulation artifacts often missed in both spatial and frequency domains, the INR heatmaps supply crucial cues that underpin the performance gains of our approach.

5 Configurations and additional plots

The supplementary materials include detailed explanations of the network configurations used in the INR framework. These cover the selection of activation functions, the reasoning behind specific choices for network depth and the number of hidden neurons, as well as an analysis of why PCA provides better feature representations than L_2 norm-based maps. Moreover, additional visualizations are provided that demonstrate the INR's ability to capture multiscale structural information through its hierarchical decomposition. These materials offer further insight into the design choices and effectiveness of the proposed method.

6 Conclusion

In this work, we propose *INFER*, a deepfake detection framework that synergistically combines semantic embeddings from CLIP with spatial–spectral cues extracted from Implicit Neural Representations (INRs). Unlike traditional approaches that rely solely on either pixel or frequency-domain features, our method leverages INR-derived heatmaps, which capture multiscale structural patterns through a learned continuous implicit function. These heatmaps expose subtle inconsistencies often overlooked by CLIP and conventional CNN-based features. Through extensive experiments across standard deepfake detection benchmarks, we show that INR features significantly boost performance when fused with CLIP embeddings. Compared to standalone CLIP models, *INFER* achieves an average AUC improvement of **+11.24%**, and outperforms other CLIP-based variants such as RGB+FFT by **+6.06%**. These results underscore the complementary nature of INR-derived representations, which offer a richer and more discriminative feature space for detecting manipulated content. Our findings not only demonstrate the efficacy of INR-guided feature decomposition for deepfake detection but also open up new opportunities for applying INRs to other forensic tasks where subtle structural cues are critical. We believe this work lays the foundation for further exploration of implicit representations as a powerful modality in real-world multimedia integrity verification.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[2] Mary Aiyetigbo, Wanqi Yuan, Feng Luo, and Nianyi Li. Implicit neural representation for video and image super-resolution. *arXiv preprint arXiv:2503.04665*, 2025.

[3] Hamed Alqahtani, Manolya Kavakli-Thorne, and Gulshan Kumar. Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, 28:525–552, 2021.

[4] Mohammed Sahib Mahdi Altaei et al. Detection of deep fake in face images based machine learning. *Al-Salam Journal for Engineering and Technology*, 2(2):1–12, 2023.

[5] Andrea Asperti, Leonardo Dessì, Maria Chiara Tonetti, and Nico Wu. Does clip perceive art the same way we do? *arXiv preprint arXiv:2505.05229*, 2025.

[6] Aashutosh AV, Srijan Das, Abhijit Das, et al. Latent flow diffusion for deepfake video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3781–3790, 2024.

[7] Lalith Bharadwaj Baru, Shilhora Akshay Patel, and Rohit Boddeda. Harnessing wavelet transformations for generalizable deepfake forgery detection. *arXiv preprint arXiv:2409.18301*, 2024.

[8] Nuri Benbarka, Timon Höfer, Andreas Zell, et al. Seeing implicit neural representations as fourier series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2041–2050, 2022.

[9] Chaitali Bhattacharyya, Hanxiao Wang, Feng Zhang, Sungho Kim, and Xiatian Zhu. Diffusion deepfake. *arXiv preprint arXiv:2404.01579*, 2024.

[10] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4113–4122, 2022.

[11] Huan Chen, Wangcai Zhao, Tingfa Xu, Guokai Shi, Shiyun Zhou, Peifu Liu, and Jianan Li. Spectral-wise implicit neural representation for hyperspectral image reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3714–3727, 2023.

[12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[13] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[14] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil 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.

[15] Audrey de Rancourt-Raymond and Nadia Smaili. The unethical use of deepfakes. *Journal of Financial Crime*, 30(4):1066–1077, 2023.

[16] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

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

[18] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022.

[19] Alisha Gilbert and Zhigang Gong. Digital identity theft using deepfakes. In *Information Technology Security and Risk Management*, pages 307–314. CRC Press, 2024.

[20] John Goutsias and Henk JAM Heijmans. Nonlinear multiresolution signal decomposition schemes. i. morphological pyramids. *IEEE Transactions on image processing*, 9(11):1862–1876, 2000.

