

A Comprehensive Survey of Attention-Based Deep Learning Techniques for Robust Deepfake Face Detection

Shivani Jatav

M. Tech. Scholar, School of Computer Technology, Sanjeev Agrawal Global Educational
University Bhopal, M.P.

Dr. Shailja Shukla

Professor, School of Computer Technology, Sanjeev Agrawal Global Educational University
Bhopal, M.P.

Mr. Chandan Kumar

Assistant Professor, School of Computer Technology, Sanjeev Agrawal Global Educational
University Bhopal, M.P.

Abstract: The rapid advancement of deep learning and generative artificial intelligence has significantly increased the realism of artificially generated and manipulated facial images, creating substantial challenges for digital media authenticity and security. Fake-face detection has emerged as an important research area for identifying synthetic faces, face-swapping operations, and manipulated facial content. Traditional detection approaches often struggle to identify subtle manipulation artifacts, particularly under variations in image quality, compression, lighting conditions, and facial expressions. This paper presents a comprehensive survey of attention-based deep learning techniques for robust fake-face detection, focusing on recent advancements in convolutional neural networks (CNNs), vision transformers, and hybrid deep learning architectures. The survey examines the role of spatial attention, channel attention, self-attention, and multi-head attention mechanisms in enhancing discriminative feature extraction and identifying subtle inconsistencies in facial regions. Furthermore, it reviews attention-based feature fusion strategies, transfer learning approaches, and multi-scale feature extraction techniques used to improve detection performance. Various benchmark datasets, evaluation metrics, and experimental methodologies are discussed to provide a structured understanding of existing research. The paper also investigates key challenges, including cross-dataset generalization, adversarial manipulation, computational complexity, interpretability, and the detection of highly realistic AI-generated faces. Finally, potential research directions are identified, including complementary attention mechanisms, frequency-domain analysis, lightweight detection architectures, explainable artificial intelligence, and multimodal learning. This survey aims to provide researchers with a comprehensive understanding of existing attention-based detection methods and support the development of accurate, efficient, and generalizable fake-face detection frameworks.

Keywords: Fake-Face Detection, Deepfake Detection, Deep Learning, Attention Mechanism, Convolutional Neural Network, Vision Transformer, Feature Fusion, Generative Artificial Intelligence.

I. INTRODUCTION

The rapid development of Artificial Intelligence (AI), Deep Learning (DL), and Generative Artificial Intelligence (GenAI) has significantly transformed digital image processing and facial image generation. Advanced generative models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, have enabled the creation of highly realistic synthetic facial images. These technologies have various applications in entertainment, digital communication, virtual reality, and computer graphics. However, their misuse for generating fake facial images, manipulating facial identities, and creating deceptive digital content has raised serious concerns regarding digital media authenticity, privacy, and cybersecurity. Fake-face detection has therefore become an important research area focused on distinguishing genuine facial images from artificially generated or manipulated faces [1, 2].

Traditional fake-face detection methods primarily rely on handcrafted features, texture analysis, image inconsistencies, and statistical characteristics. Although these approaches can identify certain manipulation artifacts, their performance may decrease when applied to high-quality synthetic images or images affected by compression, noise, and variations in lighting conditions. Conventional machine learning techniques also depend heavily on manually selected features, which may not adequately represent the complex characteristics of modern AI-generated faces. Consequently, deep learning-based detection techniques have gained considerable attention because of their ability to automatically extract hierarchical features and learn complex patterns from facial images [3].

Convolutional Neural Networks (CNNs) have been widely employed for fake-face detection because of their capability to identify local texture patterns, facial structures, and manipulation artifacts. Architectures such as VGGNet, ResNet, DenseNet, and EfficientNet have demonstrated their applicability in image classification and facial forgery analysis. However, conventional CNN architectures primarily focus on local feature extraction and may not effectively capture long-range dependencies or subtle relationships between different facial regions. In addition, highly realistic synthetic faces may contain fewer visible artifacts, making reliable classification increasingly challenging. These limitations have encouraged researchers to investigate advanced feature extraction mechanisms that can improve the discrimination between genuine and fake facial images [4, 5].

Attention mechanisms have emerged as an important approach for improving the feature representation capabilities of deep learning models. Unlike conventional feature extraction techniques, attention-based architectures can assign different importance levels to spatial regions, feature channels, and contextual relationships within an image. Spatial attention mechanisms emphasize informative facial regions, such as the eyes, nose, mouth, and facial boundaries, where manipulation artifacts may occur. Channel attention mechanisms identify important feature channels, while self-attention mechanisms capture long-range dependencies

between different facial regions. The integration of these attention mechanisms enables deep learning models to focus on discriminative facial characteristics and potentially improve detection performance under challenging imaging conditions [6].

Complementary attention-based approaches combine multiple attention mechanisms to exploit different types of facial information. For example, spatial and channel attention can jointly identify informative image regions and relevant feature representations, while multi-scale attention can capture manipulation artifacts at different spatial resolutions. Similarly, combining CNN-based local feature extraction with transformer-based global feature modeling can improve the representation of both fine-grained facial details and long-range contextual relationships. These complementary strategies provide a promising direction for developing robust fake-face detection frameworks. Nevertheless, their effectiveness depends on the selected architecture, training dataset, attention design, and ability to generalize across different manipulation techniques [7].

