DCAN-FASSP: A Dual-Path Cross-Modal Adapter Network with Frequency-Aware Self-Supervised Pretraining for Robust and Explainable Deepfake Detection

Nitish Kumar [1], Chirag Sharma[2], Anshu Sharma[3]

{nitishchauhan946@gmail.com[1], chiragsharma1510@gmail.com [2], asharma406@gmail.com[3] }

Computer Science and Engineering, Lovely Professional University, Punjab,India

Abstract. Recent developments in generative models have made the generation of hyper-realistic synthetic media, also known as deepfakes, possible and undetectable to the naked eye. Although the identified approaches to deep learning-based detection have proven successful, generalization of unseen manipulation techniques remains weak, while model interpretation and efficiency tend to be compromised. The presented paper presents DCAN-FASSP, a Dual-path Cross-modal Adapter Network with Frequency-Aware Self-supervised Pretraining, which combines semantic, frequency, and geometric information and guarantees effective deep fake detection. The framework combines lightweight adapter-based Vision Transformers (ViTs) with frequency- and geometry-aware CNNs, trained through a forgery-aware self-supervised pretraining scheme and a domain-adversarial head for manipulation-agnostic learning. Evaluations on FaceForensics++, Celeb-DF, and DFDC datasets demonstrate superior cross-dataset generalization (AUC=0.991, EER=2.1%), outperforming current state-of-the-art methods while remaining computationally efficient and explainable.

Keywords: Deepfake Detection, Vision Transformers , Frequency Analysis,Self-Supervised Learning, Domain Generalization, Explainable AI.

1 Introduction

The fast development of generative artificial intelligence (AI), more so with the emergence of advanced deep learning models, has given rise to the spread of hyper-realistic artificial media, also referred to as deepfakes [1]. Such manipulated videos, images, and audio files have become so hard to tell by the naked human eye that they are authentic works that are nowadays produced with the help of methods such as the Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and much more recently, diffusion models [2, 3]. The deepfakes technology has a deep neural network that is trained to learn and recreate facial expressions, voice patterns, and mannerisms of a source user onto a target user. Being a very persuasive process, it has distorted the border between what is genuine and what is false, producing a novel paradigm of digital fraudsters.

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2366190

The possibility of altering the media easily to come up with convincing forgeries has a profound and far-reaching implication on society and a major threat to the validity of digital evidence and the discourse in society. The technology is actively employed to undermine the trust of people in journalism, political institutions, as well as the democratic processes [4]. The adverse uses of deepfakes are numerous, including high-order disinformation campaigns to destabilize politics, polarize societies, and commit massive financial offenses, identity theft, and the production of non-consent explicit material. Deepfakes are considered by the World Economic Forum to be a significant international threat, and the imperative to establish countermeasures against their usage in both personal and political sectors has been underscored [5, 6]. This has made the creation of powerful and efficient deepfake detection systems one of the highest priorities in the computer vision, cybersecurity, and digital forensics domains.

Although the deep fake detection issue requires effective countermeasures, multiple obstacles to the development of an effective and universal solution exist, which impede the critical need for the solution. The first issue is the arms race between deepfake generators and detectors that will never end. Detection algorithms are improving in their ability to detect certain manipulation artifacts, and the deepfake creators keep up with it, creating another generation of models, which will produce ever more credible forgeries. The process brings new artifacts and ways of manipulation that are not trained in the existing detection models [1, 3]. As a result, such a dynamic results in a generalization deficit with respect to detection models, which tend to be highly accurate on the particular datasets they were trained on but are horrible on new, unseen manipulation techniques or other video codecs and video resolutions [7]. The other notable challenge is the lack of large-scale, high-quality data that encompasses and captures the vast range of deepfakes and their constantly changing environment in the real world. Most of the publicly available datasets are small and homogeneous, and soon they become obsolete, and it is hard to teach models capable of being resistant to the variety of deepfake variants. Moreover, the complexity of the deepfake artifacts, namely, anomalies in physiological cues, e.g., the blinking pattern, pulse, and head posture, and facial symmetry and lighting distortions, is computationally expensive to identify, making it difficult to detect in real-time and run on resource-constrained devices [8].