[21] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14950–14962, 2022.

[22] Jeffrey T Hancock and Jeremy N Bailenson. The social impact of deepfakes, 2021.

[23] Paul Heckbert. Fourier transforms and the fast fourier transform (fft) algorithm. *Computer Graphics*, 2(1995):15–463, 1995.

[24] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520, 2024.

[25] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023.

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

[27] Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57(6):159, 2024.

[28] Chaewon Kim, Jaeho Lee, and Jinwoo Shin. Zero-shot blind image denoising via implicit neural representations. *arXiv preprint arXiv:2204.02405*, 2022.

[29] Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. Faster than lies: Real-time deepfake detection using binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3780, 2024.

[30] Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21011–21021, 2023.

[31] Lauri Lehtonen. Implicit neural representations for non-blind depth-aware image deblurring. 2024.

[32] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *arXiv preprint arXiv:2404.13872*, 2024.

[33] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining 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.

[34] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[35] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df (v2): a new dataset for deepfake forensics [j]. *arXiv preprint arXiv*, 2019.

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[37] Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16815–16825, 2024.

[38] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.

[39] Thibault Malherbe. Implicit neural representation as vectorizer for classification task applied to diverse data structures. In *First ContinualAI Unconference-Preregistration Track Second Stage*, 2024.

[40] Amirali Molaei, Amirhossein Aminimehr, Armin Tavakoli, Amirhossein Kazerouni, Bobby Azad, Reza Azad, and Dorit Merhof. Implicit neural representation in medical imaging: A comparative survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2381–2391, 2023.

[41] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2307–2311. IEEE, 2019.

[42] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21, 2022.

[43] Jeremy M. Norman. Video rewrite, origins of deepfakes. `https://www.historyofinformation.com/detail.php?id=4792`, 2025. Accessed: 2025-05-15.

[44] Utkarsh Ojha, Yuheng Li, and Yong Jae 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.

[45] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[46] Majid Rabbani and Paul W Jones. *Digital image compression techniques*, volume 7. SPIE press, 1991.

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

[48] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[49] Sameera Ramasinghe and Simon Lucey. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *European Conference on Computer Vision*, pages 142–158. Springer, 2022.

[50] Ali Raza, Kashif Munir, and Mubarak Almutairi. A novel deep learning approach for deepfake image detection. *Applied Sciences*, 12(19):9820, 2022.

[51] T Mitchell Roddenberry, Vishwanath Saragadam, Maarten V de Hoop, and Richard G Baraniuk. Implicit neural representations and the algebra of complex wavelets. *arXiv preprint arXiv:2310.00545*, 2023.

[52] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[53] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18507–18516, 2023.

[54] Liang Shi, Jie Zhang, Zhilong Ji, Jinfeng Bai, and Shiguang Shan. Real face foundation representation learning for generalized deepfake detection. *Pattern Recognition*, 161:111299, 2025.

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

[56] Simranjeet Singh, Rajneesh Sharma, and Alan F Smeaton. Using gans to synthesise minimum training data for deepfake generation. *arXiv preprint arXiv:2011.05421*, 2020.

[57] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

[58] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil Jain. On the detection of digital face manipulation. *arXiv*, pages arXiv–1910, 2019.

[59] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. Implicit neural representations for image compression. In *European Conference on Computer Vision*, pages 74–91. Springer, 2022.

[60] Shraddha Suratkar and Faruk Kazi. Deep fake video detection using transfer learning approach. *Arabian Journal for Science and Engineering*, 48(8):9727–9737, 2023.

[61] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao 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.

[62] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[63] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

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

[65] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4129–4138, 2023.

[66] Dejia Xu, Peihao Wang, Yifan Jiang, Zhiwen Fan, and Zhangyang Wang. Signal processing for implicit neural representations. *Advances in Neural Information Processing Systems*, 35:13404–13418, 2022.

[67] Wentian Xu and Jianbo Jiao. Revisiting implicit neural representations in low-level vision. *arXiv preprint arXiv:2304.10250*, 2023.

[68] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023.

[69] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024.

[70] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023.

