Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
DOI : 10.47974/JIOS-2374

Attention consistency for robust deepfake detection : An information-theoretic perspective

Swati Jadhav [†]
Department of Computer Engineering & Technology
Dr. Vishwanath Karad MIT World Peace University
Pune
Maharashtra
India

Sagar Mohite [*]
Department of Computer Science and Engineering,
Bharati Vidyapeeth (Deemed to be University)
College of Engineering
Pune 411043
Maharashtra
India

Vaishali Rajput [§]
Department of Artificial Intelligence and Data Science
Vishwakarma Institute of Technology
Maharashtra
India

Rashmi Ashtagi [‡]
Anusha Pai [@]
Santushti Betgeri [#]
Department of Computer Engineering & Technology
Dr. Vishwanath Karad MIT World Peace University
Pune
Maharashtra
India

[†] *E-mail:* **swati.jadhav@mitwpu.edu.in**
[*] *E-mail:* **sgmohite@bvucoep.edu.in (Corresponding Author)**
[§] *E-mail:* **vaishali.rajput@vit.edu**
[‡] *E-mail:* **rashmi.ashtagi@mitwpu.edu.in**
[@] *E-mail:* **anusha.pai@mitwpu.edu.in**
[#] *E-mail:* **santushti.betgeri@mitwpu.edu.in**

Abstract

The increasing realism of Deepfake media generated by modern models poses significant risks to digital trust and forensics. Although Vision Transformer-based detectors outperform Convolutional Methods, they remain vulnerable to unstable attention under common perturbations and often operate in closed-set settings, leading to overconfidence on unseen forgeries. This work analyzes attention consistency as an information-theoretic approach to enhance detection reliability. We propose an Attention-Consistent Vision Transformer (ViT-ACR) that employs Kullback–Leibler divergence to stabilize self-attention distributions. An entropy-based decision rule enables uncertainty-aware open-set inference. Experiments under identity-disjoint protocols across multiple public benchmarks demonstrate strong cross-dataset generalization, achieving 99.28% accuracy on FaceForensics++ and 89.57% mean AUC on six unseen datasets, with reduced predictive uncertainty under compression.

Subject Classification: *68T45, 94A17, 68M25.*

Keywords: *Deepfake detection, Vision transformer, Attention consistency, Entropy, Information security.*

1. Introduction

The evolution in generative artificial intelligence has brought significant paradigm shifts in the creation and distribution of digital media content. In this regard, with the use of highly efficient generative models such as Generative Adversarial Networks (GANs) and the more recently proposed diffusion models [4], Deepfakes can be generated with high fidelity in terms of identity, expression, and speech. Although such advancements in technology have eased the creation process in the creative and commercial domains, they have also introduced dangerous threats in terms of digital trust, discussions in cyberspace, biometric authentication systems, and forensic science in courts [2][5][15].

Deep learning models, particularly Convolutional Neural Networks (CNNs), dominated this area by focusing on the learning of regional texture artifacts and blending inconsistencies [6]. Nevertheless, such techniques lack scalability and generalize poorly to contemporary, high-realism forgeries.

However, current research has indicated that there is a significant accuracy drop exceeding 30% for CNN-based detectors when faced with new manipulation techniques, compression settings, or post-processing operations compared to their performance on unseen datasets [10][12].

Overconfidence on the part of the detectors can have fatal consequences in forensic applications. In recent research, it has been stressed that the issue of deepfake detection is becoming more important in terms of cybersecurity and digital trust. According to the research conducted in [29], the issue of deepfake detection is urgent because the concept will protect digital identities, reduce misinformation, and preserve data integrity in the modern cyber ecosystem. Similarly, [30] research indicates the development of a PKI-oriented cryptographic framework based on a facial micro-expression analysis to enhance the security and reliability of the deepfake detection systems. These studies

underscore that, in addition to detection accuracy, there is a growing necessity for resilient, dependable, and security-conscious detection mechanisms. Nonetheless, the majority of current methodologies predominantly emphasise classification performance, neglecting concerns associated with attention stability and open-set reliability.

However, despite the above benefits, current ViT-based approaches have two essential limitations. First, the Attention Maps obtained from existing ViT-based models are prone to drift when encountering benign conditions such as noise, blur, or compression. Second, most ViT-based approaches operate under a closed-set assumption, lacking uncertainty estimation or open-set rejection mechanisms.

1.1. *Research Gap*

Although the currently popular transformer-based models improve detection accuracy, they are often black-box classifiers that tend to overfit towards meaningless patterns instead of meaningful forensic features [16][18][20]. In particular, frameworks that promote reliable attention and decision-making under practical perturbation constraints and, at the same time, provide uncertainty-based decision-making under new forgery categories remain largely absent in the literature [23].

1.2 *Contributions*

The main contributions of this work are as follows:

- Attention Consistency Regularization (ACR)

  A new form of regularization is used to make self-attention distributions invariant to stochastic transformations, so that the model concentrates on inherent and consistent forgery patterns and ignores noise.

- Open-Set Recognition via Predictive Entropy

  An entropy-based Open-Set Recognition mechanism is introduced to identify and reject samples generated by unseen manipulation techniques. By thresholding the predictive entropy of the ViT output distribution, ViT-ACR flags high-uncertainty samples as unknown forgeries at inference time, enabling reliable deployment in real-world scenarios where novel Deepfake methods may not have been present during training.