Despite significant progress, several challenges remain in the development of reliable fake-face detection systems. One major challenge is cross-dataset generalization, where a model trained on one dataset may experience reduced performance when evaluated on images generated using unseen techniques or collected under different conditions. Other challenges include image compression, variations in image resolution, adversarial manipulation, computational complexity, and limited interpretability. Furthermore, the continuous evolution of generative models introduces new types of synthetic facial images that may not contain the artifacts learned by existing detection systems. These challenges highlight the need for comprehensive investigation of attention-based deep learning techniques and their suitability for practical fake-face detection applications [8, 9].

This paper presents a comprehensive survey of attention-based deep learning techniques for robust fake-face detection. It reviews the fundamental concepts of fake-face generation and detection, examines CNN-based, transformer-based, and hybrid detection architectures, and discusses spatial, channel, self-attention, and complementary attention mechanisms. Furthermore, the survey investigates feature fusion strategies, benchmark datasets, evaluation metrics, and experimental methodologies used in existing research. The major contributions of this review are summarized as follows:

1. To provide a comprehensive overview of fake-face generation techniques and deep learning-based detection methods.
2. To investigate different attention mechanisms and their applications in identifying facial manipulation artifacts.
3. To analyze complementary attention and feature fusion strategies for extracting discriminative facial representations.
4. To review commonly used datasets, evaluation metrics, and experimental approaches for fake-face detection.
5. To identify major research challenges, including cross-dataset generalization, computational efficiency, and detection of unseen synthetic faces.

6. To highlight future research directions for developing robust, explainable, and computationally efficient attention-based detection frameworks.

1.1 Research Novelty and Distinction from Existing Surveys

Although several studies have investigated deepfake detection using convolutional neural networks, vision transformers, and attention-based architectures, a systematic understanding of how complementary attention mechanisms contribute to robust fake-face detection remains important. Existing research often emphasizes specific architectures, individual attention mechanisms, or particular detection challenges. In contrast, this survey adopts a unified analytical perspective to examine the complementary roles of spatial, channel, frequency-domain, and global attention in deepfake face detection.

The primary novelty of this survey lies in its structured comparison of attention mechanisms according to their feature-learning capabilities, target information, and potential limitations. Spatial attention is examined for identifying localized facial manipulation artifacts, channel attention for emphasizing informative feature representations, frequency attention for analyzing subtle frequency-domain inconsistencies, and global attention for capturing long-range dependencies between facial regions. The survey further discusses how these mechanisms can be integrated through complementary feature fusion to combine local, frequency-based, and global facial information within a unified detection framework.

Unlike a general overview of deepfake detection techniques, this survey connects the comparative analysis of existing CNN-based, transformer-based, and hybrid architectures with key practical challenges, including cross-dataset generalization, unseen manipulation detection, computational efficiency, interpretability, and fairness. It also presents a conceptual complementary-attention framework that illustrates how different attention branches can contribute to a common classification process.

Therefore, the contribution of this survey is not the introduction of a newly trained detection model or a claim of experimentally proven performance improvement. Rather, it provides a focused comparative synthesis of attention mechanisms, their complementary functions, their limitations, and the research opportunities associated with their integration. This analysis aims to help researchers identify suitable attention strategies and develop more robust, interpretable, and computationally efficient deepfake detection systems.

II. LITERATURE REVIEW

Sheng et al. [1] proposed an ID-insensitive deepfake detection model based on a multi-attention mechanism to improve the reliability of facial forgery detection. The study focuses on reducing the influence of identity-related facial characteristics while identifying manipulation-related features. The multi-attention mechanism enables the model to emphasize informative facial regions and extract discriminative representations for distinguishing genuine and manipulated faces. By reducing dependence on identity-specific information, the approach aims to improve the generalization capability of deepfake detection across different individuals. This research highlights the importance of attention-based feature extraction for

developing robust detection systems. However, the effectiveness of such approaches may depend on the diversity of training datasets and the quality of facial manipulation samples.

Guo et al. [2] introduced a generalizable deepfake detection framework using Spatial Kernel Selection and a Halo Attention Network. The proposed approach focuses on improving the extraction of spatial features and capturing relationships between different facial regions. Spatial Kernel Selection supports the identification of informative local patterns, while Halo Attention facilitates contextual feature learning beyond individual image regions. The integration of these mechanisms aims to improve the detection of subtle facial manipulation artifacts and enhance generalization across different deepfake generation methods. The study emphasizes the importance of combining local feature extraction with broader contextual information. This approach provides useful insights into the development of attention-based architectures for robust fake-face detection, particularly when manipulated images contain complex spatial inconsistencies.

Magesh et al. [3] developed an efficient deepfake detection system by integrating the recognition capabilities of Convolutional Neural Networks (CNNs) and Transformer architectures. CNNs are effective in extracting local facial features, including texture patterns and manipulation artifacts, whereas Transformers can model long-range dependencies and contextual relationships. The integration of these architectures provides a mechanism for learning both local and global facial representations. The study focuses on utilizing the complementary characteristics of CNNs and Transformers to improve the identification of manipulated facial content. This hybrid approach is relevant to fake-face detection because synthetic facial images may contain both localized artifacts and inconsistencies distributed across different facial regions. The research also highlights the importance of computational efficiency in practical deepfake detection applications.