[71] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624, 2021.

[72] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19228–19238, 2022.

[73] Dengsheng Zhang. Wavelet transform. In *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pages 35–44. Springer, 2019.

[74] Kaiwei Zhang, Dandan Zhu, Xiongkuo Min, and Guangtao Zhai. Implicit neural representation learning for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2022.

[75] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. Self-supervised transformer for deepfake detection. *arXiv preprint arXiv:2203.01265*, 2022.

A Supplementary Material

A.1 Choosing the Most Effective Activation Function

As discussed in the main text, the core of an INR lies in its activation function. An inappropriate or conventional activation can often lead to degraded performance in image representation tasks. To assess the most effective activation function, we randomly sampled 100 real and 100 fake images from the FaceForensics++ dataset, following the preprocessing steps outlined in Section 3.1. INRs were then trained using sinusoidal activations from SIREN [57], as well as those introduced in Gauss [49] and WIRE [53].

The table below summarizes the average Peak Signal-to-Noise Ratio (PSNR, in dB) obtained for both real and fake images across the different activation types:

Activation Function	PSNR (Real)	PSNR (Fake)
SIREN	37.41	38.18
Gauss	29.41	29.71
WIRE	20.01	19.73

Table 3: Average PSNR values for real and fake images across different activation functions.

As shown in Table 3, the SIREN model with sinusoidal activation significantly outperforms both Gauss and WIRE across real and fake image reconstructions. Due to its superior performance, SIREN was adopted as the default activation function for all INR-based experiments in this work.

A.2 Choosing the Number of Hidden Neurons

Another important design choice in INRs is the number of hidden neurons in each layer. Increasing this number generally enhances the network's representation capacity, enabling it to capture more complex structures and finer details. However, beyond a certain point, increasing the hidden neuron count may no longer lead to meaningful improvements in reconstruction quality. Specifically, the PSNR often plateaus once the network has reached its capacity to represent the target signal, indicating diminishing returns with further increases in model size. It is worth noting that this behavior can also depend on the type of activation function used.

Similar to the procedure described in Appendix A.1, we randomly sampled 100 real and 100 fake images from the FaceForensics++ dataset and varied the hidden neuron count from 32 to 160 in increments of 32 while keeping the number of hidden layers as two. The resulting average PSNR values for both real and fake images are presented in the left side of Fig. 5.

A.3 Choosing the Number of Hidden Layers

In addition to the number of hidden neurons, the depth of the network, defined by the number of hidden layers, is another key factor that influences the expressiveness of INRs. Deeper networks are generally capable of modeling more intricate patterns and hierarchical structures, potentially leading to better reconstruction quality. However, similar to increasing the number of neurons, increasing the number of hidden layers may also yield no further improvements in reconstruction quality. This phenomenon can be attributed to the combined effects of the activation function and other network parameters.

To analyze the impact of network depth, we varied the number of hidden layers from 1 to 3 while keeping the number of hidden neuron count as 128. Following the same evaluation protocol as before, we randomly sampled 100 real and 100 fake images from the FaceForensics++ dataset and trained INRs under each configuration. The average PSNR values obtained for both real and fake images are summarized in the right side of Fig. 5.

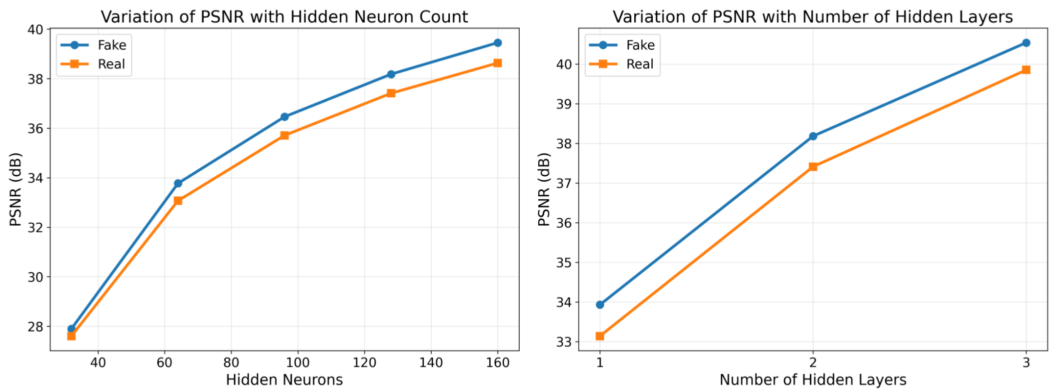

Figure 5: **Average PSNR Variation for both Real and Fake Samples:** Left side plot shows how the average PSNR varies with hidden neuron count while the Right side plot shows how the average PSNR varies with the number of hidden layers