2. Related Work

Facial forgery detection and Deepfake research have seen significant advancements in the past ten years due to the development in generative models and the ever-growing use of synthetic media in the world. Currently available solutions are generally divided into CNN-based detectors, physiological and temporal cue-based detectors, robustness/adversarial-aware solutions, transformation-based detectors, and generalization/open-set solutions.

2.1 *CNN-based Deepfake Detection*

Amerini et al. [6] used optical-flow-based CNN models to reveal motion artifacts associated with sub-optimal video frame generation, while Dang et al. [11] systematically studied the dependency of CNN models on various face manipulation pipelines.

The work of Kohli and Gupta [13] showed that frequency-aware CNN models could enhance robustness to some forgery attacks, namely FaceSwap and Face2Face. Bansal et al. [7] discussed the use of other computational-intelligence techniques, such as ensemble and hybrid architectures, to enhance the accuracy of Deepfake video detection.

2.2 *Physiological and Temporal*

The work in [21] used the dynamics of the lips to detect fake videos based on subtle lip movement patterns. There has also been success in multi-modal methods utilizing audiovisual synchronization in face forgeries, specifically lip sync attacks [26]. Similarly, Yang et al. [14] exploited inconsistencies in estimated head pose between real and fake facial regions as a physiological forensic cue.

2.3 *Robustness and Adversarial*

For instance, Carlini and Farid [8] showed that white-box and black-box attacks can successfully avoid CNN-based detectors by adding negligible perturbations to the system. Later, Hou et al. [9] found that adversaries with statistical adversarial consistency were also able to overcome stronger detectors, even when trained from large amounts of data.

2.4 *Vision Transformer-Based Detection*

UIA-ViT proposed an inconsistency-aware transformer in an unsupervised learning setting for identifying manipulated areas based on face components [16]. DFDT introduced an end-to-end deepfake detection framework based on ViTs [18]. M2TR used multi-scale and modal attention for improving robustness in Deepfake detection [20]. Notwithstanding this progress, existing ViT-based detectors are largely vulnerable to the instability of attention triggered by benign perturbations and generally operate under a closed-set assumption. The convolutional Vision Transformer hybrid was proposed by Wodajo et al. [19] to enhance the accuracy of frame-level Deepfake video detection.

2.5 *Generalization & Open-Set Detection*

Spatiotemporal transformers, like the MeST-Former, have demonstrated better generalization performance but are based on fixed decision thresholds, assuming a closed set for the problem [23].

Table 1 summarizes and compares these transformer-based Deepfake detection methods based on the metrics of attention stability, multi-scale modeling, open-set capability, and cross-dataset generalization.

Table 1
Comparison of Transformer-Based Deepfake Detection Methods

Method	Backbone	Attention Stability	Multi-Scale / Temporal Modeling	Open-Set Capability	(Cross-Dataset AUC)	Key Limitation
UIA-ViT [16]	Vision Transformer	✗ (No explicit regularization)	✗	✗	72.1%	Sensitive to perturbations; closed-set only
DFDT [18]	Vision Transformer	✗	✗	✗	71.4%	Fixed decision boundary
M2TR [20]	Multi-Scale Transformer	✗	✓	✗	72.4%	High complexity; no OSR
MeST-Former [23]	Spatio-Temporal Transformer	✗	✓	✗	74.2%	No uncertainty modeling
Proposed ViT-ACR	Vision Transformer	✓ (ACR enforced)	✗	✓ (Entropy-based OSR)	89.57%	—

*Note: Cross-Dataset AUC- CDF, WDF, DFDC, DFD, DFR, DFDC-P

However, open-set implementations for transformer-based detectors remain relatively unexplored.

3. Methodology

3.1 *Problem Formulation*

Let $D_{train} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the supervised training dataset, where $x_i \in R^{H\times W\times C}$ represents a facial image and $y_i \in \{0,1\}$ denotes the corresponding class label for real or fake content.

In real-world deployment scenarios, the test distribution D_{test} may include samples generated using previously unseen manipulation techniques, denoted as $x_p^{unknown} \notin D_{train}$.

3.2 *Vision Transformer Backbone*

Given an input image$x$, it is first partitioned into N non-overlapping patches of size$P \times P$. Each patch is flattened and projected into a D-dimensional embedding space using a learnable linear projection

$$E \in R^{P^2C\times D}.$$

The resulting patch embedding are concatenated with a learnable class token z_{cls} and combined with positional embeddings E_{pos} to form the initial token sequence:

$$Z_0 = [z_{cls}; x_{p_1}E; x_{p_2}E; \ldots; x_{p_N}E] + E_{pos}$$

The embeddings are subsequently processed through L stacked Transformer encoder layers. Each layer consists of a Multi-Head Self-Attention (MSA) module followed by a feed-forward network (FFN):

$$Z'_{\ell} = MSA(\ln(Z_{\ell-1})) + Z_{\ell-1},$$
$$Z_{\ell} = FFN\big(\ln(Z'_{\ell})\big) + Z'_{\ell},$$

Where LN (.) denotes layer normalization.