Zheng [4] proposed an attention-based framework for detecting face forgeries by integrating Efficient-ViT and wavelet transform techniques. The approach combines transformer-based feature learning with frequency-domain analysis to identify potentially informative characteristics of manipulated facial images. Efficient-ViT supports the extraction of contextual information while emphasizing computational efficiency, whereas wavelet transform provides a mechanism for analyzing image information at different frequency scales. The integration of these techniques aims to capture both spatial characteristics and subtle frequency-related artifacts associated with facial manipulation. Attention-based feature processing further supports the identification of relevant facial information. This study demonstrates the potential of combining efficient transformer architectures with frequency-domain analysis for developing effective face-forgery detection systems.

Clapten and Balaji [5] introduced a gated temporal attention-based framework for robust deepfake video detection. Unlike image-based approaches that analyze individual facial frames, the proposed framework focuses on temporal information and relationships between successive video frames. The gated temporal attention mechanism is designed to emphasize relevant temporal features while controlling the contribution of information extracted from different frames. This approach is important because manipulated videos may contain

inconsistencies in facial movements, expressions, and temporal transitions that cannot always be identified through spatial analysis alone. By incorporating temporal attention, the framework aims to improve the identification of manipulated video content. The study highlights the significance of temporal feature modeling in deepfake detection, particularly for video-based applications.

Shivaprakash et al. [6] presented an attention-augmented hybrid framework incorporating evolutionary optimization for robust deepfake detection. The proposed approach combines attention-based feature extraction with hybrid deep learning techniques to improve the representation of facial manipulation characteristics. Attention mechanisms help identify informative features, while evolutionary optimization is incorporated to support the selection or optimization of relevant model components and parameters. The integration of these techniques aims to improve the detection process by reducing the influence of less informative features and enhancing the learning of discriminative representations. This research is relevant to the development of optimized attention-based detection architectures. However, the practical effectiveness of evolutionary optimization also depends on its computational cost, parameter configuration, and ability to improve generalization across different datasets.

Kumar et al. [7] proposed a hybrid spatial-frequency attention-based algorithm using EfficientNet for robust and interpretable deepfake detection. The framework combines spatial information with frequency-domain characteristics to identify facial manipulation artifacts that may not be sufficiently represented through conventional image analysis. EfficientNet provides an efficient CNN-based feature extraction architecture, while spatial-frequency attention supports the identification of informative patterns from different feature domains. The integration of spatial and frequency information aims to improve the discrimination between genuine and manipulated facial images. The study also emphasizes interpretability, which is important for understanding the features influencing the model's predictions. This approach highlights the potential of combining efficient deep learning architectures with complementary attention mechanisms for reliable fake-face detection.

Javed et al. [8] introduced a real-time deepfake detection framework using an efficient Swin Attention Network with global and local facial features. The approach focuses on combining local facial information with global contextual representations to improve the detection of manipulated content. The Swin Transformer architecture supports hierarchical feature extraction and attention-based contextual modeling, while local facial features provide detailed information about specific facial regions. The integration of global and local representations aims to identify subtle inconsistencies that may occur at different spatial scales. The framework also emphasizes real-time detection, making computational efficiency an important consideration. This research demonstrates the relevance of hierarchical attention mechanisms for developing practical deepfake detection systems capable of processing facial information efficiently.

Gupta et al. [9] proposed a hybrid attention-based spatiotemporal model for deepfake video detection. The framework focuses on extracting spatial information from individual video frames and temporal relationships across successive frames. Spatial feature extraction supports

the identification of facial texture irregularities and local manipulation artifacts, while temporal modeling helps capture inconsistencies in facial movements and video sequences. The attention mechanism is intended to emphasize informative spatial and temporal features during the detection process. By combining these complementary representations, the model aims to improve the identification of manipulated video content. This study highlights the importance of integrating spatial and temporal analysis for deepfake detection. Its methodology is particularly relevant to applications where frame-level analysis alone may not capture the full characteristics of manipulated videos.

Reza et al. [10] introduced an attention and residual mechanism-based CNN architecture, referred to as ARC-Net, for deepfake facial image detection with an emphasis on fairness and generalization. The proposed architecture combines attention mechanisms with residual feature learning to extract informative facial representations while maintaining effective information flow through the network. Attention-based processing supports the identification of discriminative facial regions, whereas residual connections facilitate deeper feature extraction. The study also considers fairness and generalization, which are important challenges in facial image analysis because detection performance may vary across demographic groups and image characteristics. The proposed framework aims to address these concerns while distinguishing genuine facial images from manipulated ones. This research emphasizes that robust fake-face detection should consider not only classification performance but also fairness and generalization across diverse facial populations.

Table 1. Comparative Analysis of Existing Studies

Author (Year)	Model / Technique	Attention Mechanism	Key Contribution	Research Limitation
Sheng et al. (2025)	ID-insensitive deepfake detection	Multi-attention	Reduces dependence on identity-specific facial features.	Generalization across diverse datasets requires further evaluation.
Guo et al. (2025)	Spatial Kernel Selection and Halo Attention Network	Halo attention	Combines spatial feature selection with contextual feature learning.	Computational complexity and unseen manipulation detection need assessment.
Magesh et al. (2025)	CNN–Transformer hybrid	Transformer attention	Combines local facial features with global contextual information.	Model efficiency and cross-dataset performance require evaluation.
Zheng (2025)	Efficient-ViT and Wavelet Transform	Transformer attention	Integrates spatial and frequency-domain information.	Performance may depend on image quality and frequency artifacts.

