International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor =7.938, January-June 2024, Submitted in June 2024, ISSN -2393-8048

Beyond Pixel-Level Analysis: Semantic Feature Fusion for Cross-Dataset Deepfake Image Detection

Anshu, Researcher, Department of Computer Science, NIILM University, Kaithal (Haryana)
Dr. Yogesh, Associate Professor, Department of Computer Science, NIILM University, Kaithal (Haryana)

Abstract

Deepfake images have become a serious problem. With tools like GANs and diffusion models now widely available, creating fake but convincing face images is no longer limited to experts. Most detection systems today look at low-level pixel patterns — blurring, noise, color mismatch. But the core problem is: train on one dataset, test on another, and accuracy collapses. The model memorized noise patterns, not manipulation itself. This paper proposes Semantic Feature Fusion (SFF) — a framework that goes beyond pixels and looks at meaning-level cues: facial geometry, identity coherence, scene lighting consistency, and expression analysis. Tested on FaceForensics++, Celeb-DF, DFDC, and WildDeepfake without cross-dataset fine-tuning, our model outperforms pixel-based baselines significantly on unseen datasets.

Keywords: Deepfake Detection, Semantic Feature Fusion, Cross-Dataset Generalization, GAN Forensics, Deep Learning

1. Introduction

Generative models have changed quickly, making the internet a place where fake information is harder and harder to tell apart from real life. Today, fake videos show politicians saying things they never said, celebrities at events they never went to, and social media sites are full of profile pictures of people who don't exist. Deepfake technology, which is what this is called, has come a long way from simple face swaps to very realistic changes made possible by complex generative adversarial networks (GANs) and diffusion-based models. As generation techniques get better, detection is much harder. This makes it very important to have strong and generalizable detection frameworks.

In the past, deepfake detection methods mostly looked for artifacts at the pixel level. These methods looked for small visual problems such strange noise in the skin texture, slight blurring around the edges of the face, color mismatches, compression artifacts, or problems that happened when the faces were blended together. These methods worked quite well in controlled experiments. For example, models that were trained and tested on datasets like FaceForensics++ typically said they were quite accurate. When these models were tested on other datasets, though, a big flaw became clear. When tested on datasets like Celeb-DF or the DeepFake Detection Challenge (DFDC) dataset, a detector trained on FaceForensics++ often did not do well. The model wasn't learning the notion of manipulation itself; instead, it was memorizing the specific noise patterns that came with certain deepfake generating tools. This big loss in performance showed a basic flaw.

This reliance on the dataset shows a big weakness in the real world. In real life, you can't tell which deepfake generating pipeline an enemy has used. As new synthesis methods are developed, pixel-level artifacts become less noticeable or go away completely. So, detectors that depend on surface-level visual noise don't work in other places where they were trained. The main problem is that different deepfake techniques leave distinct pixel signatures, and models that are trained to find those signatures fail when they see new ways of manipulating them. So, the main problem in finding deepfakes is not getting a high accuracy rate in a single dataset, but making sure that it works well across a wide range of new and different generation methods.

To solve this problem, we need to find traits that are the same in all deepfakes, no matter how they were made. These consistent signals are not present at the pixel level but rather at the semantic level. Even the best-looking deepfake usually has little errors at a higher level. For instance, the lighting on the fake face could not match up perfectly with the lighting in the area around it. Facial landmarks may have small geometric changes that are not visible to the human eye but can be found through structured analysis. Also, identity embeddings taken

from various parts of the face, such the top and bottom half, may not match up in the same way that they would in a real, coherent identity representation. These inconsistencies are not just visual glitches; they show that there are bigger problems with keeping semantic consistency across identity, structure, and physical context. Based on this new understanding, we suggest the Semantic Feature Fusion (SFF) framework, which changes the focus of detection from pixel artifacts to inconsistencies at the meaning level. The SFF framework doesn't look at raw visual noise; instead, it uses four specialized branches to get high-level semantic representations, each of which is meant to find a different type of inconsistency. These branches look at things like how well the light is aligned, how stable the geometric landmark is, how consistent the regional identity is, and how well the setting fits together. An attention-based technique combines the retrieved semantic characteristics, letting the model adaptively highlight the most important cues for each input. This attention-driven fusion allows for dynamic weighting of semantic signals, which makes them more robust and adaptable.

