Volume 23, Issue 1: KUS: 1664: 124-137: 13062026: 2026

KHULNA UNIVERSITY
STUDIES

journal homepage: www.kus.ku.ac.bd

Research article

FIDF: A Hypergraph Convolutional Framework for Detection of Social Media Image Forgery in Compression Artefacts

Md. Mehedi Rahman Rana[1*], Md Anisur Rahman[1], Kamrul Hasan Talukder[1] and Rohul Amin[2]

[1]*Computer Science and Engineering Discipline, Science Engineering and Technology School, Khulna University, Khulna – 9208, Bangladesh*
[2]*Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Siromoni, 9402, Khulna, Bangladesh*

ABSTRACT

The rapid growth of manipulated images on social media has posed the challenging problem of robust image forgery detection. Existing methods either rely on content-driven deep models or single-view forgery cues, which limit their generalization ability under different types of manipulations and imaging conditions. In this paper, we propose FIDF (Fake Image Detection Framework), a novel multi-view discrepancy-gated hypergraph framework for detecting social media image forgeries. This motivates the proposed method based on the key observation that forged images introduce inconsistencies across multiple physically grounded representations. To take advantage of this, FIDF combines three complementary streams of features: a semantic RGB encoder, a learnable spectral representation based on a learnable DCT basis and a noise residual stream initialized from Spatial Rich Model (SRM) filters. These streams are fused by a Discrepancy-Gated Cross-Attention (DGCA) mechanism, which explicitly boosts the inter-view disagreement signals indicative of tampering. The k-nearest-neighbor hypergraph convolution module captures the non-local relations between spatially disjoint, but semantically similar regions and enables effective reasoning over complex manipulation patterns. We also present FIDD-13000 (Fake Image Detection Dataset-13000), a large-scale benchmark that reflects realistic social media forgery scenarios with different manipulation types and compression artefacts. Extensive experiments on FIDD-13000 and four public benchmarks (CASIA v1, CASIA 2.0, Columbia and MICC-F2000) demonstrate that FIDF outperforms the state-of-the-art methods consistently, with superior accuracy, F1-score and AUC under challenging conditions. Additional ablation studies further confirm the roles of multi-view streams, DGCA fusion, and hypergraph reasoning modules.

ARTICLE INFO

Article timeline:
Date of Submission:
13 June, 2026
Date of Acceptance:
24 June, 2026
Article available online:
28 June, 2026

***Keywords*:**
Image Forgery Detection
Hypergraph Neural Networks
Deep Learning-Based Forgery
Social Media Image Forgery
Frequency Domain Analysis

Introduction

Images are one of the main sources of medium and the way we communicate in images, shapes opinions, imparts decision making power as well as impacts social discourse (Chakraborty et al., 2022). However, despite this explosion of visual content, it has also made the rapid proliferation of manipulated images possible — a major digital trust and information integrity challenge. Copyright abuse, misinformation dissemination, deepfake generation and privacy invasion via forged images make image forgery detection mechanisms indispensable (Dell'Anna et al., 2025).

There are several aspects that can make the task of detecting forged images on social media more difficult. For starters, JPEG or other lossy standards require significant compression and format conversion which may obliterate some forensic artifacts (Gorle & Guttavelli, 2025). Second, social media images involve a rich variety of manipulations such as splicing, copy-move operations and the most recent generation of ever more powerful generative models (e.g., GANs; Rossler et al., 2019). Third, images come from heterogeneous devices which have different quality, resolution and noise properties further making it hard to trace down manipulations (Gardel et al., 2021). All these aspects create a rather complex and dynamic context, which is why classical forensic methods are not suitable to be reliable and robust anymore.

*Corresponding author: <mehediranacse11@gmail.com>
DOI: https://doi.org/10.53808/KUS.2026.23.01.1664-se

Approaches Detecting Image Forgery Existing distinct. Since content-driven deep learning models concentrate on the semantic feature of images, their performance can be impaired under compression artifacts or small manipulations. On the other hand, single-view forensic cues (e.g., noise residuals, frequency-domain inconsistencies or metadata analysis) are only able to detect limited types of tampering techniques individually (Farid 2019). In addition, multiple methods are unable to generalize over a large number of manipulation types and the complex impact of social media compression (Bayar & Stamm, 2016; Bayar & Stamm, 2018). This reveals an important gap in the existing literature: Current frameworks fail to model multi-view inconsistencies jointly nor capture spatial relationships that are non-local, which is a key characteristic of complex forgeries.

A central observation motivating this work is that forged images, by their nature, create a small amount of misalignment across several physically grounded representations. Manipulations commonly appear as perturbed semantic content (RGB features), spectral domain (frequency-based inconsistencies) and noise residuals (sensor pattern anomalies). These discrepancies, while subtle and individually difficult to detect, can indicate tampering across dimensions when examined in unison. It is thus indispensable to leverage these multi-view inconsistencies for robust image forgery detection, especially in presence of social media-oriented compression artefacts.

In order to solve these challenges we present: one of them is FIDF (Fake Image Detection Framework), a new multi-view discrepancy-gated hypergraph convolutional approach for social media image forgery detection. FIDF consists of three diverse feature modalities: a traditional semantic encoder on RGB, a spectral representation learnable via a discrete cosine transform (DCT) basis, and the noise-residual stream with filters given by the Spatial Rich Model (SRM). These streams are then combined using a Discrepancy-Gated Cross-Attention (DGCA) mechanism that highlights inter-view disagreement signals which indicate tampering. Moreover, a k-nearest-neighbor hypergraph convolution module encodes long-term relationships between spatially non-contiguous but semantically correlated regions, establishing reasoning for challenging manipulation modes. The application of this multi-view fusion strategy alongside hypergraph reasoning elevates the robustness and generalization performance of our detection approach, even against very challenging compression and manipulation cases.