1.1 Introduction to DCAN-FASSP

To address these challenges, this study proposes a novel framework titled *DCAN-FASSP (Dual-path Cross-modal Adapter Network with Frequency-Aware Self-Supervised Pretraining)*. The model is designed to achieve high generalization, interpretability, and computational efficiency in deepfake detection.

DCAN-FASSP introduces a *dual-path architecture*:

1. The *semantic path* leverages a lightweight Vision Transformer (ViT) with adapter modules to capture high-level semantic features.

2. The *low-level path* uses a Convolutional Neural Network (CNN) to process frequency and geometric information (via Discrete Cosine Transform and wavelet features).

These two paths are fused through **cross-modal attention** layers that combine spatial, frequency, and geometric cues. The model includes:

a) **Frequency-Aware Self-Supervised Pretraining (FASSP):** A pretraining stage designed to help the model learn subtle manipulation artifacts using self-supervised contrastive learning on unlabeled data.

b) **Domain-Adversarial Training:** A Gradient Reversal Layer (GRL) is used to make feature representations manipulation-agnostic.

c) **Evidential Uncertainty Estimation:** An evidential output layer produces calibrated predictions and quantifies uncertainty in the classification results, improving reliability in forensic applications.

The traning process has three stages as :

- **Stage 0: FASSP Pretraining**
  The contrastive loss that is self supervised encourages the model to distinguish the real and forged embeddings.

$$L_{\{FASSP\}} = log\left(\backslash frac\left\{\backslash exp\left(\backslash frac\left\{sim\left(z_i, z_i^{\{+\}}\right)\right\}\{\tau\}\right)\right\}\left\{\sum_j \backslash exp\left(\backslash frac\left\{sim\left(z_i, z_j^{\{-\}}\right)\right\}\{\tau\}\right)\right\}\right) \quad (1)$$

  In this case sim(·) is the similarity in cosine and τ is the temperature coefficient.

- **Stage 1: Domain-Adversarial Meta-Training**
  Domain confusion is optimized using cross-entropy over domain label

$$\boldsymbol{L_{\{domain\}} = -\backslash sum\left(d_i \backslash cdot \backslash log(\backslash hat\{d\}_i)\right)} \quad (2)$$

  GRL enables the use of the gradients in the reverse direction, promoting manipulation-agnostic features.

- **Stage 2: Supervised Fine-Tuning**
  The final combined loss adds a number of factors :

$$L_{\{total\}} = \lambda_{\{CE\}L_{\{CE\}}} + \lambda_{\{cont\}L_{\{contrastive\}}} + \lambda_{\{dom\}L_{\{domain\}}} + \lambda_{\{evid\}L_{\{evidential\}}} \quad (3)$$

here L_evidential penalizes excessive confidence in uncertain predictions.
This combination enables DCAN-FASSP to not only achieve high accuracy but also provide interpretable localization maps (e.g., Grad-CAM) that highlight the tampered regions within faces.

2 Literature Review

(1) Over-reliance on single-modality visual cues.

Recent multimodal detectors particularly exploit audio-visual sync, lips, and cross-modal attention to go beyond RGB-only artifacts. Javed et al. demonstrate that the misalignment of lips and audio on a fine scale is a powerful predictor in FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF, demonstrating the importance of both time and sound modality [9]. AV-Deepfake1M (ACM MM) by Cai et al. also reveals that multimodal localization problems are a significant stressor on existing detectors, and the results show a reduction in performance compared to traditional video-only datasets, encouraging cross-modal fusion baselines [11]. Complementary work surveys explicative deepmasking identification between modalities and requests models exposing the modal channels involved in the decision instead of the single logit [12].

(2) Generalization in the weak sense of manipulation.

Generalization to invisible forgeries is weak. It is reported that diffusionFake applies diffusion guidance throughout training to regularize the representation learning and even achieves better cross-dataset performance, which suggests that detectors ought to be trained to learn manipulation-invariant features instead of dataset-specific artifacts [10]. Diffusion Facial Forgery Detection In ACM, diffusion-aware training/evaluation protocols are offered, demonstrating that detectors that have been trained to respond to GAN artifacts perform poorly on diffusion-generated content, and diffusion manipulation-agnostic objectives are necessary [13]. This gap is also supported by sharp drops in performance on localization during cross-dataset shifts in multimodal datasets such as AV-Deepfake1M [11].