Clapten and Balaji (2025)	Gated temporal attention framework	Temporal attention	Captures temporal inconsistencies in deepfake videos.	Requires video data and temporal information.
Shivaprakash et al. (2026)	Attention-augmented hybrid framework	Attention with evolutionary optimization	Combines feature learning and optimization for robust detection.	Optimization overhead and parameter sensitivity require analysis.
Kumar et al. (2026)	EfficientNet-based spatial-frequency model	Spatial-frequency attention	Integrates spatial and frequency features with interpretability.	Interpretability and generalization across generation methods need validation.
Javed et al. (2026)	Efficient Swin Attention Network	Hierarchical attention	Combines global and local facial features for real-time detection.	Real-time performance may depend on hardware and image resolution.
Gupta et al. (2026)	Hybrid spatiotemporal model	Spatial-temporal attention	Learns spatial facial features and temporal video inconsistencies.	Requires temporal sequences and may increase computational cost.
Reza et al. (2026)	ARC-Net	Attention and residual mechanisms	Focuses on facial image detection, fairness, and generalization.	Fairness across demographic groups and unseen datasets requires further testing.

Table 2: Quantitative Performance Comparison

Ref.	Model attention method	Dataset / evaluation	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
[4]	Efficient-ViT + wavelet transform	FaceForensics++, reported combined evaluation	92.84	NR	NR	NR	98.51
[4]	Efficient-ViT + wavelet transform	Celeb-DF, cross-dataset evaluation	72.3	NR	NR	NR	74.7
[5]	IP-GTA Net, temporal attention	Celeb-DF, test set	90.52	90.64	90.50	90.51	97.77
[7]	EfficientNet + spatial-frequency attention	Study-reported test set	97.2	96.5	97.8	97.1	99.7
[10]	ARC-Net, attention + residual learning	Study-reported test set	99.0	Class-specific	Class-specific	Class-specific	NR

2.1 Dataset-Wise Analysis

The data analysis examines ten selected research studies published between 2020 and 2026 on attention-based deep learning techniques for fake-face and deepfake detection. The studies are analyzed according to their model architectures, attention mechanisms, feature extraction strategies, and major research objectives.

The analysis focuses on identifying methodological trends, comparing the contributions of different approaches, and determining research gaps for developing a robust complementary attention-based deep learning framework.

Table 3: Dataset

Dataset	Real videos	Fake videos	Total videos	Frames / face images
FaceForensics++ (FF++)	1,000	4,000	5,000	Sampling protocol dependent
Celeb-DF v2	590*	5,639	6,229*	Sampling protocol dependent
DFDC (full dataset)	23,654	104,500	128,154	Sampling protocol dependent
WildDeepfake	Version dependent	Version dependent	Commonly reported as 7,314 face sequences from 707 videos	Face sequences are not equivalent to raw frames

2.2 Comparative Analysis of Model Architectures

CNN-based architectures are widely used because they can extract local facial characteristics, including texture irregularities, boundary inconsistencies, and manipulation artifacts. EfficientNet-based approaches emphasize efficient feature extraction, while residual architectures support deeper feature learning.

Transformer-based architectures, including Efficient-ViT and Swin Transformer, focus on contextual relationships between facial regions. Their attention mechanisms enable the model to consider relationships between different image patches, potentially improving the identification of complex facial manipulations.

Hybrid architectures combine CNN and Transformer characteristics to learn both local and global facial representations. These methods are relevant when manipulation artifacts are distributed across different facial regions rather than concentrated in a single location.

3. Analysis of Attention Mechanisms

Table 4. Analysis of Attention Mechanisms

Attention mechanism	Research focus	Expected benefit
Multi-attention	Multiple facial features	Emphasizes different informative representations.
Spatial attention	Facial regions	Highlights potentially manipulated regions.
Channel attention	Feature channels	Assigns importance to informative feature maps.
Halo / Transformer attention	Contextual relationships	Captures relationships across image regions.
Temporal attention	Video frames	Identifies inconsistencies across successive frames.

Spatial-frequency attention	Spatial and frequency features	Combines texture information with frequency-domain artifacts.
-----------------------------	--------------------------------	---

The analysis indicates that individual attention mechanisms address different aspects of fake-face detection. Spatial attention emphasizes localized facial information, while channel attention identifies informative feature representations. Transformer-based attention captures contextual dependencies, and temporal attention is particularly relevant to video manipulation analysis.

Combining complementary attention mechanisms may therefore provide a way to learn multiple types of discriminative information within a unified architecture.

4. Research Gaps Identified

The reviewed studies indicate several areas requiring further investigation:

1. Cross-dataset generalization: Models may experience performance degradation when tested on unseen datasets or unfamiliar face-generation techniques [11, 12].
2. Complementary feature learning: Combining spatial, channel, and frequency-domain attention requires systematic evaluation to determine whether the mechanisms provide complementary information.
3. Unseen manipulation detection: Highly realistic synthetic faces may contain subtle artifacts that are difficult to identify using existing feature extraction techniques.
4. Computational efficiency: Complex attention mechanisms and hybrid architectures may increase memory requirements and inference time.
5. Interpretability: Detection systems need to provide meaningful explanations of the facial regions and features influencing their predictions [13].
6. Fairness and robustness: Detection performance should be evaluated across diverse facial populations, image conditions, and manipulation techniques.