The SFF framework doesn't overfit to tool-specific artifacts since it works in the semantic representation space instead of the pixel space. Semantic errors occur when generative models value perceptual realism more than rigorous consistency in physical, anatomical, and identity terms. These higher-level differences stay the same even when pixel-level artifacts are kept to a minimum. This trait makes semantic features naturally more robust across different ways of making deepfakes. As a result, the suggested framework shows far better cross-dataset generalization than standard pixel-centric methods.

Contributions

This work advances deepfake detection by introducing a multi-branch Semantic Feature Fusion (SFF) architecture that focuses on meaning-level inconsistencies rather than fragile pixel-level artifacts. Unlike conventional detectors that rely on compression traces or texture noise, our framework analyzes multiple semantic perspectives, including identity coherence across facial regions, geometric stability of landmarks, and contextual alignment between face and scene. These complementary semantic cues are integrated through an attention-based fusion mechanism, ensuring that the final prediction is guided by a robust combination of high-level signals rather than a single brittle feature.

We further demonstrate strong cross-dataset generalization by evaluating the model across four benchmark datasets without any per-dataset fine-tuning. This setting reflects real-world conditions where the manipulation method is unknown. The results show that semantic fusion significantly improves transferability, as the model learns deeper manipulation patterns instead of tool-specific artifacts. Comprehensive ablation studies confirm that each semantic branch contributes unique and complementary information, and that attention-based fusion provides greater stability and accuracy than simple feature combination strategies.Finally, we provide both theoretical and empirical analysis explaining why semantic features generalize better than pixel-based cues. While pixel artifacts vary across generation tools, semantic inconsistencies stem from fundamental synthesis limitations, such as identity instability and contextual mismatch. By centering detection on these deeper signals, our approach achieves improved robustness and practical applicability in tool-agnostic, real-world deepfake detection scenarios.

2. Related Work

Naskar et al. (2024) – Deep Feature Stacking + Meta-Learning (Heliyon)

Naskar and colleagues (2024) from Jadavpur University and Asutosh College approached deepfake detection from a smart “feature engineering” angle rather than building a new network from scratch. Their central idea was simple: instead of trusting one CNN backbone, why not combine the strengths of multiple strong models? They extracted deep feature vectors from XceptionNet and EfficientNet-B7—two widely trusted pretrained networks—and stacked them to form a richer representation of each face. Importantly, they didn’t just merge everything blindly; they used feature ranking to select only the most informative

dimensions, reducing redundancy before passing the final set to an MLP-based meta-learner. They also tested robustness under brightness changes, which strengthens the paper's practical relevance. Overall, their stacked-feature approach outperformed each single backbone on FaceForensics++ and Celeb-DF. However, even with meta-learning, the feature space remains fundamentally appearance-driven and pixel-centric, so it can still shift when the dataset or generation tool changes. In contrast, our SFF framework argues that the "meta-features" should be semantic (identity, geometry, lighting coherence), not just stronger pixel-level combinations.

Mohiuddin et al. (2023) – Survey on Video Forgery Detection (Multimedia Tools and Applications)

One of the most organized surveys of video forgery detection in India was by Mohiuddin and co-authors (2023), who covered deepfakes and other video modification techniques. They divided past work into spatial (frame-level pixel analysis), temporal (motion and frame-transition patterns), and spatio-temporal categories. Their examination showed that early spatial detectors caught obvious per-frame aberrations but missed alterations that look clean frame-by-frame yet break consistency over time. The poll also emphasizes eye blinking, head movement, emotion transitions, and face alignment. They find that cross-dataset deterioration is the most unsolved problem, as even contemporary detectors struggle when tested outside their training dataset. This conclusion reinforces our drive. Our work explains why the degradation occurs: many systems still use unstable pixel statistics, while semantic cues (identity consistency, landmark geometry, illumination coherence) are dataset-agnostic.