In response to this challenge, we present FIDD-13000, a newly introduced, balanced dataset of 13,000 ground-truth and forgery images (access details provided in the Data Availability Statement) with four categories: splicing, copy-move, inpainting and retouching. The images are then augmented through general platform-level factors associated with social media including recompression, resizing and filtering (see Figure 1) to align more closely with realistic conditions and cases.

Research Objectives

It aims at designing a deep learning based framework to alleviate the adverse effects of lossy compression and post-processing common in social media environments, on image forgery detection. The aims of this study are as the following:

- To develop FIDD-13000 data set containing different types of manipulation content after the changes similar to those in social media.
- To develop FIDF, a multi-view discrepancy-gated hypergraph framework integrating RGB, frequency-domain, and noise-residual representations with robustness against social media distortions.
- To capture non-local forgery evidence through k-NN hypergraph convolution.
- To comprehensively evaluate both the dataset and framework against conventional hand-crafted and deep learning baselines.

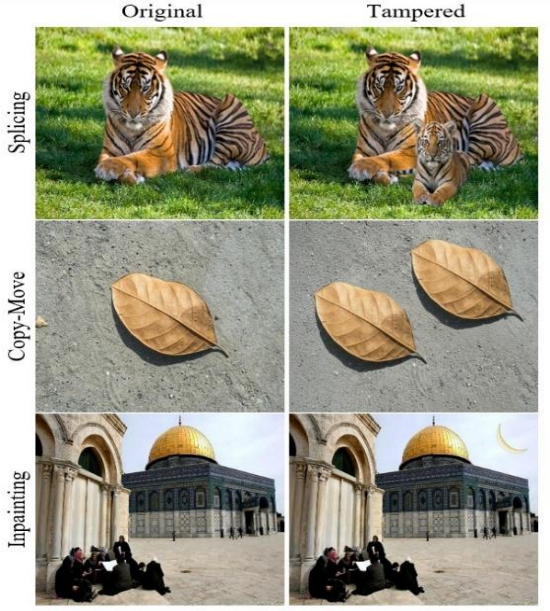

Figure 1: Representative examples of original and tampered images in the FIDD-13000 across splicing, copy-move, and inpainting manipulation types.

Key Contributions of This Research

In this paper, we propose FIDF, a multi-view discrepancy-gated hypergraph framework for image forgery detection, and introduce the FIDD-13000. Our main contributions are summarized as follows:

- **FIDD-13000:** A large-scale dataset of 13,000 images with four manipulation types (splicing, copy-move, inpainting, and retouching), incorporating realistic social media transformations.
- **FIDF:** A novel end-to-end image forgery detection framework that jointly exploits RGB, frequency-domain, and noise-residual representations to capture complementary forgery evidence.
- **Discrepancy-Gated Cross-Attention:** A cross-view fusion strategy that explicitly models disagreement between image representations and amplifies tampering-related inconsistencies.
- **Hypergraph-Based Non-Local Reasoning:** A k-NN hypergraph convolution module that captures higher-order relationships among spatially distant but feature-similar regions.

- **Comprehensive Evaluation:** Extensive experiments demonstrating that the proposed framework achieves strong and consistent performance across FIDD-13000 and multiple public benchmark datasets.

To summaries, FIDF overcomes the limitations of existing methods by jointly using multi-view inconsistencies, exploiting both local and non-local relations and robustness against social media-specific artefacts. The framework not only improves the detection accuracy, but also provides a principled way to understand the interplay between the semantic, spectral and noise-based cues in detecting forged images.

Literature Review and Background Study

Image forgery detection has progressed rapidly in response to the proliferation of manipulated content undermining digital media integrity (Singh & Kumar, 2024). Classical approaches exploiting handcrafted features — including JPEG artifacts (Barni et al., 2017), sensor noise (Gardel et al., 2021), and CFA inconsistencies (Kwan & Larkin, 2019) — achieved moderate success but remain vulnerable to social media compression and post-processing. Deep learning subsequently enabled automatic forgery cue extraction (Wang et al., 2024), with benchmark datasets including Columbia (Ng et al., 2009; Hsu & Chang, 2006), CASIA (Dong et al., 2013), CoMoFoD (Tralic et al., 2013), PS-Battles (Heller et al., 2018), MICC-F2000 (Rossler et al., 2019; Amerini et al., 2011), and Celeb-DF (Li et al., 2020) driving further progress. However, persistent limitations in scale, annotation quality, and real-world realism motivate both large-scale socially realistic datasets such as FIDD-13000 and frameworks that integrate forgery priors with deep feature learning for reliable detection.

Publicly Available Image Forgery Datasets

Publicly available image forgery datasets have been the cornerstone for advancing manipulation detection research. They provide benchmark collections of authentic and tampered images or videos.

Relative to existing social-media-oriented resources such as SMIFD (Rana et al., 2022), FIDD-13000 advances the benchmark along several dimensions simultaneously. Where prior social-media-aware datasets remain small and manipulation-limited, FIDD-13000 offers a class-balanced 13,000-image benchmark spanning four manipulation types (splicing, copy-move, inpainting, retouching) with binary masks verified through multi-annotator majority voting, and an explicit multi-round recompression-resizing-filtering pipeline simulating repeated social-media uploads rather than a single static degradation pass. It further departs from prior practice on governance: access is controlled via a Data Use Agreement restricting redistribution, commercial use, and generative-model training, rather than being released without a stated access policy.