A.4 Utilized INR

For the image reconstruction task through INR, our objective is to achieve at least 35 dB PSNR, as this level reflects high signal fidelity and indicates that the INR has effectively captured the essential structural content of the image. Such a threshold helps ensure that the reconstruction is stable and reliable for downstream analysis, including feature extraction and classification. At the same time, we aimed to avoid overly complex networks with a large number of trainable parameters. To balance reconstruction quality and model efficiency, we selected an INR architecture with sinusoidal activation [57], consisting of 128 hidden neurons and 2 hidden layers.

A.5 INR reconstructions

In addition to proving the quantitative results for INR reconstruction, Figure 6, and Figure 7 showcase how the INR reconstruction quality looks for six different real and fake samples respectively.

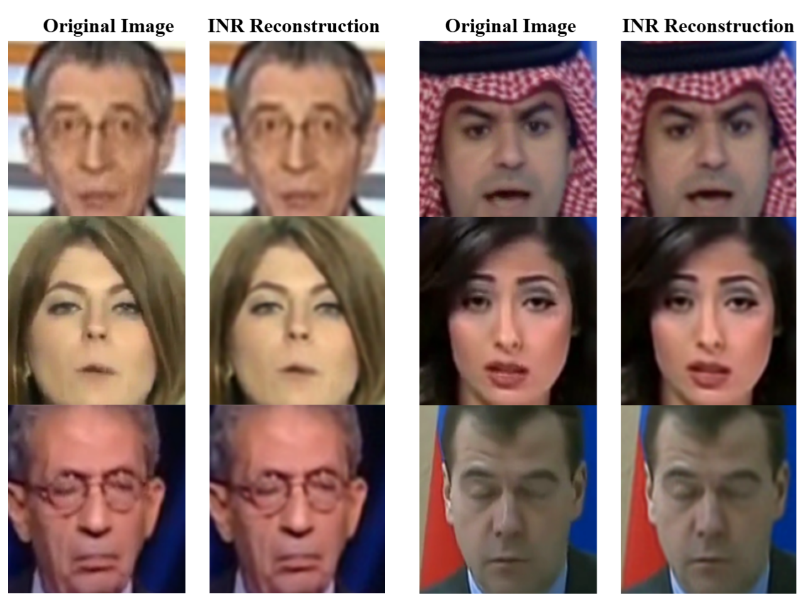

Figure 6: **Original Images and INR Reconstructions for Real Samples**: This figure presents side-by-side comparisons of original real images and their corresponding reconstructions produced by INRs.