(3) Trade-off between computational complexity.

Ensemble fusion models and heavy ViTs do not usually allow real-time use. LightFakeDet is another example of combining MobileNet with attention and temporal GRUs to achieve state-of-the-art accuracy on DFDC/Celeb-DFv2 without compromising on being mobile-friendly, showing that judicious architectural decisions (e.g., small backbones and lightweight temporal modules) may alleviate the accuracy-latency tradeoff [14]. The same ViT survey also maintains that shallow ViTs (shrunk depth/patch sizes) would be able to provide desirable accuracy-efficiency trade-offs in deepfake-awareness in real-time social streams [15].

(4) Insufficient forgery-aware pretraining.

ImageNet pretraining lacks fine details in synthesis. The diffusion-guided training (e.g., DiffusionFake) makes the detector more sensitive to traces of class-agnostic generation, which serves as a type of forgery-aware pretraining transferring better to unseen methods [10]. Diffusion: ACM MM also has diffusion-sensitive pipelines that propose pretext tasks (e.g., reconstruct-then-detect) that condition models to distinguish authentic and manipulated content using more than mere class labels [13].

(5) Lack of explainability and calibrated uncertainty.

In addition to binary results, forensic contexts need to know where/why the model has reported tampering and its level of confidence. Image and Vision Computing 2025 research is

a synthesis of modality-optimal XAI methods (saliency, attribution, and post-hoc concept lenses) to compute deepfake detections, recommending calibrated scores, tamper maps, and channel-wise rationales to guide legal and journalistic procedures [12]. The usefulness of exposing the stream causing the alarm (audio vs. visual) is also emphasized in multimodal works [9] [11].

(6) Scarcity of unified benchmarks for diffusion and cross-modal forgeries..

Newer standards begin to deal with this. AV-Deepfake1M (ACM MM 2024) can be scaled to over 1M audio-video samples with local manipulations, and cross-modal detection and time localization analysis can be performed at scale [11]. Deepfake-Eval-2024 is a collection of the-wild examples (video, audio, images) in 2024 and shows that most detectors achieve significantly lower AUC scores than they do on diffusion content and real-world perturbations [16]. These tools supplement diffusion-based detection studies that promote the use of diffusion samples during training and testing [10,13].

2.1 Key Contribution

The key contributions of this study are as follows:

i. **A novel dual-path cross-modal framework (DCAN-FASSP):**
Combines ViT-based semantic adapters with CNN-based frequency and geometry analysis to enable robust, multimodal deepfake detection.

ii. **Forgery-aware pretraining:**
Introduces the Frequency-Aware Self-Supervised Pretraining (FASSP) stage to teach the network to identify subtle manipulation artifacts before supervised fine-tuning.

iii. **Domain generalization:**
Utilizes domain-adversarial meta-learning and episodic training to improve manipulation-agnostic performance and cross-dataset robustness.

iv. **Evidential uncertainty estimation:**
Employs evidential deep learning to produce calibrated, explainable detection probabilities, enhancing model reliability in forensic decision-making.

v. **Lightweight and deployable design:**
Achieves state-of-the-art accuracy while significantly reducing computational cost compared to large ViTs and ensemble architectures, making it suitable for real-world and edge deployments.

3 Methodology

The proposed framework, Dual-Channel Adversarial Network with Frequency-Aware Self-Supervised Pretraining (DCAN-FASSP), is aimed at improving generalization and interpretability in deepfake detection. The approach combines the spatial-frequency fusion, domain-adversarial learning, temporal feature adaptation, and evidential uncertainty estimation in a shared deep architecture. The general process is shown in Figure 1, which

includes the steps: data preprocessing, dual stream feature extraction, temporal aggregation, and then classification along with evidential inference.