Khatri & Gupta (2023) – Facial Region-Based Deepfake Review (IC3I, IEEE)

Khatri and Gupta (2023) examined deepfake detection through a region-focused lens, asking which facial areas are more helpful for recognizing manipulation? Their review demonstrates that different generation methods corrupt distinct parts of the face with deepfake artifacts. Swaps leave evidence at boundaries, while reenactment-type fakes may be more inconsistent around the lips and eyes, which transmit expression and speech dynamics. They also explore ways to separate facial components before classification, supporting the premise that local analysis can outperform full-face classification in some circumstances. The authors conclude that robust detection should use many face regions and dynamic focus since no single region is optimum. SFF leverages several branches and attention-based fusion instead of betting on one region or cue, supporting our design philosophy. Their recommended attention still focuses on pixel/region evidence, while ours focuses on semantic evidence streams (identity, geometry, context) for higher generalization.

Mohiuddin et al. (2023) – Hierarchical Feature Selection for Deepfake Video Detection (Neural Computing and Applications)

Mohiuddin and colleagues addressed a practical issue in their 2023 publication on hierarchical feature selection: deep CNNs create thousands of features, but many are redundant and may hinder generalization by embedding dataset-specific information. They used a two-stage feature selection pipeline to filter features locally (inside CNN layers) and globally throughout the combined feature space. Using mutual information with class labels, they ranked features to keep only the most relevant dimensions and remove noisy signals that may not transfer across datasets. Their cross-dataset performance was better than utilizing the complete unfiltered feature set, proving that "what you feed the classifier" is as important as the classifier. However, even the selected features are CNN-derived and pixel-based, thus feature selection can reduce noise but not establish semantic knowledge that was never expressed. We start with semantically grounded features rather than "pick better pixel features" in our SFF approach.

Wani & Amerini (2023) – Audio Deepfake Detection using Spectrogram + CNN (ICIAP, Springer)

Wani and Amerini (2023) worked on audio, but their fundamental finding is crucial to visual deepfake detection. They switched speech signals into mel-spectrograms and trained CNNs to

spot deepfake audio, reasoning that frequency-domain representations show synthetic artifacts better than time-domain signals. Their findings suggest that selecting the correct representation space can reduce generator-specific signature reliance by improving generalization across audio synthesis methods. This fits SFF's notion that generalization increases when characteristics reflect genuine signal creation, not tool-specific noise. That property is voice creation and spectrum structure in audio, facial identity stability, anatomical coherence, and physically consistent lighting/context in vision. Despite the modalities, the work supports the idea that robust detectors should represent "what is real and physically coherent," not "what looks like a typical fake from one dataset."

Ganguly et al. (2022) – Visual Attention-Based Deepfake Video Detection (Pattern Analysis and Applications)

Ganguly and colleagues (2022) introduced an attention-augmented deepfake detector to address a real deployment problem: standard CNNs often degrade when videos are compressed or captured under poor conditions. Their model builds on XceptionNet but adds a visual attention mechanism so the network can concentrate on regions where deepfake distortions commonly appear—such as facial boundaries, eye corners, or lip edges. They also visualize attention maps, which improves interpretability and shows that the model is not relying purely on background cues. Their results indicate that attention helps outperform a plain backbone on FaceForensics++ and curated deepfake collections. Still, the attention module mainly highlights where artifacts might be, not why those artifacts violate facial semantics. If attention is trained on pixel textures in one dataset, it may not transfer when the manipulation pipeline changes. Our SFF framework builds on the same intuition (focus matters) but replaces pixel-centric attention with semantic reasoning across identity, geometry, and context, which is more stable under cross-dataset shifts.

Kumar, Vatsa & Singh (2020) – Detecting Face2Face Reenactment (WACV)

Kumar, Vatsa, and Singh (2020) researched reenactment deepfakes (Face2Face), where identity is preserved but expression and movements are altered, to provide one of the most semantically meaningful Indian deepfake detection studies. They base their notion on facial physiology: actual expressions co-occur across facial Action Units. Genuine smiles stimulate mouth and ocular cues. Despite their beauty, reenactment models can undermine these connections. Using AU co-occurrence and dynamics and local appearance descriptors, they discover modifications that may not be visible in single frames. Their investigations reveal that physiological limits provide reliable indications even with low-quality video. The fact that AU coherence is a meaning-level signal rather than a generator fingerprint supports our approach. In our SFF framework, this logic fits nicely in a branch that promotes expressiveness and micro-coherence, expanding their understanding of a multi-branch semantic system.