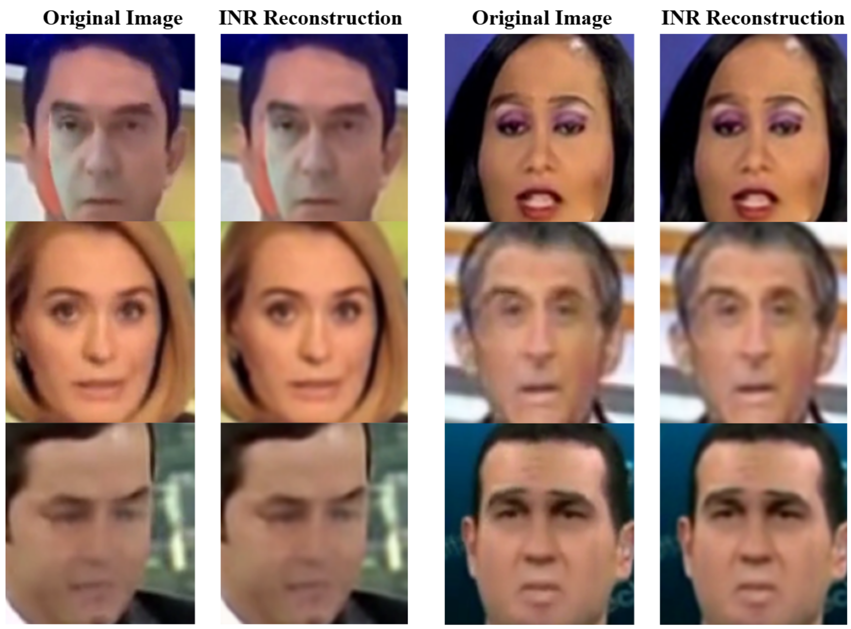

Figure 7: **Original Images and INR Reconstructions for Fake Samples**: Side-by-side comparisons of original fake images and their corresponding INR reconstructions.

A.6 Heatmap Analysis for Different Datasets

In addition to the heatmap visualizations from the CDF_{v2} dataset in the main text, we also present INR-derived heatmaps for CDF_{v1}, DFD, and FSh. These additional visualizations further highlight the ability of INRs to capture structural inconsistencies across different manipulation methods and datasets.

A.6.1 CDF_{v1}

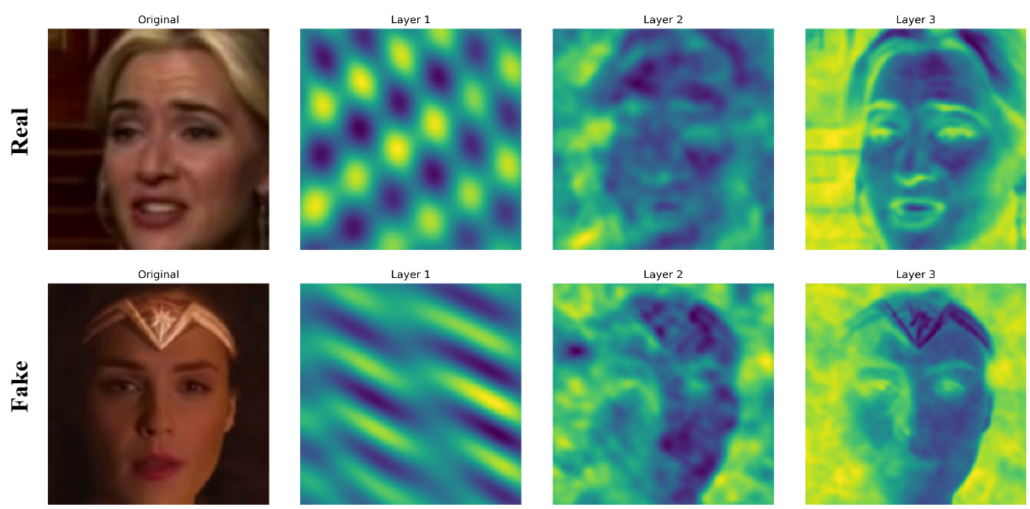

Figure 8: **INR Feature Heatmap Progression for Real and Fake Images (CDF_{v1})**

As can be seen from Figure 8, the first row corresponds to a real image, while the second row shows a deepfake. In the real image, the INR learns progressively meaningful representations: the first layer captures periodic frequency patterns, the second begins to reveal coarse facial structure, and the third cleanly delineates key semantic features such as eyes, nose, and mouth with sharp transitions and spatial coherence. This reflects a natural multiscale decomposition that can be commonly observed in INRs trained on natural content. In contrast, the heatmaps from the deepfake image reveal subtle inconsistencies. While the initial layer shows strong frequency bands, the second and third layers display noisier, less structured activations, particularly in regions like the cheek and jawline. Notably, the third-layer features lack the same spatial sharpness and exhibit localized overactivation near synthetic textures (e.g., the forehead accessory). These differences highlight how INR activations implicitly encode artifacts introduced by manipulation, supporting their utility in forensic analysis.