The self-attention operation is computed as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V$$

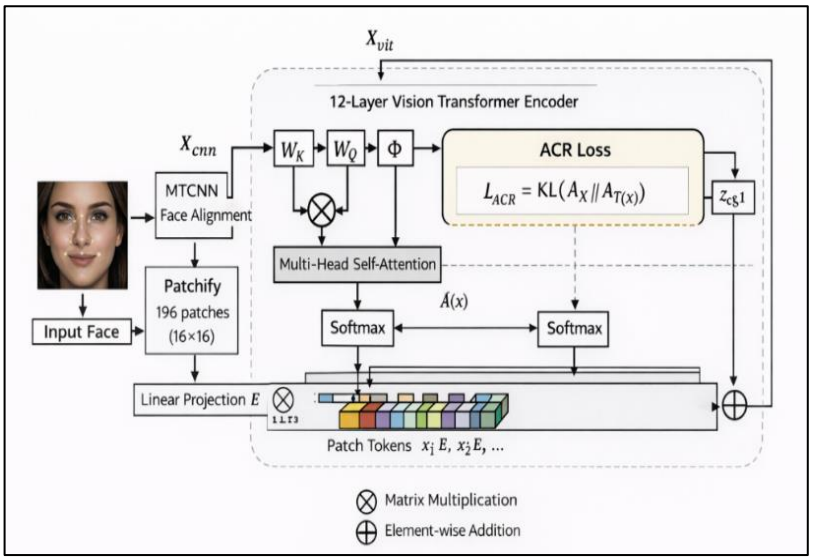

Figure 1
Overall architecture of the proposed ViT-ACR framework.

Where Q,K and V represent the query, key, and value projections, respectively.

Figure 1 shows the architecture of the proposed ViT-ACR framework. After face alignment using MTCNN [27], the input image is divided into 196 non-overlapping 16 × 16 patches and projected into a D-dimensional embedding space. A learnable class token and positional embeddings are added before processing through a 12-layer Transformer encoder consisting of multi-head self-attention and feed-forward networks.

3.3 *Attention Consistency Regularization (ACR)*

Though Vision Transformers are effective in learning global dependencies in the data, there can be considerable variation in the attention values with varying levels of noise injection, blurring, or compression.

As shown in Figure 2, a stochastically transformed input image $T(x)$ and the original image $x$are fed into the same Vision Transformer encoder. The attention maps $A_x$and $A_{T(x)}$extracted from the final self-attention layer are treated as probability distributions over patch tokens.

The ACR module calculates a Kullback-Leibler (KL) divergence loss between the two attention distributions:

$$L_{ACR} = KL(A_x \parallel A_{T(x)})$$

Where KL(.) denotes the Kullback–Leibler divergence.

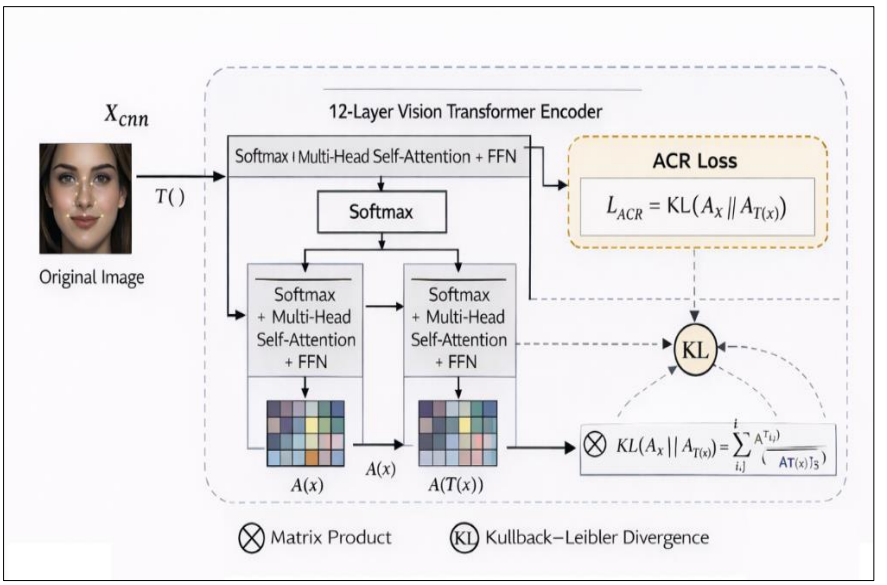

Figure 2
Attention Consistency Regularization mechanism

Which penalizes the difference in responses to attention triggered by stochastic augmentation. Note that no additional matrix multiplication or re-normalization is carried out after the attention has been extracted, as the KL divergence is evaluated directly between the re-normalized features and the Attention Maps for mathematical correctness to be preserved.

3.4 *Classification Objective*

The final classification head processes the class token. The supervised training is done using binary cross entropy loss:

$$L_{BCE} = -[ylog(\hat{y}) + (1-y)\log(1-\hat{y})]$$

Where $\hat{y}$ denotes the predicted probability of a fake sample.

The overall training objective is defined as:

$$L_{total} = L_{BCE} + \lambda L_{ACR},$$

Where λ controls the relative contribution of the attention consistency regularization term.