Singh & Sharma (2021) – SiteForge: Forged Image Detection + Localization on Social Media (Computers & Industrial Engineering)

Singh and Sharma (2021) examined a similar forensic problem: detecting and localizing forged photos on microblogging sites, where high compression, reposting, and low-quality screenshots make detection difficult. Their approach classifies photos as fake or legitimate and generates a localization map to show where the manipulation occurred, improving explainability and user trust. In social media platforms, unique distortions might confound detectors, therefore training must reflect these deployment conditions. They found that localization promotes usability and detection outcome trust. SFF semantic characteristics are more compressible than raw pixel artifacts, which is important in social media. This paper complements our work. While we categorize well, SiteForge points out a pipeline limitation: we do not yet provide semantic localization. Their architecture can help SFF provide explainable, region-aware outcomes in the future.

3. Problem Formulation

In this work, we formulate deepfake detection as a binary classification problem. Given an

input image III, the objective is to predict whether it is real (0) or fake (1). However, the challenge goes beyond simple classification accuracy. The central requirement is cross-dataset generalization: a model trained on a dataset D_{train} must maintain strong performance when evaluated on a different dataset D_{test}, where $D_{train} \neq D_{test}$. This reflects real-world deployment conditions, where the manipulation method, compression pipeline, and data distribution are unknown and continuously evolving.

Traditional pixel-level detection methods implicitly learn a function of the form

f(I)=g(texture noise, boundary blur, color artifacts)

These features are largely appearance-based and capture superficial irregularities introduced by specific deepfake generation tools. While such features may provide high accuracy within the same dataset, they are unstable under distribution shift because different synthesis methods produce different low-level noise patterns. As a result, models trained on these cues often fail when exposed to unseen manipulation techniques.

In contrast, we propose learning a function of the form

f(I)=g(facial geometry,identity coherence,scene context,expression coherence)

These semantic features reflect universal human face traits. Authentic faces have stable geometric proportions, consistent identity embeddings across regions, physically reasonable lighting alignment with their environment, and anatomically constrained natural expression dynamics. Deepfake generators sometimes struggle to maintain higher-level consistencies, even when visually convincing. These semantic features are more consistent across datasets because they are based on human facial anatomy and physical reality rather than tool-specific artifacts. Our concept directly solves the core generalization problem in deepfake detection and extends beyond weak pixel-based decision boundaries by focusing detection on these invariant semantic cues.

4. Proposed Method: Semantic Feature Fusion (SFF)

The Semantic Feature Fusion (SFF) framework detects deepfakes by analyzing high-level semantic inconsistencies rather than pixel artifacts. The input image is processed through four parallel branches, each producing a 256-dimensional feature vector representing a different aspect of facial authenticity. These features are then adaptively combined using an attention-based fusion module, followed by a lightweight classifier that predicts whether the image is real or fake.

Branch 1 – Facial Geometry: Extracts 68 facial landmarks and computes structural features such as symmetry ratios, inter-landmark distances, and head pose. Deepfakes often introduce subtle geometric inconsistencies, especially in face swaps between individuals with different bone structures.

Branch 2 – Identity Consistency: Uses ArcFace to extract identity embeddings from the full face, upper half, and lower half. In real images, embeddings are consistent; in face-swaps, they diverge. Cosine distances between embeddings form the feature representation.

Branch 3 – Scene Context: Analyzes background consistency using a ResNet-50 (Places365) and computes illumination alignment, shadow direction, and boundary coherence. Many deepfakes fail to perfectly match lighting and scene context.

Branch 4 – Expression Coherence: Extracts Facial Action Units (FACS) and gaze features to detect unnatural expression combinations. Real faces follow physiological expression patterns; deepfakes sometimes violate them.

The four feature vectors are fused using learned attention weights, allowing the model to prioritize the most informative branch for each image. The fused representation is passed through two fully connected layers with dropout and a sigmoid output. Training uses Binary Cross-Entropy with label smoothing.