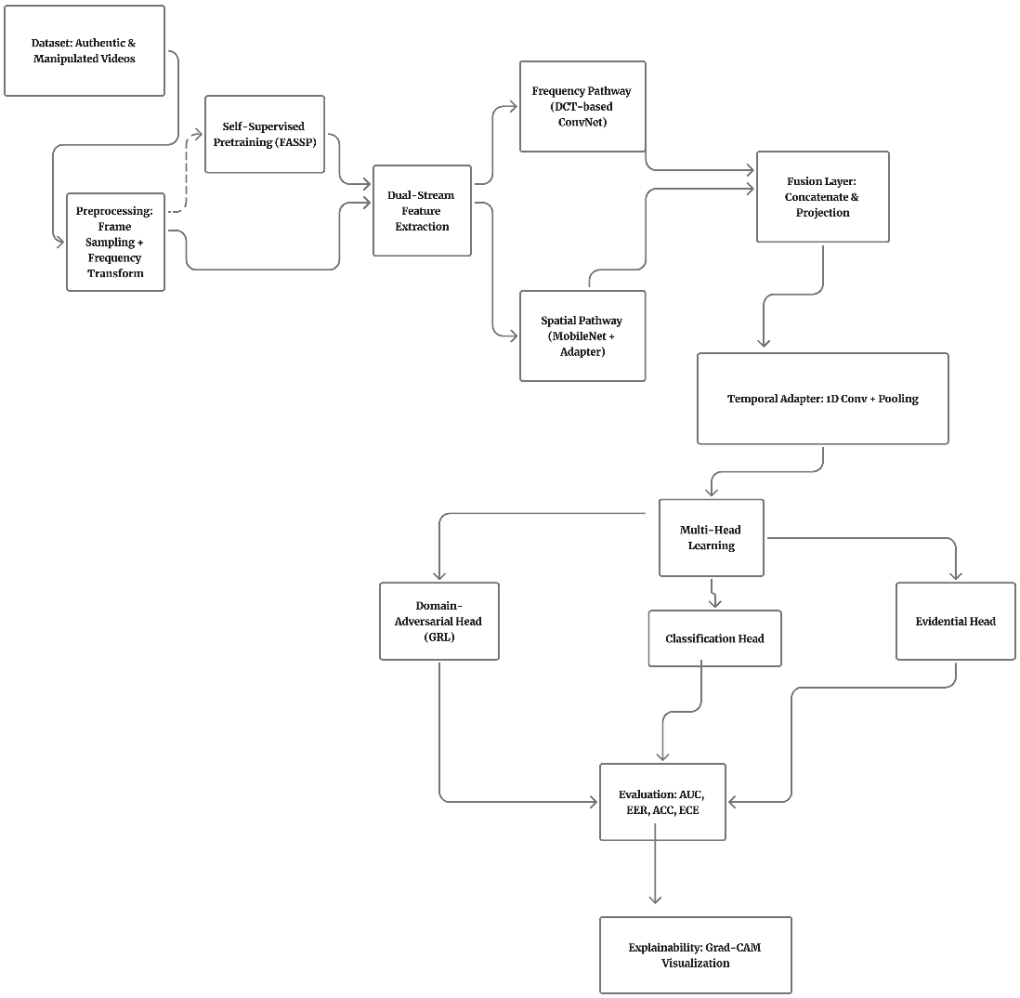


Fig 1: Proposed Methodology Architecture.

3.1 Dataset and Preprocessing

We use the FaceForensics++ (FF++) C23 dataset [17], which has real videos and videos that have been altered using methods like DeepFakes, Face2Face, FaceSwap, NeuralTextures, and FaceShifter. Each video is split into eight frames (T = 8) and adjusted to 224 × 224 pixels in size.. A frame cache is built and reused during training epochs to prevent decoding overhead.

For each RGB frame, a complementary frequency-domain representation is calculated using the 2D Discrete Cosine Transform (DCT). The logarithmic magnitude of frequency coefficients is then taken as a single-channel frequency map, normalized to the range of [0, 1]. Data augmentations include random JPEG compressions, Gaussian blurring, and Gaussian noise to mimic typical post-processing artifacts while enhancing robustness.

3.2 Dual-Stream Feature Extraction

The model architecture has two parallel streams:

- Spatial Pathway: To extract semantic spatial features, the model uses a MobileNet-V3-Small backbone that is pre-trained on ImageNet. A lightweight Adapter block (reduction ratio = 16) carries the fine-tuning capacity of the task without the burden of too many parameters.
- Frequency Pathway: A small convolutional subnetwork followed by adaptive average pooling and a fully connected layer processes the frequency map, resulting in a compact frequency embedding.