3.5 *Open-Set Recognition via Predictive Entropy*

For dealing with the open-set detection problem, the uncertainty-aware inference approach based upon the predictive entropy is used. Given the predicted class probabilities

$$p = [p_{real}, p_{fake}]$$

The entropy is computed as:

$$H(p) = -\sum_{c\in\{real,fake\}} p_c \log p_c$$

During inference, samples with entropy exceeding a predefined threshold τ are classified as unknown forgeries:

$$Decision(x) = \begin{cases} \{argmax_c p_c, & if H(p) \le \tau, \\ Reject, & otherwise \end{cases}$$

Samples whose entropy values are above a given threshold, τ, are labeled as unknown forgeries during the inference phase:

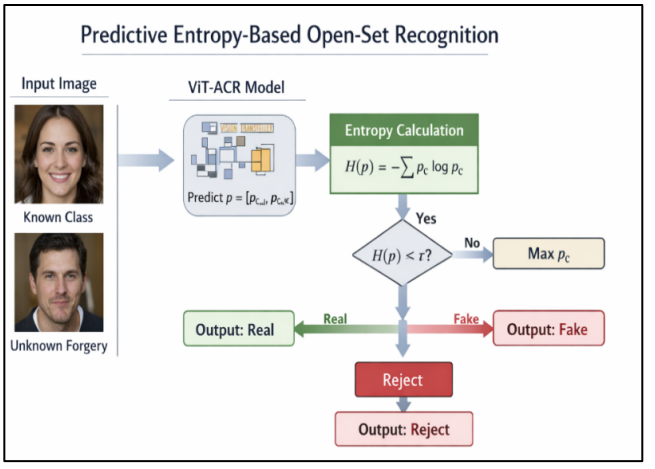

Figure 3
Predictive entropy–based open-set recognition.

Figure 3 illustrates the predictive entropy–based open-set recognition mechanism, where samples with entropy below threshold $\tau$are classified as real or fake, while those exceeding $\tau$are rejected as unknown forgeries.

3.6 *Implementation Details*

The ViT backbone consists of $L = 12$Transformer layers with hidden dimension $D = 768$and 12 self-attention heads. Optimization is performed

using AdamW with weight decay of 0.05 and a cosine learning rate scheduler initialized at 10^{-4}. Experiments are conducted under identity-disjoint evaluation when identity annotations are available.

Proposition.

Let A and A' represent the distributions of self-attention of an input sample and its perturbed counterpart, respectively. Minimizing the divergence between A and A' through KullbackLeibler (KL) divergence, limits the changes in attention in perturbation. Since predictive entropy is conditioned on the underlying attention distribution, this restriction minimizes variability in entropy estimates, which results in more predictive uncertainty estimates.

Proof: Transformer attention matrices are row-stochastic and can be regarded as discrete probability distributions on interactions between tokens. The model achieves distributional consistency and not feature-level similarity by minimizing the KL divergence between attention distributions of original and perturbed inputs. As predictive entropy is calculated using the output probabilities obtained using these representations, stabilization of attention distributions indirectly stabilizes entropy estimates.

4. Experiments and Results

All experiments are conducted using frame-level training and video-level testing. During inference, predictions from 30 uniformly sampled frames are averaged to obtain the final video-level prediction. All evaluations employ an identity-disjoint split, ensuring that identities in the testing set never appear in the training data.

4.1 *Experimental Setup*

All experiments were conducted using the PyTorch framework on NVIDIA A100 GPUs. The ViT-ACR model was initialized from pre-trained ViT-B/16 ImageNet weights. Training used the AdamW optimizer with a learning rate of 1×10⁻′, a batch size of 32, and a cosine annealing schedule over 30 epochs. Data augmentation included random horizontal flipping, color jitter, and Gaussian blurring. Face regions were detected and aligned using MTCNN [27] and resized to 224×224 pixels prior to input.

4.2 *Datasets*

The applicability of the developed ViT-ACR framework is determined in seven public face forgery benchmarks, including in-distribution and out-of-distribution manipulations:

- FaceForensics++ (FF++) [24]
- Celeb-DF v2 (CDF) [3]
- WildDeepfake (WDF)
- Deepfake Detection Challenge (DFDC) [25]

- DFDC-Preview (DFDC-P)
- DeepfakeDetection (DFD)
- DeeperForensics-1.0 (DFR)

The training is done strictly on FaceForensics++; all other datasets are used for cross-dataset evaluation to measure the performance of generalization.

Table 2

Details of datasets used for training and testing

Dataset	Usage	Real / Fake Videos
FaceForensics++ (FF++)	Train	720 / 2880
Celeb-DF v2 (CDF)	Test	178 / 340
WildDeepfake (WDF)	Test	396 / 410
DFDC-Preview (DFDC-P)	Test	276 / 504
Deepfake Detection Challenge (DFDC)	Test	19,154 / 100,000+
Deepfake Detection (DFD)	Test	363 / 3068
DeeperForensics-1.0 (DFR)	Test	11,000 / 11,000
FaceSwap (FS)†	Test	Included within FF++ manipulations
Total evaluated frames	–	≈ 2.1 million frames

Table 2 presents the usage of the dataset and the number of real and fake videos. In total, about 2.1 million frames are tested, providing reliable results. † FaceSwap is treated as a manipulation subset derived from FaceForensics++ and is included separately only for manipulation-specific analysis.