A.6.2 DFD

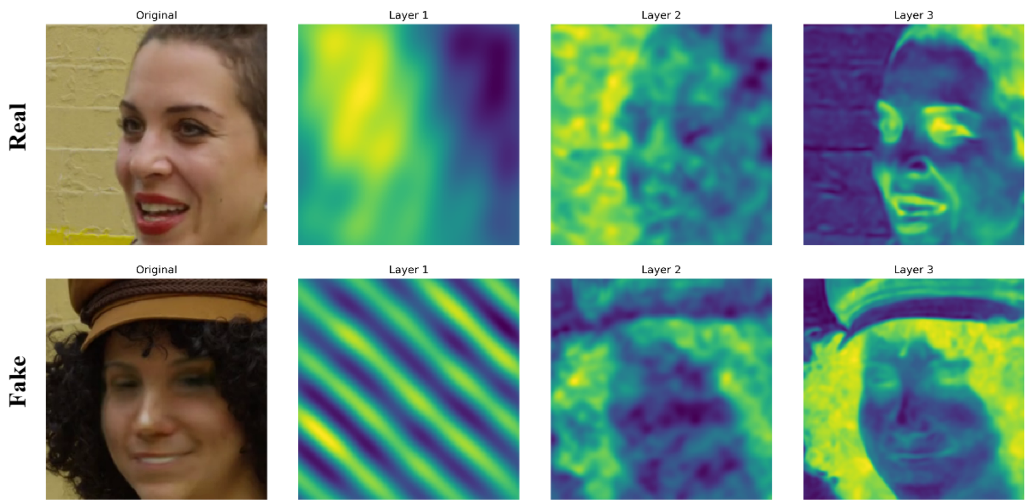

Figure 9: **INR Feature Heatmap Progression for Real and Fake Images (DFD)**

As can be seen from Figure 9, in the real sample (top row), the network exhibits a natural decomposition: the first layer encodes smooth, low-frequency gradients, while subsequent layers progressively extract spatial structure aligned with facial semantics. By the third layer, the representation distinctly highlights the subject's facial features and background texture in a spatially coherent manner. On the other hand, the fake sample reveals signatures of overactivation and structural inconsistency. As the depth increases, the heatmaps become increasingly noisy, with attention distributed unevenly across irrelevant regions such as the background or accessories (e.g., hat, hair). The third layer lacks the focused delineation observed in the real case, underscoring the INR's struggle to generalize to synthetic artifacts. These observations highlight the discriminative potential of INR-derived features in distinguishing real from fake content.

A.6.3 FSh

As can be seen from Figure 10, in the real image (top row), the network exhibits a natural and structured activation flow. The first layer encodes smooth, diagonal sinusoidal frequencies. By the second layer, coherent facial structures begin to emerge. In the third layer, semantic features such as the eyes, mouth, and hairline become sharply defined, with strong localization and contrast — indicating confident learning of meaningful spatial content. In contrast, for the fake image, the deep layers tend to be spatially noisy and less well-formed activations in layers 2 and 3. Although the overall face layout is still present, the details are less distinct. Key features like the mouth and eyes appear blurred or over-smoothed, and the network spreads attention more uniformly, suggesting difficulty in modeling fine-grained semantics. These differences align with patterns observed across fake content, where subtle inconsistencies in structure and texture impede robust INR representation learning. This highlights the sensitivity of INR-derived heatmaps to manipulation artifacts.