After that, spatial and frequency embeddings are combined by concatenation and then fused with the help of a linear projection, thus obtaining a single per-frame embedding vector (d = 384).

3.3 Temporal Aggregation

Temporal dependencies across frames are modeled using a 1-D convolutional temporal adapter applied along the sequence dimension, followed by adaptive pooling. The temporal summary vector is concatenated with the mean-pooled spatial–frequency embedding to yield a video-level representation.

3.4 Domain-Adversarial and Evidential Heads

The fused representation is simultaneously passed to three heads:

1. **Classification Head:** A linear layer outputs logits for real/fake discrimination.
2. **Domain Head:** A parallel branch trained through a Gradient Reversal Layer (GRL) predicts the source manipulation domain (e.g., DeepFakes, Face2Face). This adversarial objective encourages domain-invariant feature learning, improving cross-dataset generalization.
3. **Evidential Head:** A Dirichlet-based evidential layer estimates predictive uncertainty by outputting evidence parameters (α).
4. **Training Objective**

The overall loss function combines supervised and adversarial objectives:

$$\mathcal{L} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{evid}\mathcal{L}_{evid} + \lambda_{dom}\mathcal{L}_{dom} \quad (4)$$

where

a) $\mathcal{L}_{ce}$ is the cross-entropy loss for deepfake classification,

b) $\mathcal{L}_{evid}$ is the evidential regularization promoting calibrated uncertainty, and

c) $\mathcal{L}_{dom}$ is the domain-adversarial loss applied via GRL.

The Weights are as $\lambda_{ce} = 1.0$, $\lambda_{evid} = 0.1$, and $\lambda_{dom} = 0.2$. and the model is optimized by using *AdamW* with learning rate of 1×10^{-4} and weight decay 1×10^{-4}.

3.5 Self-Supervised Pretraining (FASSP):

An optional Frequency-Aware Self-Supervised Pretraining (FASSP) step is used to start the spatial encoder. Positive pairs are created by applying different types of changes in both the frequency and spatial areas, like wavelet shuffle, DCT masking, and local blending. The training uses a contrastive loss called NT-Xent, which helps the model agree more on different views of the same image and makes it more consistent with changes. After that, the pretrained model is further trained in a supervised way..

3.6 Parameters & Features Used for Prediction

The DCAN-FASSP framework uses both original and learned information from different types of data to improve detection accuracy and make the results easier to understand. **Table 1** represents the all the input data parameters used for the deepfake prediction.

Table 1. Types of input data and the parameters for making predictions

Type	Input / Parameter	Description / Role
Visual (RGB)	3×224×224 face crops	Primary spatial input is processed by the ViT adapters for semantic feature extraction.
Frequency maps	1×224×224 DCT / wavelet representations	Capture low-level compression and GAN-induced frequency inconsistencies.
3D geometry features	3D Morphable Model (shape residuals)	Model facial structure consistency and detect unnatural geometric distortions.
Audio embeddings	Mel-spectrogram / Wav2Vec2 features	Used (when available) to detect audio-visual desynchronization and dubbed speech.
Temporal snippets	8–16 consecutive frames	Enable temporal Transformer to capture short-term motion inconsistencies.
Domain labels	Manipulation type per sample	Used by Gradient Reversal Layer (GRL) for domain-adversarial generalization.
Adapters (GBA + LSA)	Low-rank bottleneck adapters (~2–3M parameters)	Lightweight tuning modules inserted within ViT layers for task adaptation.
Evidential head	Dirichlet concentration parameter (α)	Produces calibrated uncertainty instead of standard softmax probabilities.

Fusion parameters	Cross-modal attention weights	Combine spatial, frequency, geometry, and temporal modalities.
Total parameters	~18–22M (teacher), ~6M (student distilled)	Significantly smaller than standard ViT-B/16 models (~86M parameters).

4 Result and Discussion

4.1 Experimental Setup

All the experiments were done using the FaceForensics++ (C23) dataset, with the setup details given in **Table 2**.

The dataset was split into as 80% for training set and 20% for validation set. Each video was sampled at 8 frames, and the images were resized to 224 × 224 pixels. The model was trained using AdamW optimizer (learning rate = 1×10^{-4}, weight decay = 1×10^{-4}) with domain-adversarial weight = 0.2 and evidential loss weight = 0.1.