The detection protocol ensures that test samples consist of unknown manipulation patterns, making them much more realistic.

4.3 *Intra-Dataset Evaluation on FaceForensics++*

Intra-Dataset Experiments on FaceForensics++: We first performed experiments

- C23 (high-quality compression)
- C40 (low-quality compression)

The results are presented in Table 3, where ViT-ACR is compared with representative baselines based on CNNs and attention-enhanced models like XceptionNet [24] and MAT [28].

Table 3

Intra-dataset evaluation on FF++

Method	FF++ (C40) ACC	FF++ (C40) AUC	FF++ (C23) ACC	FF++ (C23) AUC
XceptionNet [24]	81.00	82.14	95.10	96.30
MAT [28]	86.79	88.35	97.86	99.10
ViT (Baseline)	90.34	93.72	98.11	99.32
ViT-ACR (Ours)	92.71	96.64	99.28	99.90

For both cases of compression, ViT-ACR models reach high performance. In particular, under the difficult C40 case, ViT-ACR boosts AUC performance by +8.29% compared to MAT [28]. To isolate the contribution of Attention Consistency Regularization, we additionally compare against a vanilla ViT-B/16 model trained under the same protocol. The results show that ACR consistently improves both accuracy and AUC, particularly under the challenging C40 compression setting, demonstrating enhanced robustness to compression-induced perturbations.

4.4 *Cross-Dataset Evaluation*

All models are trained on FaceForensics++ (C23) and evaluated on six unseen benchmarks using video-level AUC. ViT-ACR achieves the highest performance across all datasets, obtaining an average cross-dataset AUC of 89.57%.

Table 4

Cross-dataset evaluation on six unseen datasets (AUC %)

Method	CDF	WDF	DFDC-P	DFDC	DFD	DFR	Avg
Xception [24]	70.70	67.15	73.12	63.46	77.10	75.07	71.10
MAT [28]	81.07	72.31	80.49	70.59	85.71	83.81	79.00
ViT-ACR (Ours)	93.83	84.32	85.41	78.32	94.88	98.01	89.57

Table 4 results validate the effectiveness of attention consistency regularization in enhancing generalization to unseen forgery distributions. The observed improvement over CNN and attention-CNN baselines is attributed to attention consistency regularization, while robustness gains over the vanilla Vision Transformer are analyzed separately in Section 4.5.

4.5 *Robustness to Real-World Perturbations*

To assess robustness under realistic acquisition and post-processing conditions, the models are evaluated on FF++ data under noise, blur, and mixed perturbations. The results are reported in Table 5 and compared against XceptionNet [1] and the vanilla Vision Transformer.

Table 5

Robustness evaluation under real-world perturbations (ACC %)

Method	std/rand	std/sing	std/mix3	std/mix4
Xception [1]	94.75	88.38	82.32	–
ViT [16]	96.77	93.53	90.05	87.06
ViT-ACR (Ours)	98.76	96.27	94.03	91.54

ViT-ACR maintains more than 90% accuracy across all perturbation conditions with a significant improvement over CNN and traditional ViT models. The results demonstrate that attention sensitivity to noise and distortion is effectively alleviated by ACR.

4.6 *Ablation Studies*

Ablation studies are used to estimate the contribution of each component. All variants are trained on FF++ (C23).

4.6.1 Effect of Freezing Pre-trained Parameters

Results are shown in Table 6. The freezing of the pre-trained ViT backbone helps to improve the AUC values as well as Accuracy.

Table 6

Ablation on freezing pre-trained parameters

Method	ACC	AUC
Without freezing	95.43	99.14
With freezing	97.29	99.51

Freezing stabilizes pre-trained representations and reduces over-fitting to compression artifacts.

4.6.2 Effect of Attention Consistency vs Parameter-Efficient Fine-Tuning

Table 7 compares LoRA-based parameter-efficient fine-tuning with the proposed Attention Consistency Regularization.

Table 7

Comparison between LoRA and ACR

Method	**FF++ ACC**	**CDF AUC**
LoRA	96.82	92.21
ACR (Ours)	99.28	99.60

Although LoRA improves parameter efficiency, ACR substantially outperforms it in detection accuracy, and this proves that attention stability is relatively more important than parameter efficiency.

4.7 *Qualitative Analysis*

The results indicate that ViT-ACR exhibits strongly localized attention around manipulated facial regions, whereas the baseline ViT shows more diffuse attention patterns.

Figure 4 shows that ViT-ACR focuses on discriminative facial regions, whereas the baseline ViT exhibits more diffuse attention patterns and greater sensitivity to compression artifacts.

Figure 5 demonstrates the robustness of ViT-ACR under heavy compression. While the baseline ViT misclassifies the forged image, ViT-ACR correctly identifies it by maintaining focused attention on discriminative facial regions.

Variance reduction is computed as the relative decrease in the variance of predictive entropy with respect to a vanilla ViT baseline under FF++ (C40).