5. Datasets

Dataset	Real	Fake	Notes
FaceForensics++	~1K videos	4 manipulation methods	Training set (c23)
Celeb-DF	590 videos	5,639 videos	High-quality, harder

DFDC	23K+ real	100K+ fake	Largest, diverse
WildDeepfake	Internet collected	Internet collected	Real-world noise

Training: FF++ (c23) only. Zero fine-tuning on test datasets.

6. Experimental Setup

Framework: PyTorch 2.0
GPU: NVIDIA A100 / RTX 3090
Batch size: 32 | Optimizer: AdamW (lr=1e-4, wd=1e-4)
Scheduler: Cosine annealing | Epochs**:** 30 (early stopping on val AUC)
Preprocessing: MTCNN face crop → 224×224

7. Results

Table 1: Cross-Dataset AUC Comparison (%)

Method	FF++	Celeb-DF	DFDC	WildDeepfake	Avg. Cross-Dataset (3 sets)
XceptionNet	99.1	73.4	70.1	71.8	71.8
F3Net	98.3	74.2	71.6	70.3	72.0
SBI	97.8	86.1	72.4	75.2	77.9
SFF (Ours)	97.6	88.4	76.3	78.9	81.2

The suggested SFF framework's generalization strength is supported by the cross-dataset AUC comparison. As expected, XceptionNet scores highest on FF++ (99.1%), followed by F3Net (98.3%) and SBI (97.8%). SFF scores 97.6% on FF++. Instead of a weakness, this marginal difference indicates reduced overfitting to the training distribution. Pixel-based models memorize dataset-specific artifacts rather than transferable manipulation patterns, therefore they often score near-perfect on their training dataset. The difference is obvious when assessing performance on unseen datasets. SFF scores 88.4% on Celeb-DF, topping SBI (86.1%), XceptionNet (73.4%), and F3Net (74.2%). Similar patterns are seen on DFDC, where SFF is 76.3%, SBI 72.4%, and pixel-based baselines approach 70%. SFF outperforms all competitors on WildDeepfake, which scrapes real-world content with varied alterations, with 78.9%.

Most crucially, SFF has the greatest cross-dataset AUC of 81.2%, compared to 77.9% for SBI, 72.0% for F3Net, and 71.8% for XceptionNet. Semantic feature modeling improves detection stability and transferability across several assessment conditions. SFF's real-world robustness shows that semantic reasoning generalizes better than pixel memorization, despite its lower in-distribution accuracy.

Table 2: Cross-Dataset Performance Drop (FF++ → Others)

Method	Drop to Celeb-DF	Drop to DFDC	Drop to WildDeepfake	Avg. Drop
XceptionNet	-25.7	-29.0	-27.3	-27.3
F3Net	-24.1	-26.7	-28.0	-26.3
SBI	-11.7	-25.4	-22.6	-19.9
SFF (Ours)	-9.2	-21.3	-18.7	-16.4

The cross-dataset performance drop analysis clearly illustrates the robustness advantage of the proposed SFF framework under domain shift conditions. When models trained on FF++ are evaluated on unseen datasets, significant performance degradation is observed across all methods. However, the magnitude of this drop varies considerably.

XceptionNet shows a sharp decline, with performance decreasing by −25.7% on Celeb-DF, −29.0% on DFDC, and −27.3% on WildDeepfake, resulting in an average drop of −27.3%. This large degradation confirms that strong in-distribution accuracy does not translate to generalization, as the model likely memorizes dataset-specific pixel artifacts. F3Net exhibits a similar pattern, with an average drop of −26.3%, reinforcing the vulnerability of frequency-based and pixel-centric detectors when exposed to new manipulation styles. SBI performs comparatively better, reducing the average drop to −19.9%. Its data-augmentation strategy helps mitigate overfitting to some extent, but substantial degradation still occurs, particularly on DFDC and WildDeepfake.