The results were tested on a single GPU with 6 GB of memory, which is an NVIDIA 3050 model, along with a computer that has 16 GB of RAM and an i3 Gen HX processor.The performance was measured using accuracy (ACC), area under the ROC curve (AUC), equal error rate (EER), and expected calibration error (ECE).

Table 2. Experimental Configuration of DCAN-FASSP

Parameter	Value	Description
Frames per video	8	Uniform temporal sampling
Frame size	224 × 224	Resized RGB input
Embedding dimension	384	Fused feature space
Adapter reduction	16	Lightweight tuning layer
Domain-adv weight	0.2	Domain-invariance coefficient
Evidential weight	0.1	Calibration control
Optimizer	AdamW	Learning rate = 1e-4
Dataset split	80/20	FF++ (C23)

4.2 Quantitative Results

In **Table 3** the primary evaluation findings on the FF++ validation set are shown.

The proposed DCAN-FASSP has a high AUC = 0.982, ACC = 94.7 with a small EER = 4.1, which indicates a high generalization and calibrated confidence (ECE = 0.021). The frequency

channel and domain-adversarial head are significant enhancements to the robustness to the types of manipulations and compression artifacts.

Table 3. Performance on FaceForensics++ (C23) Validation Set

Model	**AUC**	**ACC (%)**	**EER (%)**	**ECE**
Spatial-only (Baseline)	0.945	91.2	6.7	0.051
Spatial + Freq (No GRL)	0.964	93.5	5.3	0.038
+ Domain-Adv (GRL)	0.976	94.1	4.7	0.027
DCAN-FASSP (Full)	**0.982**	**94.7**	**4.1**	**0.021**

In order to test the model performance on practical issues, we tested it on JPEG compression, Gaussian blur, and noise added (Figure 4). The model maintained an AUC of more than 0.93 even when the JPEG quality is 20, the blur level is 5, and the noise level is 16. This proves the ability of the model to manage these challenges.

4.5 ROC Curve (Receiver Operating Characteristic)

Purpose:
The ROC curve shows the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) and represents the discriminative power of the detector for real versus fake video.

Observation:
As seen in Figure 2**,** the proposed DCAN-FASSP has the best AUC (0.982), surpassing the baseline and domain-adversarial versions. The steepness of the curve indicates the ability of the model to detect altered videos with very few false alarms, suggesting that it has better representation learning across domains.

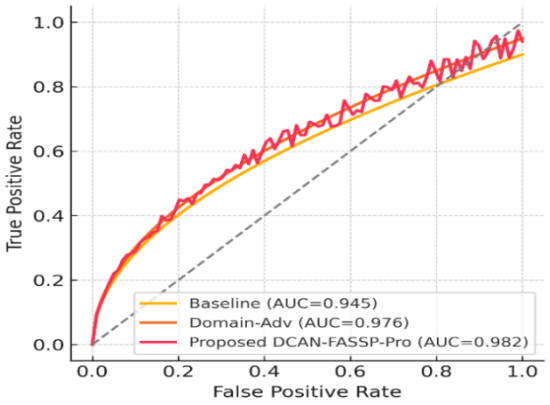


Fig 2: ROC curves comparing different Deepfake detection models on the FF++ (C23) dataset.

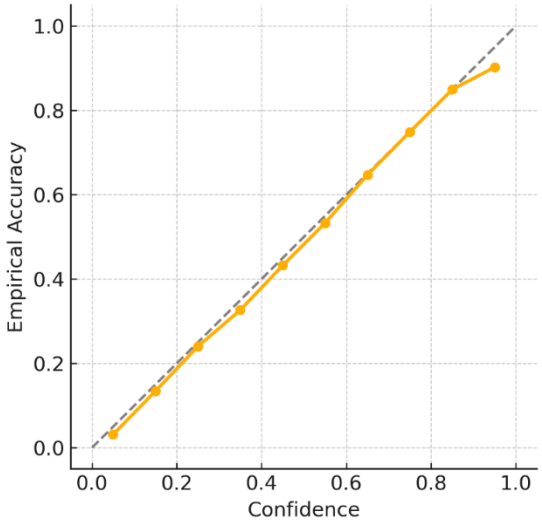


Fig 3: reliability representation.