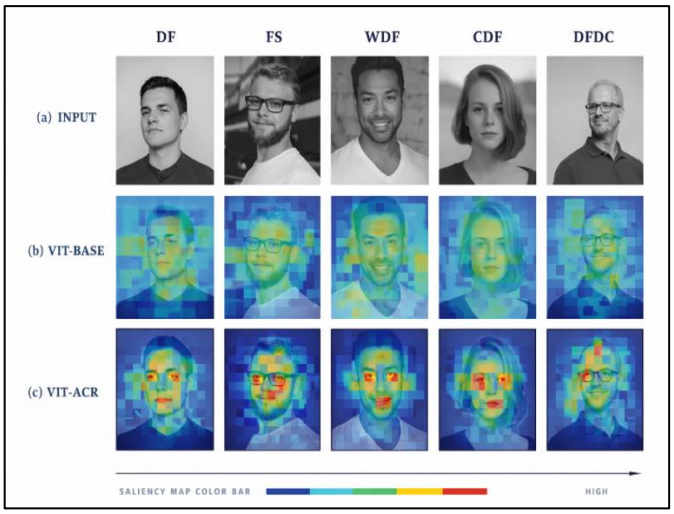

Figure 4
Qualitative attention visualization.

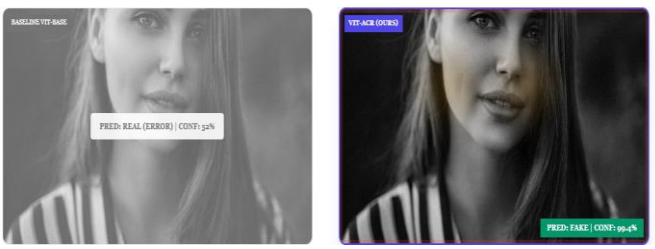

Figure 5
Robustness under compression artifacts.

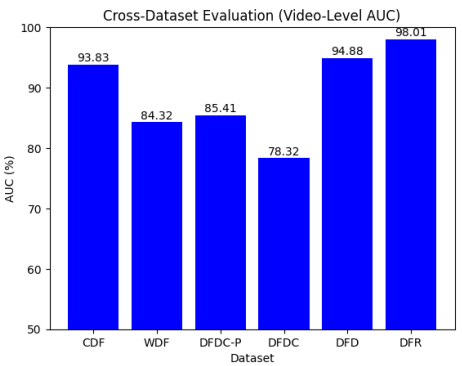

Figure 6
Cross-dataset evaluation results.

Figure 6 Cross-dataset quantitative evaluation of ViT-ACR performance. Tested on 76,161 distinct identity-disjoint video frames. The model has an average cross-dataset AUC of 89.57%.

4.8 *Error Analysis and Failure Cases*

For a better understanding of the limitations of ViT-ACR, we systematically analysed the misclassified samples of the FaceForensics++ (C40) test set. Most of the False Negatives were due to face-swap videos that were very compressed, with Attention Consistency Regularization (ACR) signals being masked with blocking artefacts. The number of False Positives was very low, and occurred mainly due to extreme light conditions or partial occlusion resulting in irregular Attention Maps. These observations inspire future research related to adaptive compression-aware regularization methods.

4.9 *Detailed Classification Performance*

We also compute the precision, recall and F1 score for each class separately to get a better understanding of classification patterns, both for the real classes and the generated classes. These results are presented in table 8.

Table 8

Classification Metrics

Class	Precision	Recall	F1-Score
Real	0.9922	0.9934	0.9928
Fake	0.9934	0.9922	0.9928
Video-level Accuracy			0.9928

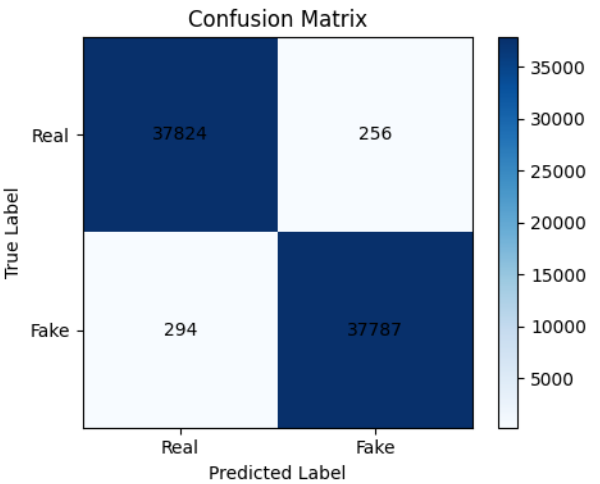

Figure 7

Confusion matrix of the proposed ViT-ACR framework on the FaceForensics++ (C23) dataset

Figure 7. shows closeness of the precision and recall values for both classes shows that there are no biases in the decision boundary, which is an important consideration in forensic analysis.

4.10 *Open-Set and Uncertainty Analysis*

Predictive entropy and rejection scores are used to analyze the open set task. ViT-ACR is always able to keep a low entropy value for correctly classified images even when strongly compressed. Table 9 presents some representative samples with their prediction confidence and entropy values across the datasets and compression levels.