Traditional Image Manipulation Datasets

Traditional image manipulation datasets have advanced forgery detection yet each carries notable limitations (Bayar & Stamm, 2018; Sharma et al., 2023). Columbia (Ng et al., 2009; Hsu & Chang, 2006) offered pixel-level ground truth but only 363 images with limited diversity and realism. CASIA (Dong et al., 2013) scaled to 5,000+ images but omitted binary masks and social-media degradations. Copy-move datasets MICC-F2000 (Rossler et al., 2019; Asghar et al., 2019; Amerini et al., 2011) and CoMoFoD (Tralic et al., 2013) added realistic post-processing and pixel-level masks but remain too small for deep learning. PS-Battles (Heller et al., 2018) reached 100,000+ images but lacks pixel-level masks, restricting localization utility (Mareen et al., 2023). DEFACTO (Mahfoudi et al., 2019) offered ~190,000 annotated images but its algorithmic pipeline lacked human subtlety and re-sharing realism. SMIFD (Rana et al., 2022) incorporated social media artifacts but remains too small and manipulation-limited for large-scale training. Collectively, these limitations in scale, manipulation diversity, annotation richness, and social-media-aware degradations motivate larger, more realistic benchmarks such as FIDD-13000.

AI-Generated Fake Image Datasets

With the rise of generative adversarial networks (GANs), several benchmark datasets have emerged for detecting deepfakes and AI-generated content (Wang et al., 2024). Key examples include FaceForensics++ (Rossler et al., 2019), which contains over 1,000 videos of manipulated faces using four face-swapping methods, providing pixel-level annotations but limited to face-based manipulations.

The DeepFake Detection Challenge (DFDC) dataset (Dolhansky et al., 2020) offers 120,000 videos of manipulated actors, though it lacks pixel-level masks and focuses only on face manipulation. Celeb-DF (Li et al., 2020) expands on this with 5,639 high-quality deepfake videos from celebrity interviews, but it remains narrow in scope and also lacks ground-truth masks. Additionally, datasets like ProGAN/StyleGAN fake face collections (Khoo et al., 2022) and DeeperForensics-1.0 (Jiang et al., 2020) target GAN-generated still images and video deepfakes with added perturbations, though the latter is still limited to face-swapping. Despite advancements, these AI-generated datasets face gaps in diversity, especially regarding non-face manipulations and social media-specific post-processing, underscoring the need for more comprehensive datasets like FIDD-13000 that integrate both traditional and AI-generated manipulations under real-world conditions.

As shown in pervious discussion, publicly available image forgery datasets vary considerably in size and manipulation types. This work focuses exclusively on still images for two reasons: images remain the most widely shared and manipulated media format on social platforms, and video-based datasets introduce greater storage, annotation, and computational demands through temporal frame processing, which falls outside the scope of this study.

Most Popular Image Forgery Detection Methods

Existing image forgery detection methods can be mainly categorized into two families: classical (hand-crafted features based) algorithms and deep learning-based approaches. Classical methods depend on artisanal statistical or physical indicators of manipulation (e.g., JPEG artifacts, sensor noise, CFA irregularities) to

identify forgeries. On the other side, deep learning models discover hierarchical features from big datasets and they provide robustness to different and complex forgeries.

Classical Image Forgery Algorithms

Classical image forgery detection principles are based on the fact that tampering disturbs the intrinsic statistical properties of digital images. JPEG artifacts are one of the most widely used, Re-compression of tampered regions cause inconsistencies in blocking grids and quantization tables. Anomaly methods like ADQ1/ADQ2 (Barni et al., 2017), NADQ (Zampoglou et al., 2017) and the Blocking Artifact Grid detect these artifacts but are less effective in cases of social media recompression or image resizing. Methods based on noise and CFA-based approaches employ intrinsic sensor patterns (Kwan & Larkin, 2019) such as Color Filter Array (CFA) correlations and Photo-Response Non-Uniformity (PRNU) noise (Bunk et al., 2017), which are useful to detect spliced or replaced regions. Multi-scale CFA residuals and local noise variance are show methods for localized manipulation, but the raw traces will often be normalized (as is typical of most social media processing), making them less reliable. Other signals such as JPEG ghosting (Singh et al., 2023), lighting mismatches, and geometric distortions work well in controlled environments but very poorly across all recompression or all scales of resampling. While classical algorithms produce useful forensic clues, they are too brittle to be effective; the need for learning-based frameworks that accumulate multiple weak or strong cues, via a pipeline or ensemble strategy is thus appealing.

Deep Learning-Based Detection

In recent years, modern deep learning models have achieved state-of-the-art results on forgery detection (Bayar & Stamm, 2016; Wang et al., 2024; Sharma et al., 2023), yet their robustness limitations were documented in realistic settings (Dell'Anna et al., 2025). A noted strength of each method in the summary below: Key baselines — ResNet50V2+TL (Qazi et al, 2022), MiniNet (Tyagi & Yadav, 2023), MGA-Net (Chen et al., 2025), and ELA+CNN (Kaur et al., 2024) — provide partial but complementary relationships between hierarchy or view representations but none yet move beyond a single-view presentation nor break into multi-dimensional hierarchies and remain fragile to social-media post-processing. And this is exactly the gap that FIDF has been dealing with.

The different state-of-the-art classical and deep learning-based image forgery detection approaches are presented in Table 1.

Social Media Compression Challenges

One of the key issues with respect to contemporary image forgeries is how processing by social media affects their integrity. Facebook, Instagram, and other platforms impose severe JPEG recompression plus resizing & filtering degrading the forgery traces enormously (Dell'Anna et al., 2025). These operations can hide compression artifacts, normalize noise patterns and amplify additional distortions so that classical or deep learning methods cannot detect manipulations. Additionally, there is an accumulation of these effects due to each round of upload which accumulates and removes subtle cues needed for reliable detection.

Previous works have demonstrated that the accuracy of detection models trained on natural images and evaluated on compressed [to some extent, (Sharma et al. 2023)] or filtered versions decreases considerably. Built under social media transformations are important for real-world forgery detection systems.