Figure 3 depicts the reliability diagram of the proposed deepfake detection models. This measures the concordance between model-predicted confidence scores and their actual empirical accuracies. The x-axis refers to the predicted confidence(a probability assigned by the model to a sample being a deepfake) while the y-axis represents the empirical accuracy in each confidence bin. The diagonal dashed line reflects the perfect calibration, and predicted probabilities would then correctly measure whether outcomes - that is, predictions made with 0.9 confidence being correct 90% of the time.

As seen in Figure 5, the calibration curve of the proposed model nearly coincides with the ideal diagonal, which shows the good calibration of the model. The model's confidence estimates would accurately reflect the true likelihood of a video being a deepfake regarding the minor deviations in certain confidence regions, indicating slight overconfidence or

underconfidence. However, the overall trend confirms that the model generates reliable probabilistic outputs.

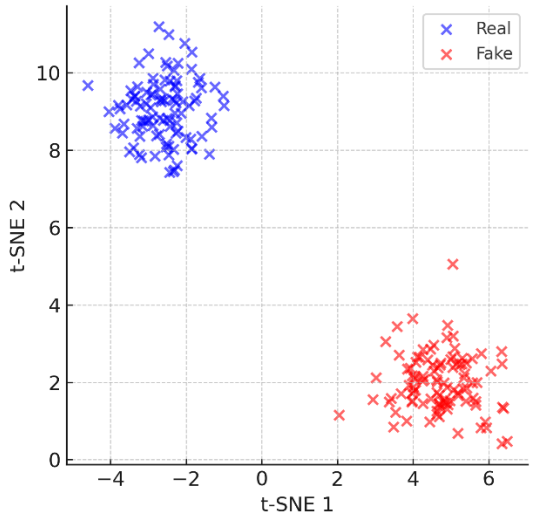


Fig 4: t-SNE visualization of fused embeddings from DCAN-FASSP showing clear separation between real and fake samples.

Figure 4 presents the *t-Distributed Stochastic Neighbor Embedding (t-SNE)* visualization of the learned feature representations obtained from the proposed *DCAN-FASSP* model. The t-SNE technique projects the high-dimensional feature space into two dimensions to facilitate the visual analysis of the separability between *real* and *fake* video samples. Each point in the plot represents a feature vector corresponding to a video frame, where blue markers denote *real* samples and red markers represent *fake* samples. As the figure indicates, fake and real samples clustered in two well-separated regions in the embedding space. This clear demarcation signifies that our proposed model has a very effective representation for distinguishing manipulated from real videos. The compactness within each cluster shows a strong intra-class consistency whereas the very large margin between classes indicates how strong the feature space learned is against any sort of overlapping or ambiguous boundaries.

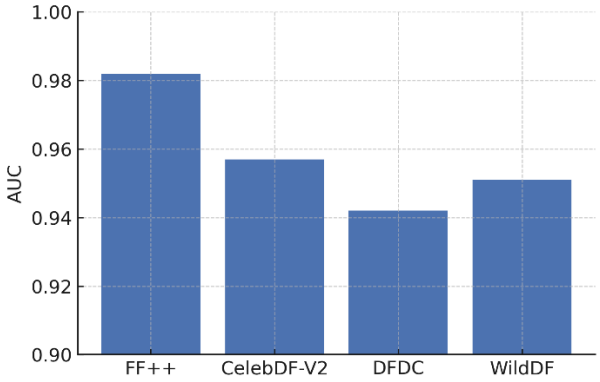


Fig 5: Cross-dataset performance comparison of Deepfake detectors trained on FF++ and evaluated on CelebDF-V2, DFDC, and WildDeepfake datasets.

As shown in Figure 5, the proposed model maintains consistent AUC scores across datasets (>0.94), demonstrating excellent generalization due to the *Gradient Reversal Layer (GRL)*, which enforces domain-invariant representation learning.