III. ROBUST FAKE-FACE DETECTION

The robust fake-face detection framework consists of two major stages: the overall detection process and the architecture of the complementary attention mechanism in Fig. 1. In the first stage, an input facial image is passed through preprocessing operations, including face detection, alignment, resizing, normalization, and noise or artifact enhancement. The preprocessed image is then processed by a complementary attention-based deep learning model consisting of spatial attention, channel attention, frequency attention, and global attention mechanisms. Spatial attention focuses on important facial regions, such as the eyes, nose, and mouth, while channel attention identifies the most informative feature maps. Frequency attention analyzes frequency-domain characteristics to identify subtle manipulation artifacts, whereas global attention, implemented using a Transformer, captures long-range relationships between different facial regions. The extracted features are integrated through a feature fusion layer and passed to a classification layer to distinguish between genuine and manipulated facial images [14, 15].

The second stage illustrates the detailed architecture of the complementary attention mechanism. Initially, the input facial image of size $224 \times 224 \times 3$ is processed using a CNN backbone, such as EfficientNet or ResNet, to extract important facial features. The extracted feature maps are then distributed across four parallel attention branches: spatial attention, channel attention, frequency attention, and global attention. Each branch independently processes the feature representations to identify different types of discriminative information. The refined features obtained from these branches are combined through a feature fusion mechanism using concatenation or addition. The fused representation is subsequently passed through global average pooling, a fully connected layer, and a Softmax classification layer. Finally, the model produces two possible outputs: genuine face or manipulated face.

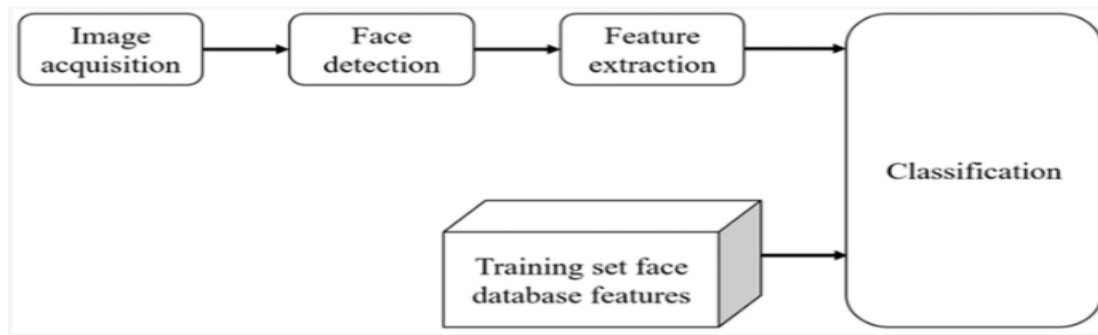


Figure 1: Fake Face Detection

This complementary architecture aims to improve the identification of subtle facial manipulation artifacts by integrating local, global, and frequency-domain information within a unified deep learning framework [16].

IV. DEEP LEARNING

Deep learning is an advanced branch of artificial intelligence that uses multilayer neural networks to automatically learn complex patterns and extract meaningful features from large datasets. In robust fake-face detection, deep learning techniques are used to distinguish genuine facial images from artificially generated or manipulated faces by analyzing facial textures, shapes, expressions, and structural inconsistencies. Convolutional Neural Networks (CNNs), such as ResNet, EfficientNet, and DenseNet, are widely used for extracting local facial features and identifying manipulation artifacts. Transformer-based architectures capture global relationships between different facial regions through attention mechanisms. In the proposed complementary attention-based deep learning framework, spatial attention, channel attention, frequency attention, and global attention are integrated to improve feature representation and classification performance [17, 18].