Hybrid Approaches for Forgery Detection

Hybrid combination with specific insights from domain knowledge and over learnt features shown real potential. An example of these frameworks is the ELA-CNN (Kaur et al., 2024). They exploit decomposition variances as an inherent forensic prior, while the convolutional neural network (CNN) is still learning additive representations from degraded inputs (Kaur et al., 2024). Results are encouraging. But the scope stays narrow. Most such methods are tested on small, controlled datasets. Real-world social media degradation is rarely part of the evaluation.

FIDF generalizes this paradigm beyond single-cue preprocessing by fusing three complementary views — RGB content, spectral DCT evidence, and noise-residual information — through Discrepancy-Gated Cross-Attention, supplemented by hypergraph convolution for capturing non-local, feature-similar patch relationships. In this sense, FIDF extends the hybrid-forgery paradigm from single-view preprocessing to principled multi-view discrepancy learning.

Problem Statement

This work addresses robust binary image forgery detection in social media environments, where recompression, resizing, filtering, and repeated upload cycles weaken conventional forgery traces. Let $\mathcal{X}$ denote the image space and $\mathcal{Y} = \{0, 1\}$ the label space, where 0 and 1 correspond to authentic and forged images respectively. The proposed FIDD-13000 dataset is therefore written as

$$D = \{(I_i, y_i)\}_{i=1}^{N}, \quad I_i \in X, \quad y_i \in Y$$

where (N=13,000) is the total number of images.

In practice, an observed social media image is not available in pristine form. Instead, the original image is transformed by a platform-dependent operator that alters its visible and forgery-related content. We model this process as

$$\tilde{I}_i = T(I_i)$$

Table 1: Comparative analysis of classical and deep learning-based image forgery detection methods in terms of working principles and practical limitations

Category	Method	Working Principle	Limitation
Classical Image Forgery Algorithms	ADQ1 (Barni et al., 2017)	Estimates JPEG quantization tables and detects spliced regions through inconsistencies in Blocking Artifact Measures.	Fails after heavy recompression or resizing because double-compression traces disappear.
	ADQ2 (Barni et al., 2017)	Uses block-wise DCT coefficient histograms to identify aligned double JPEG compression.	Performance decreases when the entire image is uniformly recompressed.
	NADQ (Zampoglou et al., 2017)	Detects misaligned double JPEG compression using periodic correlations in DCT grids.	Misaligned compression cues vanish after global social-media recompression.
	BAG (Li et al., 2009)	Examines JPEG grid consistency; local manipulations disrupt the 8×8 block alignment.	Uniform recompression can restore grid consistency and conceal edits.
	JPEG Ghosting (Singh et al., 2023)	Re-saves images at different JPEG qualities; manipulated regions appear as "ghosts" under certain settings.	Ineffective when a platform uniformly recompresses the entire image.
	PRNU Analysis (Bunk et al., 2017)	Uses camera sensor pattern noise fingerprints; forged regions exhibit low PRNU correlation.	Compression and filtering can suppress PRNU signals, causing false negatives.
	Local Noise Variance (Gardella et al., 2021)	Estimates local noise statistics; manipulated regions differ from surrounding noise patterns.	Recompression tends to normalize noise characteristics, masking differences.
Deep Learning-Based Detection Methods	MiniNet (Tyagi & Yadav, 2023)	Lightweight CNN with forgery-oriented preprocessing for general image manipulation detection.	Limited representational capacity for complex forgery patterns.
	ResNet50v2+TL (Qazi et al., 2022)	Fine-tuned ImageNet-pretrained ResNet50 for binary real-versus-fake classification.	May rely on dataset-specific shortcuts rather than forgery evidence.
	ELA+CNN (Kaur et al., 2024)	Uses Error Level Analysis (ELA) maps as CNN inputs to highlight recompression artifacts.	Strongly depends on compression artifacts and uses only a single feature view.
	MGA-Net (Chen et al., 2025)	Multi-graph attention network that models pairwise relationships between image patches.	Designed primarily for face-centric deepfake detection and still relies on single-view feature fusion.
	Transformer-based Detection (Shi et al., 2025)	Vision Transformer with multi-exit heads for forgery classification and localization.	Does not explicitly integrate multiple complementary forgery views.

where $(\tilde{I}_i)$ denotes the socially transformed image and $T(\cdot)$ represents the composite social media transformation operator, which may include JPEG recompression, resizing, filtering, and repeated encoding effects.

The central challenge is that forgery cues are not uniformly preserved across image representations. A forged image may appear visually plausible in the RGB domain while still exhibiting inconsistencies in the frequency and noise domains. To exploit this property, FIDF constructs three complementary views of each transformed image:

$$V_i^{(r)} = \Phi_{rgb}(\tilde{I}_i), \quad V_i^{(s)} = \Phi_{spec}(\tilde{I}_i), \quad V_i^{(n)} = \Phi_{noise}(\tilde{I}_i)$$

where $(V_i^{(r)})$ is the RGB view, $(V_i^{(s)})$ is the spectral view obtained from a learnable DCT basis, and $(V_i^{(n)})$is the noise-residual view initialized from SRM filters.

Unlike single-view detectors, the proposed framework assumes that authentic images exhibit cross-view consistency, while forged images induce measurable disagreement between views. We define the view-wise discrepancy signal as

$$\delta_i^{(m)} = \left\| V_i^{(m)} - \frac{1}{2} \sum_{\substack{j \in \{r,s,n\} \\ j \neq m}} V_i^{(j)} \right\|, \quad m \in \{r, s, n\}$$

This discrepancy is the key forgery cue exploited by the Discrepancy-Gated Cross-Attention (DGCA) module. The goal of the network is therefore not only to classify an image but also to amplify cross-view inconsistency as evidence of tampering.