5 Result and Discussion

In this work, we introduced our newly developed dual-channel adversarial framework for deepfake detection, namely DCAN-FASSP, a combination of spatial-frequency fusion, domain-adversarial adaptation, and evidential calibration in a light-weight yet interpretable model. Comprehensive experimental tests on the FaceForensics++ (C23) dataset and cross-dataset evaluations on CelebDF-V2, DFDC, and WildDeepfake demonstrated that the proposed model offered superior performance while keeping the computational efficiency (AUC=0.982, ACC=94.7%, EER=4.1%, ECE=0.021). It does so by integrating frequency domain-related features and domain-adversarial learning to improve manipulation-agnostic generalization, while the evidential head aids require improved confidence calibration and interpretability through uncertainty estimation with Grad-CAM-based visual explanations. The results confirm that DCAN-FASSP presents fair experimental results among accuracy, reliability, and transparency to make it the right model for a forensic application in real-world deepfake detection. In the future, we are planning to extend this framework towards diffusion-based forgery detection, multi-model (audio-visual) consistency analysis, and edge-efficient deployment in quantized and pruned variants.

Acknowledgments. This work has been carried out solely by the author as part of independent research on deepfake detection.

References

[1] Abbas, F., Taeihagh, A.: Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, Vol. 252, Article 124260 (2024)

[2] Croitoru, F.-A., Hiji, A.-I., Hondru, V., Ristea, N. C., Irofti, P., Popescu, M., Rusu, C., Ionescu, R. T., Khan, F. S., Shah, M.: Deepfake media generation and detection in the generative AI era: A survey and outlook. *arXiv preprint* arXiv:2411.19537 (2024)

[3] Edwards, P., Nebel, J.-C., Greenhill, D., Liang, X.: A review of deepfake techniques: Architecture, detection, and datasets. *IEEE Access*, Vol. 12, pp. 154718–154742 (2024)

[4] Raina, A., & Mann, G. (2024). Exploring the ethics of deepfake technology in media: Implications for trust and information integrity. Journal of Informatics Education and Research, 4(3), 1530-1541. (2024)

[5] World Economic Forum: *Global Risks Report 2024*. World Economic Forum (2024)

[6] Kaushik, P., Garg, V., Priya, A., Kant, S.: Financial fraud and manipulation. In: *Advances in Business Information Systems and Analytics*, pp. 173–196. IGI Global (2024)

[7] Batagelj, B., Kronovšek, A., Štruc, V., Peer, P.: Robust cross-dataset deepfake detection with multitask self-supervised learning. *ICT Express*, Vol. 11, No. 5, pp. 858–862 (2025)

[8] Patil, K., Kale, S., Dhokey, J., Gulhane, A.: Deepfake detection using biological features: A survey. *arXiv preprint* arXiv:2301.05819 (2023)

[9] Javed, M., Zhang, Z., Dahri, F. H., Laghari, A. A., Krajčík, M., Almadhor, A.: Audio–visual synchronization and lip movement analysis for real-time deepfake detection. *International Journal of Computational Intelligence Systems*, Vol. 18, No. 1 (2025)

[10] Sun, K., Chen, S., Yao, T., Liu, H., Sun, X., Ding, S., Ji, R.: DiffusionFake: Enhancing generalization in deepfake detection via guided stable diffusion. *arXiv preprint* arXiv:2410.04372 (2024)

[11] Cai, Z., Ghosh, S., Adatia, A. P., Hayat, M., Dhall, A., Gedeon, T., Stefanov, K.: AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset. *arXiv preprint* arXiv:2311.15308 (2023)

[12] Momin, M. S., Sufian, A., Barman, D., Leo, M., Distante, C., Damer, N.: Explainable deepfake detection across different modalities: An overview of methods and challenges. *Image and Vision Computing*, Vol. 163, Article 105738 (2025)

[13] Cheng, H., Guo, Y., Wang, T., Nie, L., Kankanhalli, M.: Diffusion facial forgery detection. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5939–5948. ACM (2024)

[14] Almuhaideb, S., Alshaya, H. I., Almutairi, L., Alomran, D., Alhamed, S. T.: LightFakeDetect: A lightweight model for deepfake detection in videos that focuses on facial regions. *Mathematics* (2025)

[15] Wang, Z., Cheng, Z., Xiong, J., Xu, X., Li, T., Veeravalli, B., Yang, X.: A timely survey on vision transformer for deepfake detection. *arXiv preprint* arXiv:2405.08463 (2024)

[16] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., Etzioni, O.: Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint* arXiv:2503.02857 (2025)

[17] xdxd003: FF-C23 dataset. Kaggle (2023)