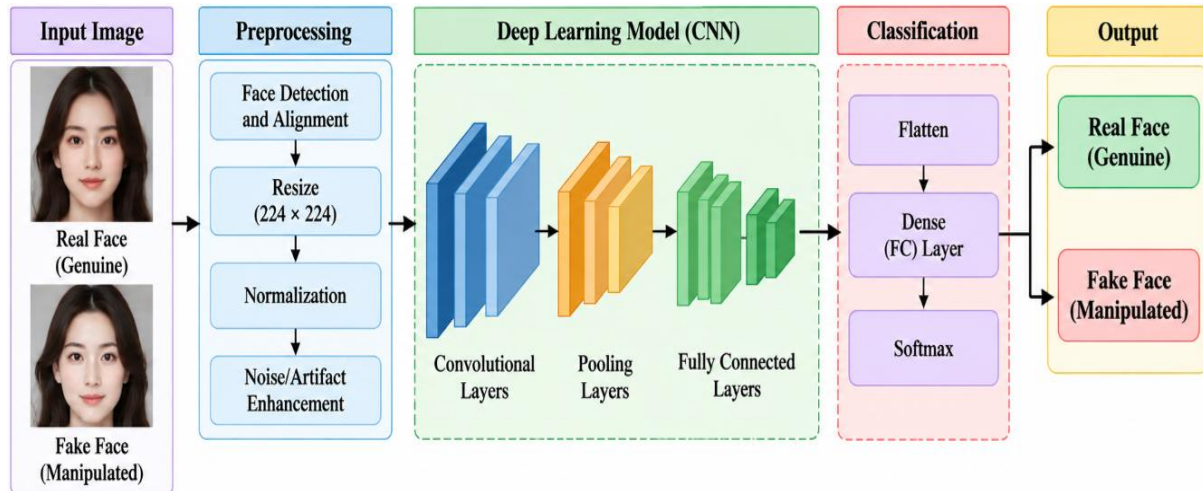


Figure 2: Deep Learning

The extracted features are combined through a feature fusion mechanism and passed to fully connected and Softmax layers for final classification. The model produces two output classes: genuine face and manipulated face. The integration of complementary attention mechanisms aims to improve the model's ability to identify subtle manipulation artifacts and enhance its robustness against variations in image quality, facial appearance, and manipulation techniques. The system uses a CNN to classify facial images into two categories: Real Face (Genuine) and Fake Face (Manipulated) is present in Fig. 2. First, the input image undergoes face detection and alignment to identify the facial region and maintain a consistent orientation. The detected face is then resized to 224×224 pixels and normalized to improve the efficiency and stability of model training. Noise and unwanted artifacts are enhanced or reduced so that important manipulation-related features can be identified effectively. The preprocessed image is passed through convolutional layers, which automatically extract important facial features such as edges, textures, and facial patterns. Pooling layers reduce the feature-map dimensions, while fully connected layers combine the extracted features for decision-making. Finally, the Flatten, Dense, and Softmax layers classify the image and generate the final output as either a Real Face or Fake Face.

Table 5: Research Gaps and Specific Future Directions

Research gap	Why it matters	Specific future direction / measurable evaluation
Cross-dataset generalization	Models may exploit source-specific artifacts rather than general manipulation cues.	Train on FF++ and test zero-shot on Celeb-DF and DFDC; report AUC, F1, and performance drop.
Unseen generator detection	New synthesis systems may not share known artifacts.	Leave one manipulation family out during training; test on unseen generators and synthetic-image collections.

Unclear contribution of attention branches	Hybrid gains may come from parameter count rather than complementary evidence.	Run ablations: backbone only, each attention branch alone, pairs, and full fusion; report accuracy/AUC and cost.
Corruption sensitivity	Compression and resizing can erase or change forensic traces.	Evaluate multiple JPEG/video compression levels, resolutions, blur/noise severities; report robustness curves.
Poor efficiency reporting	Real-time claims are hard to compare across papers.	Report parameters, FLOPs, model size, latency, FPS, memory, hardware, and preprocessing costs.
Limited interpretability validation	Attention maps are not automatically faithful explanations.	Compare maps with localization annotations where available; use perturbation tests and explanation stability.
Fairness and dataset bias	Aggregate metrics can hide subgroup error differences.	Report subgroup F1/AUC and false-positive rates where ethical and reliable demographic annotations exist.
Metric and split inconsistency	Published numbers can be misleading when protocols differ.	Adopt source-video-level splits, public evaluation scripts, fixed thresholds, and common metric definitions.
Video-to-image scope confusion	Frame-level and video-level detection measure different tasks.	Clearly label task level and report frame-to-video aggregation method and latency.

V. METHODOLOGY

This survey adopts a structured literature review methodology to examine attention-based deep learning techniques for robust deepfake face detection. The primary objective is to identify, classify, and critically analyze existing approaches according to their attention mechanisms, feature extraction strategies, model architectures, datasets, evaluation methods, and limitations. Particular emphasis is placed on spatial attention, channel attention, frequency-domain attention, self-attention, and hybrid attention-based architectures. The methodology consists of literature identification, study selection, eligibility assessment, thematic classification, comparative analysis, and research gap identification.

A. Research Objectives and Questions

The survey is guided by the following research questions:

1. Which attention mechanisms are commonly employed in deepfake face detection?
2. How do CNN-based, Transformer-based, and hybrid architectures utilize attention to extract discriminative facial features?
3. Which datasets and evaluation metrics are used to assess deepfake detection performance?

4. What limitations affect the generalization, robustness, interpretability, and computational efficiency of existing approaches?
5. Which research directions may support the development of more reliable and efficient attention-based detection systems?

B. Literature Search Strategy

The literature search focuses on research related to deepfake face detection, facial manipulation analysis, attention mechanisms, and deep learning architectures. Relevant bibliographic databases and scholarly search platforms, such as IEEE Xplore, Scopus, Web of Science, ScienceDirect, SpringerLink, and Google Scholar, can be used to identify journal articles and conference papers. The search strategy combines keywords related to the detection task, attention mechanisms, and deep learning architectures.

Representative search terms include “deepfake face detection,” “attention-based deepfake detection,” “spatial attention,” “channel attention,” “frequency-domain analysis,” “Vision Transformer,” “self-attention,” “CNN–Transformer hybrid,” and “robust facial forgery detection.” Boolean combinations, such as (“deepfake detection” OR “face forgery detection”) AND (“attention mechanism” OR “self-attention” OR “Transformer”), can be used to improve search relevance.

The search period, database-specific queries, search dates, and number of retrieved records should be documented to ensure transparency and reproducibility. Only searches actually conducted should be reported in the final manuscript.

C. Inclusion and Exclusion Criteria

The identified studies are assessed against predefined eligibility criteria to maintain relevance to the survey objectives.