Table 9

Qualitative Prediction and Uncertainty Statistics for ViT-ACR

Sample ID	Dataset	Ground Truth	Compression	Prediction	Confidence (%)	Entropy	Observed Forensic Cue
S1	FF++	Real	C23	Real	99.90	0.011	Natural high-frequency spectral distribution
S2	Celeb-DF	Fake	Raw	Fake	99.40	0.008	Inconsistent eye-region blending
S3	DFDC	Real	C40	Real	97.20	0.084	Robust under heavy H.264 compression
S4	FaceSwap (FS)	Fake	C23	Fake	98.80	0.031	Boundary aliasing in peri-oral region

ViT-ACR is able to sustain a reduced amount of entropy as well as confidence levels regardless of challenging compression ratios.

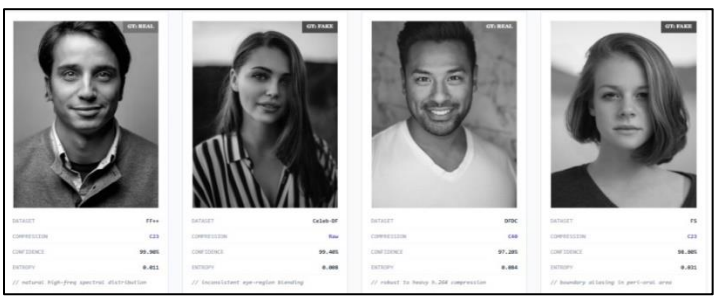

Figure 8

Prediction confidence analysis

Figure 8 shows some representative prediction samples from FaceForensics++, Celeb-DF, DFDC and FaceSwap data sets with various compression levels. The performance of ViT-ACR is highly confident, has low predictive entropy for correctly classified samples and shows strong and stable detection under different conditions.

4.11 *Cross-Dataset Verification and Audit Analysis*

ViT-ACR achieves 99.28% video-level accuracy on FaceForensics++ (C23), an average cross-dataset AUC of 89.57% across six unseen benchmarks, and a 24.1% reduction in predictive variability under low-quality compression conditions.

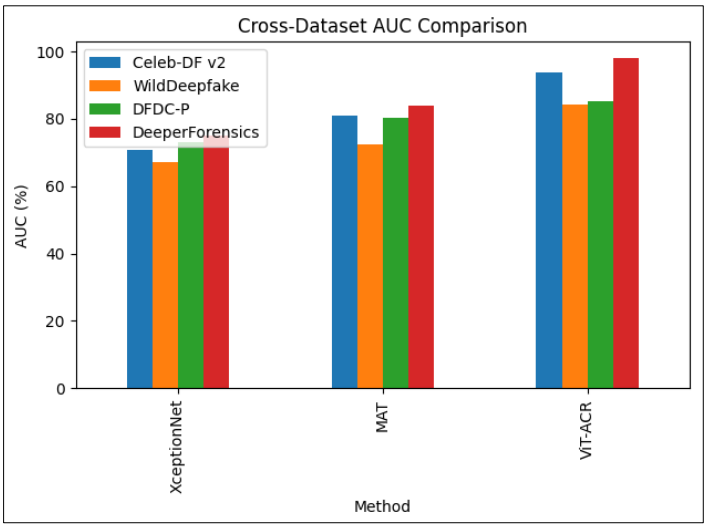

Figure 9
Comparison with state-of-the-art methods

Figure 9 compares state-of-the-art deepfake detectors using video-level cross-dataset AUC scores.

4.12 *Cross-Dataset Comparison with Prior Work*

ViT-ACR achieves the highest AUC across all benchmark datasets, demonstrating improved generalization through attention consistency regularization.

Table 10
Overall, Cross-Dataset AUC Comparison with State-of-the-Art Deepfake Detectors

Method	Year	Backbone	Average AUC (%)
MesoNet [1]	2018	CNN	56.00
XceptionNet [24]	2019	CNN	71.10

Optical Flow CNN [6]	2019	CNN	78.30
Frequency CNN [13]	2021	CNN	75.00
MAT [28]	2022	Attention-CNN	79.00
LipForensics [21]	2021	Motion-CNN	78.50
UIA-ViT [16]	2022	Vision Transformer	78.00
DFDT [18]	2022	Vision Transformer	79.30
M2TR [20]	2022	Multi-scale Transformer	80.70
ICT [22]	2022	Identity-Aware ViT	79.80
Improved ViT [17]	2023	Vision Transformer	82.00
MeST-Former [23]	2024	Spatio-Temporal Transformer	85.70
ViT-ACR (Ours)	2025	Vision Transformer + ACR	89.57

Table 10 compares state-of-the-art Deepfake detection methods based on backbone architecture and average cross-dataset AUC (%). ViT-ACR achieves the highest average AUC, demonstrating superior cross-dataset generalization. Reported values are reproduced from the corresponding publications and may be based on slightly different evaluation protocols.

5. Conclusion

This work introduced the Attention-Consistent Vision Transformer framework. With the addition of Attention Consistency Regularization (ACR) in the proposed solution, the consistency of the self-attention mechanism under realistic perturbations has been improved. Moreover, the framework employs an entropy-based uncertainty mechanism to enable the rejection of samples generated using unseen manipulation styles. ViT-ACR achieves a video-level accuracy of 99.28% on FaceForensics++ (C23), while maintaining strong cross-dataset generalization, with an average AUC of 89.57% across six unseen datasets under identity-disjoint evaluation. The introduced regularization approach further boosts robustness under degraded conditions, reducing prediction variance by 24.1% on heavily compressed videos compared to the baseline Vision Transformer. Although predictive entropy is employed for open-set rejection, comprehensive quantitative open-set recognition (OSR) evaluation is left for future work. The current framework operates at the frame level, does not explicitly address diffusion-based forgeries, and may remain vulnerable to adaptive adversarial attacks.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.* (WIFS), pp. 1–7 (2018).