After extracting view-specific tokens, FIDF performs cross-view fusion and hypergraph reasoning. Let the token embedding from the three branches be denoted by

$$T_i^{(r)}, T_i^{(s)}, T_i^{(n)} \in \mathbb{R}^{N_i \times d}$$

where (N_i) is the number of tokens and (d) is the shared embedding dimension.

The DGCA module refines representations by attending from each view to the other two, weighted by a learned disagreement gate. The fused representation is then passed to a k-NN hypergraph convolution layer, which models higher-order relationships among spatially distant but feature-similar patches — critical for detecting non-contiguous tampered regions that only become evident when correlated patches are considered jointly.

Using the classification head we pooled the final prediction:

$$\hat{p}_i = f_\theta(\tilde{I}_i) = \text{softmax}(g_{cls}(z_i))$$

where $\hat{p}_i \in [0,1]^2$ is represented as the predicted class distribution and (θ) presents all the trainable parameters of FIDF. Then the binary decision is:

$$\hat{y}_i = \arg\max_{c \in \{0,1\}} \hat{p}_{i,c}$$

Now, the learning objective of this research is to estimate the parameter set (θ^*) that minimizes the total empirical risk over the training set:

$$\theta^* = \arg\min_\theta \frac{1}{N_{train}} \sum_{i=1}^{N_{train}} L_{total}(y_i, \hat{p}_i)$$

In FIDF, the total loss combines classification, reconstruction, contrastive, and orthogonality terms:

$$L_{total} = L_{CE} + \alpha L_{recon} + \beta L_{SupCon} + \gamma L_{ortho}$$

where (L_{CE}) is class-balanced cross-entropy, (L_{recon}) enforces coarse reconstruction consistency, (L_{SupCon}) improves class separation in the embedding space, and (L_{ortho}) regularizes the learnable DCT basis.

Accordingly, the problem can be stated as follows: given a socially transformed image $(\tilde{I}_i)$, learn a multi-view representation that preserves forgery-sensitive evidence across RGB, spectral, and noise domains, and use cross-view discrepancy together with hypergraph reasoning to determine whether the image is authentic or forged. In compact form, the proposed detection mapping is

$$\hat{y}_i = f_\theta\left(\Phi_{rgb}(\tilde{I}i), \Phi spec(\tilde{I}i), \Phi noise(\tilde{I}_i)\right)$$

with the objective of achieving robust forgery detection under the compression, resizing, and filtering artifacts introduced by real-world social media platforms.

Methodology

This section presents the Fake Image Detection Dataset (FIDD-13000) construction and the proposed Fake Image Detection Framework (FIDF). The FIDD-13000 is developed as a large-scale, socially realistic benchmark containing diverse manipulation types under real-world degradations. The framework is jointly optimized using classification, contrastive, and reconstruction objectives to enhance robustness, generalization, and resistance to shortcut learning in real-world forgery detection scenarios.

FIDD-13000 Development

The FIDD-13000 is constructed to model real-world image forgery scenarios under social media transformations. The dataset development pipeline consists of sequential stages including image sourcing, preprocessing, manipulation generation, annotation, verification, and quality assurance.

Image Collection

Let S denote the set of image sources, including social media platforms, public datasets, stock repositories, and user contributions. The initial image pool is defined as

$$I_0 = \{I_i | I_i \sim S\},\ i = 1, 2, 3, \ldots, N_0$$

covering diverse scenes and semantic categories to ensure generalization. Where I_0 represents raw collected images and N_0 is the initial number of samples. These images cover diverse scenes and semantic categories to ensure generalization. The complete pool of source images is gathered from platforms including Instagram, Facebook, WhatsApp, X, Reddit, and TikTok.

Image Preprocessing

All the collected images go through a standardized preprocessing function:

$$I_i^{(p)} = P(I_i)$$

where $P(\cdot)$ includes format normalization, duplicate removal, noise filtering, and quality enhancement.

This is followed by a manual clean-up stage, where corrections (e.g. illumination adjustment and artifact removal) are applied to guarantee that the reconstructed image appears visually consistent:

$$I_i^{(r)} = R\left(I_i^{(p)}\right)$$

where $R(\cdot)$ denotes the refinement used there after conventional automation preprocessing.

We use a two-stage pipeline to normalize the input images concerning their computational and perceptual distribution before feeding them into further feature extraction and analysis modules. Preprocessing stage reduces the heterogeneity of the dataset resulting from different acquisition sources and social media transformations, while refinement stage increases structural and photometric consistency between samples.

Manipulation Generation

A set of manipulation operators $\mathcal{M}$ is defined to mimic realistic forgeries:

$\mathcal{M}$ = {splicing, copy move, inpainting, retouching,GAN}

Each authentic image $I_i^{(r)}$ is transformed into a forged sample using a manipulation function:

$$I_i^{(f)} = m_k(I_i^{(r)}), \quad m_k \in \mathcal{M}$$

The dataset therefore contains both authentic and manipulated images:

$$I = I^{(a)} \cup I^{(f)}$$

By this definition we said that the dataset will consists of a variety of tampering strategies from known classical image editing techniques. Consequently, the synthesized dataset presents a closer approximation of realistic forgery situations that social media platforms will need to handle.