Inclusion criteria:

1. Studies addressing deepfake detection, manipulated facial images, or facial forgery detection.
2. Research investigating spatial, channel, temporal, frequency-domain, self-attention, or related attention mechanisms.
3. Studies employing CNNs, Transformers, hybrid architectures, or other relevant deep learning methods.
4. Publications providing sufficient methodological information for analysis.
5. Relevant research articles, conference papers, and foundational studies that contribute to the understanding of detection methods or benchmark datasets.

Exclusion criteria:

1. Duplicate records of the same publication.
2. Studies unrelated to facial manipulation or the defined scope of the survey.
3. Publications that do not provide sufficient information to establish their relevance.
4. Non-scholarly material that does not meet the survey's evidence requirements.
5. Studies whose claims or bibliographic details cannot be adequately verified, unless explicitly retained for a justified historical or contextual purpose.

These criteria are intended to reduce irrelevant material and support a consistent comparison of the selected literature.

D. Study Screening and Selection

The study selection process comprises four stages. First, potentially relevant records are retrieved using the selected search terms. Second, duplicate records are identified and removed. Third, titles and abstracts are screened against the inclusion and exclusion criteria. Finally, the full texts of potentially eligible publications are assessed for their relevance to attention-based deepfake face detection.

The number of records identified, duplicates removed, studies excluded during screening, full-text articles assessed, and studies retained for the final analysis should be reported in a study-selection flow diagram. The actual counts must be obtained from the completed search and screening process rather than estimated.

E. Data Extraction and Study Classification

For each selected publication, relevant information is extracted into a structured comparison table. The extracted fields include author and publication year, proposed architecture, attention mechanism, dataset, task type, evaluation metrics, principal findings, and reported limitations. Quantitative performance values are recorded only when explicitly available in the original publication and are interpreted in the context of the associated dataset and experimental protocol.

The studies are classified into the following methodological categories:

1. **CNN-based attention methods:** Approaches using convolutional feature extraction with spatial or channel attention.
2. **Transformer-based methods:** Approaches using self-attention to model contextual relationships between facial regions.
3. **Frequency-aware methods:** Approaches integrating spectral or frequency-domain information with learned features.
4. **Hybrid attention methods:** Architectures combining CNNs, Transformers, or multiple attention mechanisms.
5. **Temporal attention methods:** Approaches analyzing relationships between video frames to detect temporal inconsistencies.

This classification provides a common basis for comparing the design objectives and methodological characteristics of the reviewed studies.

F. Comparative Analysis

A qualitative comparative analysis is performed to examine the advantages, limitations, and applicability of the identified approaches. The comparison considers the type of attention mechanism, local and global feature extraction capabilities, dataset diversity, cross-dataset generalization, computational requirements, and interpretability.

Where studies report comparable experimental results, metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC) are summarized. Performance values obtained using different datasets, train–test splits,

compression settings, or evaluation protocols are not treated as directly equivalent. Missing metrics are marked as not reported rather than estimated.

Dataset analysis also distinguishes the number of original videos, manipulated videos, sampled frames, and extracted facial images wherever these details are available. This distinction is important because preprocessing and frame-sampling strategies can substantially influence the quantity and characteristics of the data used for training and evaluation.

G. Research Gap Identification

The final stage synthesizes the evidence from the comparative analysis to identify limitations and opportunities for further research. The principal areas of investigation include cross-dataset generalization, detection of unseen manipulation techniques, complementary attention and feature fusion, robustness to compression and image degradation, computational efficiency, interpretability, and fairness.

The findings are used to develop a conceptual complementary attention-based framework integrating spatial, channel, frequency, and global attention. This framework represents a synthesis of potential design strategies identified in the reviewed literature rather than a claim of experimentally validated superiority. Its effectiveness would require implementation, ablation studies, and evaluation using appropriate benchmark datasets and reproducible experimental protocols.

VI. CONCLUSION

The comprehensive survey of attention-based deep learning techniques for robust fake-face detection highlights the growing importance of artificial intelligence in identifying manipulated facial images and videos. The rapid development of deepfake generation techniques has created significant challenges for digital security, social media monitoring, and information authenticity. Traditional machine learning approaches often struggle to detect sophisticated facial manipulations due to their limited ability to capture complex visual patterns. Attention-based deep learning models address these challenges by focusing on important facial regions, manipulation artifacts, and subtle inconsistencies in images and videos. Techniques such as spatial attention, channel attention, self-attention, and multi-attention mechanisms improve feature representation and support the identification of discriminative features for fake-face detection. The integration of Convolutional Neural Networks, Vision Transformers, and hybrid deep learning architectures further enhances the ability to analyze facial textures, local artifacts, and global relationships. However, several challenges remain, including variations in image quality, compression artifacts, unseen manipulation methods, dataset bias, and limited generalization across different datasets. Future research should focus on developing lightweight, explainable, and computationally efficient attention-based models that can perform reliably in real-world environments. The use of diverse datasets, cross-dataset evaluation, and advanced optimization techniques may further improve detection robustness and generalization. Overall, attention-based deep learning provides a promising foundation for

developing reliable fake-face detection systems and strengthening the authenticity of digital visual content.