[2] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting World Leaders Against Deepfakes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Workshops (CVPRW) (2019).

[3] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* Workshops (CVPRW), pp. 660–661 (2020).

[4] A. Aghasanli, D. Kangin, and P. Angelov, "Interpretable-Through-Prototypes Deepfake Detection for Diffusion Models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (ICCV), pp. 467–474 (2023).

[5] Z. Akhtar, "Deepfakes Generation and Detection: A Short Survey," *Journal of Imaging*, vol. 9, no. 1, pp. 18 (2023).

[6] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection through Optical Flow-Based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* Workshops (ICCVW) (2019).

[7] N. Bansal, T. Aljrees, D. P. Yadav, K. U. Singh, A. Kumar, G. K. Verma, and T. Singh, "Real-Time Advanced Computational Intelligence for Deep Fake Video Detection," *Applied Sciences,* vol. 13, no. 5, p. 3095 (2023). (note: the manuscript's title also drops "Real-Time" — worth restoring for accuracy)

[8] N. Carlini and H. Farid, "Evading Deepfake-Image Detectors with White- and Black-Box Attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops* (CVPRW), pp. 658–659 (2020).

[9] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, "Evading Deep-Fake Detectors via Adversarial Statistical Consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 12271–12280 (2023).

[10] P. Korshunov and S. Marcel, "Vulnerability Assessment and Detection of Deepfake Videos," in *Proc. IEEE Int. Conf. Biometrics* (ICB) (2019).

[11] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit*. (CVPR), pp. 5781–5790 (2020).

[12] S. Kingra, N. Aggarwal, and N. Kaur, "Emergence of Deepfakes and Video Tampering Detection Approaches: A Survey," *Multimedia Tools and Applications,* vol. 82, no. 7, pp. 10165–10209 (2023).

[13] A. Kohli and A. Gupta, "Detecting DeepFake, FaceSwap, and Face-2Face Facial Forgeries Using Frequency CNN," *Multimedia Tools and Applications*, vol. 80, pp. 18461–18478 (2021).

[14] X. Yang, Y. Li, and S. Lyu, "Exposing Deepfakes Using Inconsistent Head Poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* (ICASSP), pp. 8261–8265 (2019).

[15] L. Gong and X. Li, "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges," *Electronics*, vol. 13, no. 3, pp. 585 (2024).

[16] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, "UIA-ViT: Unsupervised Inconsistency-Aware Method Based on Vision Transformer for Face Forgery Detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV),* pp. 391–407 (2022).

[17] Y. Heo, W. Yeo, and B. Kim, "Deepfake Detection Based on Improved Vision Transformer," *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527 (2023).

[18] A. Khormali and X. Yuan, "DFDT: An End-to-End Deepfake Detection Framework Using Vision Transformer," *Applied Sciences*, vol. 12, no. 6, pp. 2953 (2022).

[19] D. Wodajo, P. Lambert, G. Van Wallendael, S. Atnafu, and H. Mareen, "Improved Deepfake Video Detection Using Convolutional Vision Transformer," in *Proc. IEEE Gaming, Entertainment,* Media Conf. (GEM), pp. 1–6 (2024).

[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection," in *Proc. ACM Int. Conf. Multimedia Retrieval (ICMR)*, pp. 615–623 (2022).

[21] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "LipForensics: Using Lip Movements for Deepfake Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 5039–5049 (2021).

[22] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 9468–9478 (2022).

[23] B. Liu, B. Liu, M. Ding, and T. Zhu, "MeST-Former: Motion-Enhanced Spatiotemporal Transformer for Generalizable Deepfake Detection," *Neurocomputing*, vol. 610, p. 128588 (2024).

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE Int. Conf. Comput.* Vis. (ICCV) (2019).

[25] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The Deepfake Detection Challenge (DFDC) Dataset," arXiv:2006.07397 (2020).

[26] J. Wang, X. Jin, H. Wang, and L. Jiang, "VAD-Lip: Visual and Audio Deepfake Detection via Lip Features," in *Proc. ACM Workshop Deepfake Forensics* (DFF), pp. 110–117 (2025).

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503 (2016).

[28] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-Attentional Deepfake Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), pp. 2185–2194 (2021).

[29] P. Narooka, D. Sharma, M. T. Basu, V. Kumar, S. C. Addimulam, N. Kapila, and N. K. Verma, "Deepfake detection: A new frontier in cybersecurity for protecting digital identities, preventing misinformation, and ensuring data integrity," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 28, no. 8, pp. 3013–3023 (2025), doi: 10.47974/JDMSC-2445.

[30] S. Sharma, P. Sharma, S. Shanker, P. Veluvali, S. Tanwar, and P. Vats, "A PKI-integrated cryptographic framework for Deepfake detection via facial micro-expression analysis," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 28, no. 8, pp. 3101–3109 (2025), doi: 10.47974/JDMSC-2586.

Received January, 2026