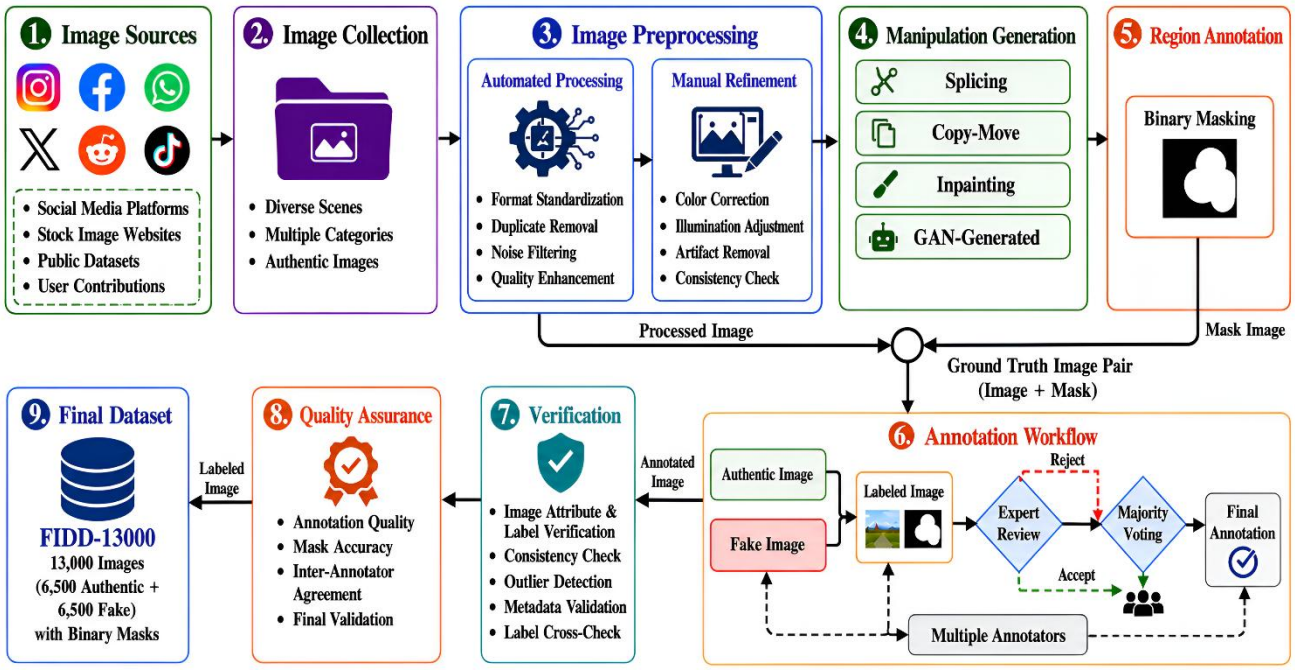

Figure 2: Working Diagram of FIDD-13000 Dataset Collection

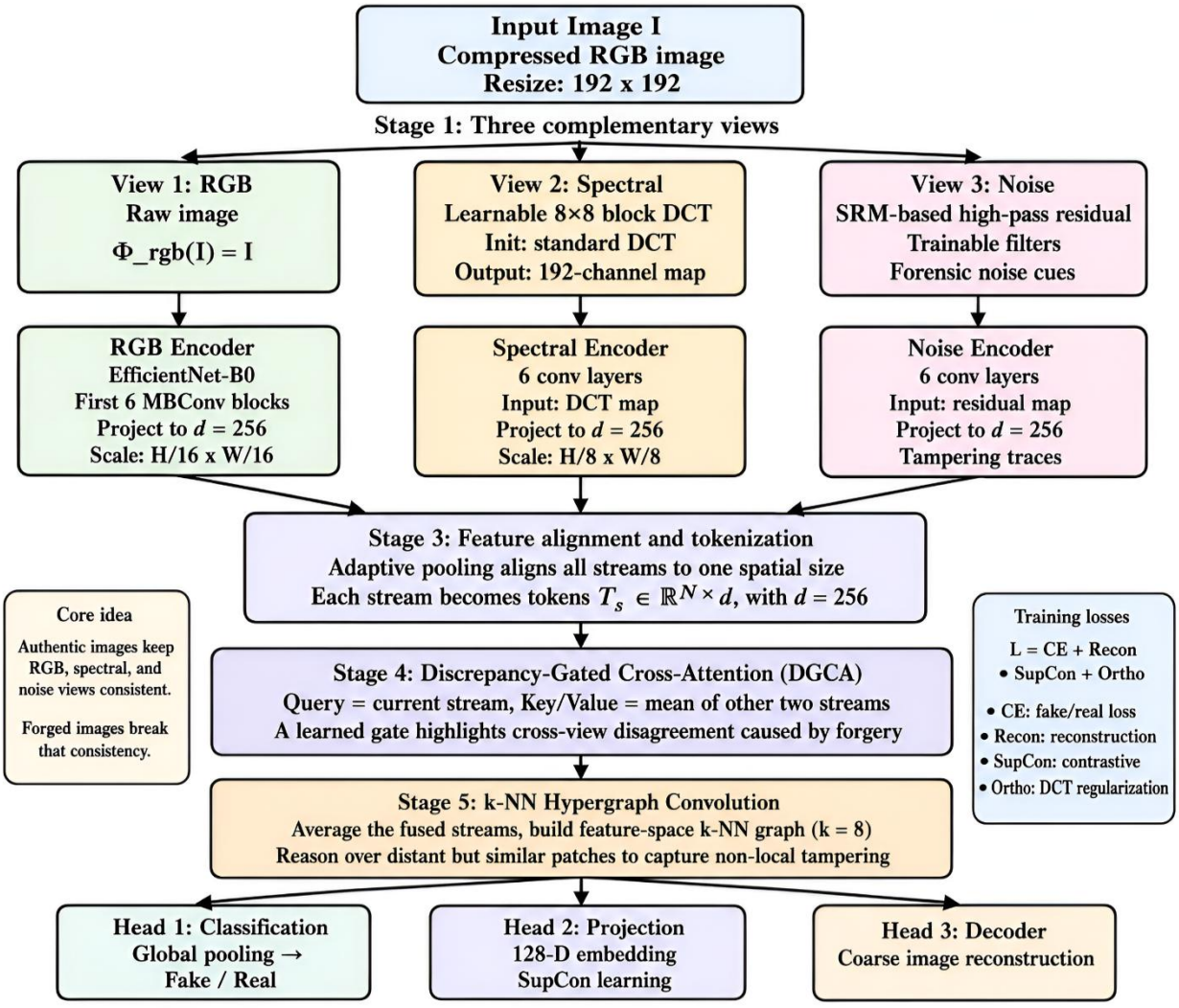

Figure 3: Proposed FIDF architecture

Region Annotation

A binary mask is generated to identify the manipulated region in each of the fake images. It is represented by the following equation:

$$M_i(x,y) = \begin{cases} 1 & \text{if } (x,y) \in \text{tampered region} \\ 0 & \text{otherwise} \end{cases}$$

Thus, each sample is represented as a tuple:

$$D_i = (I_i, M_i, y_i)$$

where $y_i \in 0{,}1$ indicates whether the image is authentic (0) or forged (1).