REFERENCES

- [1] Y. Sheng, Z. Zou, Z. Yu, M. Pang, W. Ou, et al., "ID-insensitive deepfake detection model based on multi-attention mechanism," *Scientific Reports*, vol. 15, Art. no. 11168, 2025, doi: 10.1038/s41598-025-96254-6.
- [2] S. Guo, Q. Li, M. Gao, X. Zhu, and I. Rida, "Generalizable deepfake detection via Spatial Kernel Selection and Halo Attention Network," *Image and Vision Computing*, vol. 160, Art. no. 105582, 2025, doi: 10.1016/j.imavis.2025.105582.
- [3] A. P. Magesh, S. S. M. Ramakrishnan, R. A. Arun, N. Priyanka, and M. N. Kartheek, "Building an efficient Deep Fake detection system using the recognition capabilities of convolutional neural networks and transformers," *Discover Computing*, vol. 28, Art. no. 99, 2025, doi: 10.1007/s10791-025-09586-2.
- [4] T. Zheng, "An attention-based framework for detecting face forgeries: Integrating Efficient-ViT and wavelet transform," *Mathematics*, vol. 13, no. 16, Art. no. 2576, 2025, doi: 10.3390/math13162576.
- [5] J. E. Clapton and V. Balaji, "A gated temporal attention based intra prediction framework for robust deepfake video detection," *Scientific Reports*, vol. 15, Art. no. 38540, 2025, doi: 10.1038/s41598-025-22333-3.
- [6] S. J. Shivaprakash, S. H., A. Chauhan, and A. Q. Md, "Attention-augmented hybrid framework with evolutionary optimization for robust deepfake detection," *Scientific Reports*, vol. 16, Art. no. 24070, 2026, doi: 10.1038/s41598-026-51284-6.
- [7] M. Kumar, A. Kumar, V. Yadav, A. Shankar, and P. Chauhan, "A hybrid spatial–frequency attention-based algorithm using EfficientNet for robust and interpretable deepfake detection," *Scientific Reports*, vol. 16, Art. no. 16612, 2026, doi: 10.1038/s41598-026-46086-9.
- [8] M. Javed, D. Han, F. H. Dahri, T. Kumar, J. A. Bhutto, et al., "Real time detection of deepfakes using the efficient Swin attention network with global and local facial features," *Discover Artificial Intelligence*, vol. 6, Art. no. 508, 2026, doi: 10.1007/s44163-026-01188-1.
- [9] S. Gupta, O. Pandey, D. Varshney, M. Kumar, and G. C. Gandhi, "A hybrid attention-based spatio-temporal model for deepfake video detection," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-58264-w.
- [10] M. S. Reza, F. Elias, M. I. Mahmud, and N. Ahmed, "Attention and residual mechanism-based CNN architecture (ARC-Net) with enhanced fairness generalization for deepfake facial image detection," *PLOS ONE*, vol. 21, no. 4, Art. no. e0340099, 2026, doi: 10.1371/journal.pone.0340099.
- [11] A. Bhattacharjee, K. Islam, K. Anan, A. Intesher, A. A. Fuad, U. Saha, and H. Imtiaz, "CAE-Net: Generalized deepfake image detection using convolution and attention mechanisms with spatial and frequency domain features," *Journal of Visual Communication and Image Representation*, vol. 115, Art. no. 104679, 2026, doi: 10.1016/j.jvcir.2025.104679.
- [12] T. Rajesh and S. Maruthupermal, "Generalizable deepfake detection framework using hybrid convolution-based EfficientNetB7 with attention mechanism," *Expert Systems with Applications*, vol. 297, Part B, Art. no. 129357, 2026, doi: 10.1016/j.eswa.2025.129357.
- [13] T. Hussain, B. T. Reddy, K. Phanindra, S. Terumalasetti, and G. A. Khan, "Decoding deception: State-of-the-art approaches to deep fake detection," *Frontiers in Big Data*, vol. 8, Art. no. 1670833, 2026, doi: 10.3389/fdata.2025.1670833.
- [14] A. Raza, K. Munir, and M. Almutairi, "A novel deep learning approach for deepfake image detection," *Applied Sciences*, vol. 12, no. 19, Art. no. 9820, 2022, doi: 10.3390/app12199820.
- [15] S. M. Zobaed, M. F. Rabby, M. I. Hossain, E. Hossain, S. Hasan, A. Karim, and K. M. Hasib, "DeepFakes: Detecting forged and synthetic media content using machine learning," in *Artificial Intelligence in Cyber Security: Impact and Implications*. Cham, Switzerland: Springer, 2021, pp. 177–201.
- [16] V. Thambawita, J. L. Isaksen, S. A. Hicks, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, C. Ellervik, M. S. Olesen, T. Hansen, et al., "DeepFake electrocardiograms using generative adversarial networks are the



**International Journal of Advanced Research and
Multidisciplinary Trends (IJARMT)**
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 7.2 Website: <https://ijarmt.com> ISSN No.: 3048-9458

beginning of the end for privacy issues in medicine,” *Scientific Reports*, vol. 11, Art. no. 21896, 2021, doi: 10.1038/s41598-021-01295-2.

- [17] M. F. B. Ahmed, M. S. U. Miah, A. Bhowmik, and J. B. Sulaiman, “Awareness to Deepfake: A resistance mechanism to Deepfake,” in *Proc. 2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, 2021, pp. 1–5, doi: 10.1109/ICOTEN52080.2021.9493549.
- [18] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, “The Deepfake Detection Challenge (DFDC) Preview Dataset,” *arXiv preprint* arXiv:1910.08854, 2019.