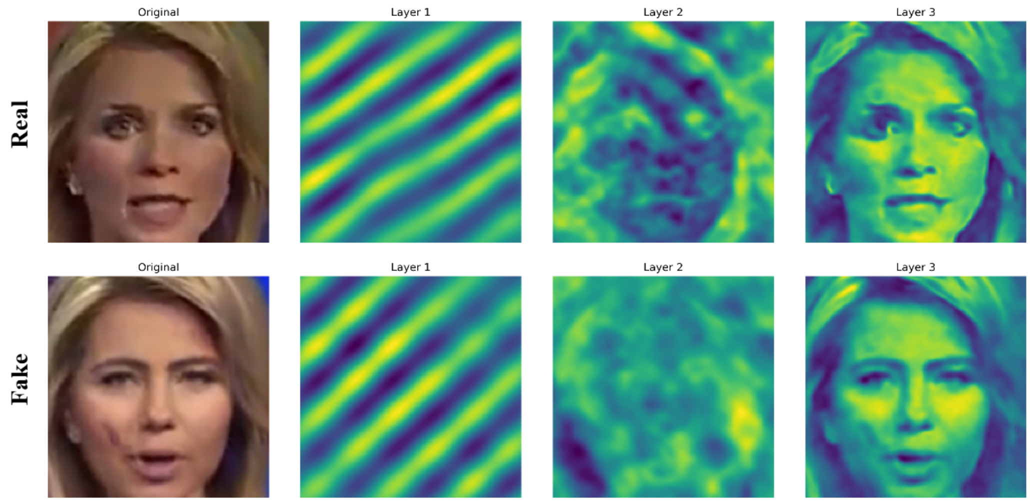

Figure 10: **INR Feature Heatmap Progression for Real and Fake Images (FSh)**

A.7 Feature Space Analysis

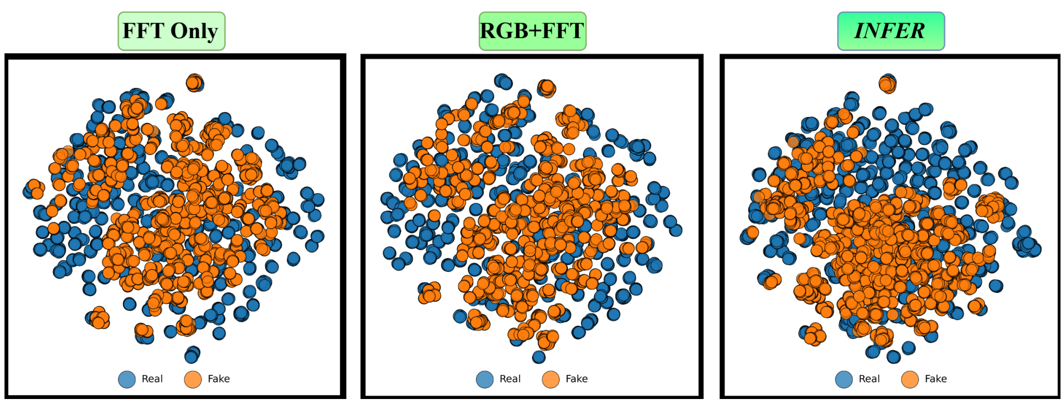

Figure 11: **t-SNE visualization of feature embeddings from the CDF_{v2} dataset using different input modalities**

To better understand how different feature combinations affect the structure of the learned representation space, we visualize the embeddings of real and fake samples using t-SNE for three configurations as shown in Figure 11. Each configuration involves concatenating the respective features before classification. These plots reveal how the choice of representation transforms the feature space and impacts class separability.

FFT Only (Left): This configuration concatenates global frequency information (via the FFT magnitude spectrum) with CLIP embeddings. The FFT captures the global energy distribution across frequencies, but discards all spatial localization. While this can detect abnormal high-frequency content typical of manipulations, it cannot tell where these signals occur which is a critical limitation for identifying local artifacts. As many fake traces are spatially sparse or structured (e.g., boundary mismatches or warped facial regions), this global representation leads to significant overlap between real and fake distributions in the t-SNE space. Moreover, FFT is phase-agnostic in this setup, meaning structural information embedded in phase is ignored. CLIP contributes semantic context but lacks pixel-level sensitivity. As a result, the combined representation fails to disentangle class boundaries effectively.

RGB + FFT (Middle): Here, raw image pixels, FFT features, and CLIP embeddings are concatenated. While this introduces spatial information through RGB and captures frequency cues through FFT, the representation is not explicitly organized to reflect multi-scale spatial-frequency patterns. Even though FFT complements this with frequency statistics, it still lacks localization. Consequently, the feature space becomes more structured than the FFT-only case, but real and fake samples still exhibit considerable intermixing, suggesting insufficient separation.