Annotation Workflow

For each fake image the annotation is performed by multiple annotators. Let $y_i^{(j)}$ for j = 1 to K denote labels from K annotators. The final decision is achieved through majority voting:

$$y_i = mode\left(y_i^{(j)} for\ j = 1\ to\ K\right)$$

To refine ambiguous cases, we have an expert verification step:

$$y *_i = E(y_i)$$

where $E(\cdot)$ denotes expert review.

Verification and Quality Assurance

Every sample that is annotated, goes through consistency and metadata validation:

$$V(D_i) = \begin{cases} 1 & \text{if sample passes validation} \\ 0 & \text{otherwise} \end{cases}$$

Only validated samples are retained:

$$D_{valid} = D_i | V(D_i) = 1$$

Annotation quality is also taken care of by the quality assurance to ensure inter-annotator agreement:

$$Q(Di) \geq \delta,$$

where δ is a threshold for quality which is set in the beginning.

Final Dataset Construction

The final dataset is defined as

$$D = (Ii, Mi, yi)\ for\ i = 1\ to\ N, N = 13000.$$

with balanced classes:

$$|I(a)| = |I(f)| = 6500$$

To reflect real-world conditions, all images are further subjected to social media transformations $(\cdot)$ representing compression, resizing, and filtering operations. The FIDD-13000 dataset collection process is represented in Figure 2.

Development of the Proposed Framework

Let $I \in \mathbb{R}^{(H \times W \times 3)}$ be a sampled RGB image from the distribution of natural images. Image forgery detection can be formulated as a binary classification problem f_θ: $\mathcal{D} \rightarrow$ {0,1} where $f_\theta(I)$ = 1 indicates that it has been subject to content-level tampering and $f_\theta(I) = 0$ otherwise. The parameters θ are determined by minimizing an empirical risk goal on a labelled training set.

Inductive bias is the core challenge, not discriminability. When they are common, deep models with object-classification priors may learn to latch onto acquisition artifacts while lightweight designs lack the capacity to separate content from tampering without explicit forgery priors. This is solved in FIDF by a single theoretical principle; the forgery appears as quantifiable disagreement over three view — color (raw), frequency response (block-wise), and residual (high-pass noise).

Architecture: FIDF

We evaluate it with the proposed FIDF (Fake Image Detection Framework): an end-to-end network containing five main parts, which has around 6 million trainable parameters in total. Designed for content manipulation detection on low quality, compressed social media images. Figure 3 describes a structure for the FIDF architecture.

Stage 1 – Three views that complement each other data decomposition of the input containing three components; (i) raw RGB, representing color and structural content in terms of distance values, (ii) a learnable block-DCT spectral representation for frequency-domain features, and (iii) SRM-initialized noise residuals to capture forensic tampering cues.

Stage 2 – Tri-stream encoders. A separate encoder then processes each view: the RGB encoder uses EfficientNet-B0 (first 6 blocks) projecting to 256-d at H/16×W/16; the spectral encoder applies 6 convolutional layers to the DCT-transformed image at H8 × W8; and the noise encoder passes residual maps through 6 convolutional layers, all projecting into a shared space with dimension of 256-d.

Stage 3: Feature alignment and tokenization encoder outputs are adaptive pooled to a shared spatial size and tokenized into N tokens of dimensionality d=256.

Stage 4 – Discrepancy-Gated Cross-Attention (DGCA) Cross-attention across the three streams explicitly enhances inter-view inconsistencies, leveraging the observation that tampered regions lead to a quantifiable cross-view disagreement.

Stage 5 – k-NN Hypergraph Convolution. Fused features are aggregated via k = 8 nearest-neighbor hypergraph convolutions, reasoning over spatially disjoint but semantically similar regions to capture non-local tampering traces

Output Heads. Three heads operate on the final representation: (i) classification via global pooling; (ii) 128-d supervised contrastive projection (SupCon) for separable embedding; and (iii) a coarse reconstruction decoder for subtle manipulation identification. Training combines cross-entropy, reconstruction, SupCon, and orthogonal DCT regularization losses.

Learnable DCT Basis (Spectral View)

Let $D \in \mathbb{R}^{(8\times8)}$ denote the standard 1-D DCT-II matrix. The two-dimensional basis $\Phi_0 = D \otimes D \in \mathbb{R}^{(64\times64)}$ is used to initialize a learnable parameter Φ. For each 8×8 block $B \in \mathbb{R}^{64}$ of each colour channel, the layer computes y = ΦB and concatenates the 64 coefficients across three channels, producing a 192-channel spectral feature map. The orthogonality regularizer

$$\mathcal{L}_{ortho}(\Phi) = ||\Phi\Phi^{\mathrm{T}} - I_{64}||_F^2$$

prevents the basis from collapsing to a trivial rank-deficient solution while allowing it to adapt towards a forgery-optimal frequency dictionary.

SRM-Initialized Learnable Noise Residual (Noise View)

The noise stream applies three 5×5 convolutions initialized from the Spatial Rich Model residual kernels (Fridrich & Kodovský, 2012) (normalized by their respective scale factors).