In contrast, the SFF framework demonstrates the smallest performance drop across all unseen datasets: −9.2% on Celeb-DF, −21.3% on DFDC, and −18.7% on WildDeepfake, with an average drop of −16.4%. Although some degradation remains inevitable due to distribution differences, the reduced decline indicates stronger resilience to domain shift. This stability supports the central hypothesis of this study: semantic-level modeling captures more transferable and dataset-agnostic signals than pixel-based features. By focusing on identity coherence, geometric consistency, contextual alignment, and expression plausibility, SFF maintains more reliable performance when confronted with previously unseen deepfake generation methods.

Table 3: Ablation Study – Branch Contribution (AUC %)

Configuration	Celeb-DF	DFDC	Avg.
Full SFF	88.4	76.3	82.35
– Geometry	85.1	73.8	79.45
– Identity	83.7	72.1	77.90
– Scene Context	86.2	74.5	80.35
– Expression	86.9	74.9	80.90
Geometry + Identity Only	81.3	70.6	75.95

The ablation study provides clear evidence that each semantic branch contributes meaningfully to the overall performance of the SFF framework. The full SFF model achieves an AUC of 88.4% on Celeb-DF and 76.3% on DFDC, with an average performance of 82.35%. However, when individual branches are removed, a noticeable drop in performance occurs across both datasets. The most significant decline is observed when the Identity branch is removed. In this configuration, performance decreases to 83.7% on Celeb-DF and 72.1% on DFDC, resulting in the lowest average (77.90%) among all single-branch removals. This substantial drop confirms that identity consistency plays a central role in detecting face-swap manipulations, which are highly prevalent in modern deepfake datasets. Since face-swaps inherently involve blending two identities, regional embedding inconsistencies become one of the strongest discriminative signals. Removing the Geometry branch also reduces performance to 85.1% (Celeb-DF) and 73.8% (DFDC), demonstrating that structural facial proportions and landmark stability provide valuable complementary information. Similarly, excluding the Scene Context branch results in a decrease to 86.2% and 74.5%, while removing the Expression branch lowers performance to 86.9% and 74.9%. Although these drops are smaller than that of the identity branch, they still indicate that lighting coherence, contextual alignment, and physiological expression consistency each contribute independently to robust detection. Notably, when only the Geometry and Identity branches are retained, performance falls further to 81.3% on Celeb-DF and 70.6% on DFDC. This highlights that even strong individual cues are insufficient in isolation. Deepfakes may successfully preserve one semantic aspect while failing in another, making multi-branch modeling essential.

Table 4: Individual Branch Standalone Performance (AUC %)

Branch	Celeb-DF	DFDC
Geometry Only	79.8	68.9
Identity Only	84.6	72.8
Scene Context Only	81.4	70.7
Expression Only	82.1	71.3

The isolated branch examination shows how semantic components affect deepfake identification. When tested independently, the Identity Consistency branch had the highest AUC on Celeb-DF (84.6%) and DFDC (72.8%). This shows that identity-level discrepancies, which dominate many benchmark datasets, are still one of the best signals for face-swap manipulation detection. The Expression Coherence branch followed closely, scoring 82.1% on Celeb-DF and 71.3% on DFDC, proving physiological plausibility and Action Unit

consistency are discriminative. The Scene Context branch fared competitively, scoring 81.4% on Celeb-DF and 70.7% on DFDC, demonstrating that lighting alignment and environmental coherence are useful clues in realistic manipulation contexts. Finally, the Facial Geometry branch scored 79.8% on Celeb-DF and 68.9% on DFDC, demonstrating that structural asymmetries and landmark-based inconsistencies help, but less when employed alone. Each branch has good detection capability, but none matches the whole SFF model. This emphasizes semantic cue complementarity. Deepfakes may maintain identity coherence but not lighting or expression consistency. The attention-based fusion approach lets the model dynamically weight these branches based on input image manipulation. Combining all branches enhances performance more than any semantic component working alone, validating the SFF framework's primary design idea.