INFER (Right): The proposed *INFER*, where INR-derived heatmaps are concatenated with CLIP embeddings, results in the most well-separated clusters. INRs reconstruct images from continuous coordinates, and the resulting heatmaps capture how different spatial positions activate the network. These activations inherently encode localized frequency responses, much like a learned multiscale basis decomposition. From a signal processing perspective, INRs offer a unique advantage: they disentangle an image's representation into a hierarchy of frequencies conditioned on position. This means they capture both what frequencies are present and where, which is similar to a spatially adaptive filter bank. Fake images, which often contain unnatural local discontinuities, exhibit distinct activation behaviors in these heatmaps compared to real images. When concatenated with CLIP, which provides semantic structure, the combined representation becomes highly expressive: local inconsistencies are aligned with global semantics, resulting in a well-structured, and a more separable space. This is visually evident from the transformation that both real and fake clusters have undergone compared to Left and Middle figures.

A.8 Grad-CAM Analysis

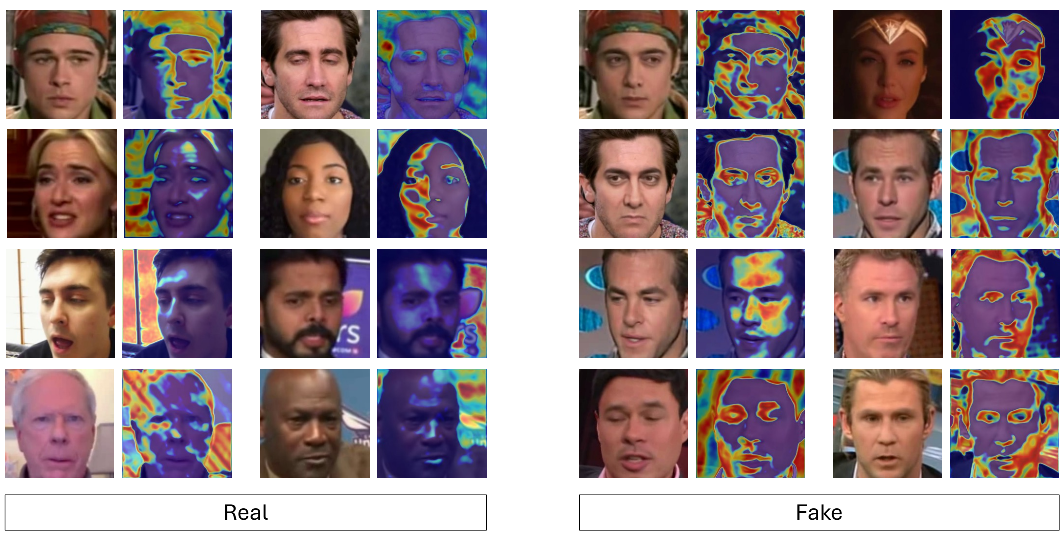

Figure 12: **Activation Maps for Real and Fake Images**

The Grad-CAM visualizations reveal distinct attention patterns for real and fake images, highlighting the complementary roles of semantic and structural cues in *INFER*, as shown in Figure 12. For real faces, the heatmaps are diffuse, with activations spilling into the background and being distributed across broad facial regions rather than tightly clustering around specific landmarks. This suggests that, in the absence of obvious distortions, the detector relies on the overall consistency of textures, both in the background and on the face, rather than on narrowly defined semantic features. In contrast, when processing deepfake images, the attention drifts outward toward peripheral zones such as the hairline boundaries and jawline contours, as well as toward landmark regions like the eyes, nose, and mouth. These are precisely the areas where synthesis artifacts commonly appear, including blending errors, texture irregularities, and subtle warping. This shift in attention arises from *INFER*'s integration of INR-derived features: by overfitting a sinusoidally activated INR to each input and extracting multiscale activation heatmaps via PCA, *INFER* captures fine-grained frequency-domain distortions that standard CNN backbones and CLIP embeddings often overlook. When these INR heatmaps are concatenated with CLIP's semantic embeddings, the downstream classifier learns to look where the fakes break, prompting Grad-CAM to highlight artifact-rich regions in fake images. Consequently, *INFER* enhances robustness by guiding the detector to attend not only to plausible facial geometry but also to the subtle structural inconsistencies that are characteristic of deepfakes.