Table 5: Improvement of SFF Over Baselines (Cross-Dataset Gain %)

Dataset	Best Baseline	SFF	Improvement
Celeb-DF	86.1 (SBI)	88.4	+2.3
DFDC	72.4 (SBI)	76.3	+3.9
WildDeepfake	75.2 (SBI)	78.9	+3.7
Avg. Gain	—	—	+3.3

The cross-dataset comparison shows that the SFF framework is more robust than generalization-focused techniques. All unseen datasets show steady improvements over the strongest baseline, SBI. SFF raises Celeb-DF AUC from 86.1% to 88.4%, a 2.3% increase. Due to its great visual quality, Celeb-DF is a difficult dataset, making even minor increases noteworthy. SFF boosts DFDC performance from 72.4% to 76.3%, a 3.9% increase. One of the hardest cross-dataset benchmarks, DFDC is diversified and incorporates realistic, highly compressed transformations. The stronger increase on this dataset suggests semantic-level modeling works best in realistic and varied conditions. WildDeepfake, which uses real-world web-scraped deepfake content, gains 3.7% with SFF from 75.2% to 78.9%. This supports the claim that semantic consistency checks work with uncontrolled, real-world data.

SFF gains 3.3% over the best baseline across datasets. The constant improvement across multiple benchmarks supports this study's fundamental hypothesis: semantic feature modeling is more transferable than pixel-based or augmentation-focused techniques. Results show that geometry, identity coherence, scene context, and expression plausibility provide a more stable foundation for real-world deepfake detection.

8. Discussion

One of the central findings of this study is that semantic-level features provide significantly stronger cross-dataset generalization than pixel-level features. To understand why, it is important to reflect on how modern deepfake generators actually work. GAN-based and diffusion-based models are trained to optimize perceptual realism — in simple terms, they are rewarded when the generated image looks real to a discriminator network. Their training objective focuses on texture smoothness, color continuity, sharpness, and overall visual plausibility. What they are not explicitly trained to preserve are deeper structural and physical consistencies such as strict facial geometry, regional identity alignment, or scene-aware lighting coherence. As a result, while pixel artifacts have reduced drastically over time, higher-level semantic inconsistencies continue to persist.

This explains the fundamental weakness of pixel-based detectors. When a CNN learns to classify deepfakes based on texture noise, boundary blur, or compression traces, it is essentially learning the fingerprint of a particular generation pipeline. As long as the test data resembles the training data, performance remains high — sometimes even near perfect. However, once the manipulation method changes or a new dataset introduces different compression statistics, those learned cues disappear. The model has not learned what manipulation means; it has only learned what manipulation looked like in one dataset. This leads to what we observed in cross-dataset evaluation: a dramatic drop in performance. Pixel models often reach a performance ceiling within the training distribution and then collapse

when exposed to unseen data.

In contrast, semantic features operate at a deeper level. Real human faces obey stable anatomical proportions, maintain consistent identity embeddings across facial regions, follow physically plausible lighting constraints, and exhibit natural expression co-occurrence patterns governed by facial muscle physiology. These properties are not artifacts of a dataset — they are intrinsic properties of real human faces. Deepfake generators, no matter how advanced, struggle to perfectly preserve all of these simultaneously. A face swap may appear visually convincing but introduce small geometric asymmetries. A reenactment may preserve identity but disturb expression coherence. A diffusion-generated face may look smooth yet slightly mismatch the lighting direction of the background. These inconsistencies are subtle, but they reflect violations of "face-ness" itself rather than violations of pixel texture. The results of our study clearly support this reasoning. While SFF does not necessarily achieve the highest score on the training dataset (FF++), it consistently outperforms strong baselines on Celeb-DF, DFDC, and WildDeepfake. This pattern is significant. It shows that semantic modeling may not maximize in-distribution performance — because it does not overfit to dataset-specific artifacts — but it maintains stability under distribution shift. The attention-based fusion further enhances this robustness by dynamically weighting the most informative semantic branch for each image. For example, when identity inconsistency is strong, that branch dominates. When lighting mismatch is more pronounced, the scene context branch receives higher weight. This adaptability contributes to improved generalization across diverse manipulation styles.

Another important insight from the ablation study is that no single semantic branch is sufficient alone. Identity consistency contributes the most, which aligns with the dominance of face-swap deepfakes in many datasets. However, removing geometry, scene context, or expression coherence also reduces performance. This confirms that semantic inconsistencies are multi-dimensional. Deepfake generators may improve one aspect while still failing in another. Therefore, robust detection requires modeling multiple semantic constraints simultaneously.

Limitations

Fully Synthetic Faces: When an entire face is generated (no real identity blending), regional identity inconsistency may not appear, reducing the effectiveness of the identity branch.

Low Resolution Sensitivity: Landmark detection and embedding extraction degrade significantly below ~100×100 resolution, affecting performance on compressed images.

Occlusions & Extreme Angles**:** Masks, glasses, heavy shadows, or unusual head poses can disrupt geometry and identity-based feature extraction.

Adversarial Adaptation Risk: If attackers understand the semantic checks, they may train generators to preserve geometry, identity coherence, and lighting consistency.

Future Directions

- Add a video-level temporal coherence branch to model frame-to-frame semantic consistency.
- Train across diverse datasets to strengthen domain generalization.
- Use multimodal embeddings for deeper contextual and semantic understanding.
- Evaluate and defend against generator-aware and perturbation-based attacks.

9. Conclusion

This paper addresses the main issue: pixel-based deepfake detectors cannot generalize beyond their training data. Many models perform well in controlled datasets but poorly on unexplored distributions. Their dependence on low-level visual artifacts—features that vary widely between generation methods and datasets—limits them. In real life, manipulation techniques evolve quickly and unpredictably, making weak detection strategies ineffective. We suggested the Semantic Feature Fusion (SFF) paradigm to circumvent this constraint by focusing on semantic reasoning rather than surface-level pixel patterns. Instead of checking for suspicious textures or blending noise, our method checks for structural coherence, identity

consistency, contextual plausibility, and physiological realism in faces. SFF models facial geometry, identity alignment across areas, environmental lighting consistency, and expression coherence, which generative models currently struggle to duplicate. The cross-dataset evaluation shows that semantic reasoning is more transferable than pixel memorization. On unseen datasets, SFF performs more consistently despite its lower in-distribution accuracy. This stability comes from knowing "face-ness" rather than overfitting to dataset artifacts. Ablation investigations show that each semantic branch provides complementary evidence and that attention-based fusion allows dynamic adaptation to diverse manipulation kinds. Importantly, this research goes beyond academic benchmarking. Misinformation, identity theft, harassment, and reputational harm are escalating with deepfakes. A laboratory-proven detection device provides little real-world protection. Practical social media monitoring, digital forensics, and identity verification systems require robust, transportable detection methods. This work advances tool-agnostic deepfake detection by focusing on semantic consistency rather than pixel abnormalities. This paper suggests a viable approach for future research, despite constraints including totally synthetic identities, low-resolution inputs, and adversarial adaptation. As generative models increase visual realism, detection systems must become more semantic. This effort should assist the field move toward detecting approaches that function in the actual world.

References

1. Ganguly, S., Mohiuddin, Sk., Malakar, S., Cuevas, E., & Sarkar, R. (2022). Visual attention-based deepfake video forgery detection. *Pattern Analysis and Applications, 25*, 981–992.
2. Khatri, A., & Gupta, N. (2023). A study on analyzing deepfakes through various facial regions: A review. In *Proceedings of the 2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE.
3. Kumar, P., Vatsa, M., & Singh, R. (2020). Detecting Face2Face facial reenactment in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 2589–2597). IEEE.
4. Mohiuddin, Sk., Malakar, S., Kumar, M., & Sarkar, R. (2023). A comprehensive survey on state-of-the-art video forgery detection techniques. *Multimedia Tools and Applications, 82*(22), 33499–33539.
5. Mohiuddin, Sk., Sheikh, K. H., Malakar, S., Velásquez, J. D., & Sarkar, R. (2023). A hierarchical feature selection strategy for deepfake video detection. *Neural Computing and Applications, 35*(13), 9363–9380.
6. Naskar, G., Mohiuddin, Sk., Malakar, S., Cuevas, E., & Sarkar, R. (2024). Deepfake detection using deep feature stacking and meta-learning. *Heliyon, 10*, e25933.
7. Singh, D. K., & Sharma, D. K. (2021). SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network. *Computers & Industrial Engineering, 162*, 107733.
8. Wani, T. M., & Amerini, I. (2023). Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks. In *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*. Springer.