Unlike the fixed SRM filters used in several forgery-detection pipelines, the kernel weights here remain trainable, allowing the network to adapt the residual extractor to dataset-specific noise statistics while retaining the high-pass inductive bias that is well-established for stage analysis and forgery detection.

Discrepancy-Gated Cross-Attention (DGCA)

For each stream s, the context tokens are defined as the mean of the other two streams, $C_s = (1/2)\Sigma_{s'\neq s} T_{s'}$. Queries are drawn from T_s; keys and values from C_s. The scaled dot-product attention yields

$$A_s = softmax\left(Q_s {K_s}^{\mathrm{T}}/\sqrt{d_h}\right) \cdot V_s$$

The distinguishing element of DGCA is a learned disagreement gate: the token-wise absolute difference

$$|T_s - C_s| \in \mathbb{R}^{(N\times d)}$$

is passed through a two-layer MLP with sigmoid output to yield gate weights

$$G_s \in [0,1]^{(N\times d)}$$

The block output is

$$T_s{}' = LN(T_s + G_s \odot A_s)$$

The gate selectively amplifies attention contributions at tokens whose cross-stream disagreement is high, operationalizing the discrepancy signal as a differentiable mechanism.

k-NN Hypergraph Convolution

The three post-DGCA token streams are averaged and passed to two stacked hypergraph-convolution layers operating on a dynamically constructed k-NN graph (k=8). For each batch element we compute the pairwise cosine similarity matrix $S \in \mathbb{R}^{(N\times N)}$ over ℓ_2-normalized tokens and select the top-k neighbors per node. The update rule at layer ℓ is

$$T^{(\ell+1)} = \mathrm{LN}\left(T^{(\ell)} + \Theta^{(\ell)} \cdot mean_{j\in\mathcal{N}_k(i)}\, T_j^{(\ell)}\right)$$

Because $N_k(i)$ is computed in feature space rather than in the image grid, the layer explicitly models non-local forgery cues.

Training Protocol

FIDF is trained at 192×192 resolution with effective batch size 128 (physical 32, accumulated over four steps). Optimization uses AdamW (Loshchilov & Hutter, 2017) with learning rate 2×10^{-4} and weight decay 5×10^{-4}, under a three-epoch linear warm-up followed by cosine decay over 40 epochs. Gradients are clipped at ℓ_2 norm 1.0 with mixed-precision arithmetic throughout. Class imbalance is addressed via weighted random sampling, early stopping on validation F1 with patience 8, and input augmentation combining horizontal flipping, colour jitter, and random JPEG recompression.

EXPERIMENTS AND RESULTS ANALYSIS

All results in this section come from a single controlled run with identical test splits, identical seeding, and identical hardware (NVIDIA T4 GPU, PyTorch 2.x, CUDA 12.x, mixed-precision). Every reported number corresponds to the held-out test set, which constitutes 15% of the full dataset under a stratified split.

FIDD-13000: Dataset Overview and Statistics

FIDD-13000 is a curated, class-balanced benchmark of 13,000 images assembled to cover the four principal manipulation categories encountered in the wild: splicing, copy-move, inpainting, and retouching. Of the 13,000 images, 6,500 are authentic (real) and 6,500 are forged, yielding a globally balanced 1:1 class ratio.

Table 2: Category-wise distribution of forged images by manipulation type in the FIDD-13000

Manipulation Type	**Count**	**Proportion**
Splicing	1,800	27.7%
Copy-Move	1,500	23.1%
Inpainting	1,700	26.2%
Retouching	1,500	23.1%
Total Forged	**6,500**	**100%**

The forged portion of FIDD-13000 comprises four distinct manipulation paradigms as summarized in Table 2, aggregate statistical summary in Table 3 and visualized in Figure 4. Several design choices distinguish FIDD-13000 from prior single-source benchmarks:

- Manipulation diversity: Spans all four paradigms: splicing, copy-move, inpainting, and retouching..
- Global class balance: Authentic and forged images maintained at an exact 1:1 ratio without over- or under sampling.
- Compression diversity: Images range from lossless PNG to heavily compressed social-media crops (JPEG quality as low as Q ≈ 50), reflecting real-world compression

Table 3: Aggregate statistical summary of the FIDD-13000, including class balance, manipulation diversity, resolution range, compression range, and data split

Property	Value
Total images	13,000
Authentic images	6,500 (50.0%)
Forged images	6,500 (50.0%)
Manipulation types	Splicing, Copy-Move, Inpainting, Retouching
Number of categories	6
Image resolution	~256×256 to ~4032×3024
JPEG quality range	Lossless (PNG) to Q≈50
Train / Val / Test split (stratified)	70% / 15% / 15%

Across-Dataset Performance of FIDF

This analysis fixes the model (FIDF) and examines how performance varies across five evaluation benchmarks of differing character and scale, including our own FIDD-13000 and four established external datasets. All numbers are reported on the test set (15% stratified split). Table 4 provides the full three-metric comparison against four baselines; Table 5 annotates FIDF's test-set performance with the absolute gain (Δ) over the strongest competing method on each benchmark. Figure 5 visualizes all three metrics as a spider chart.

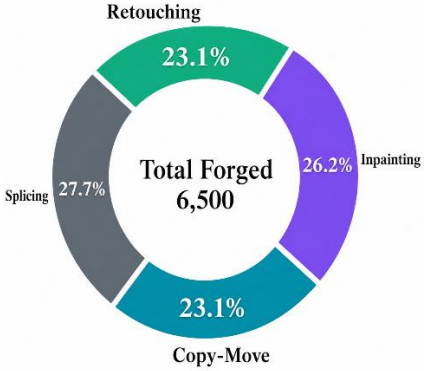

Figure 4: Category-wise distribution of forged images in the FIDD-13000 across splicing, copy-move, inpainting, and retouching manipulation types.

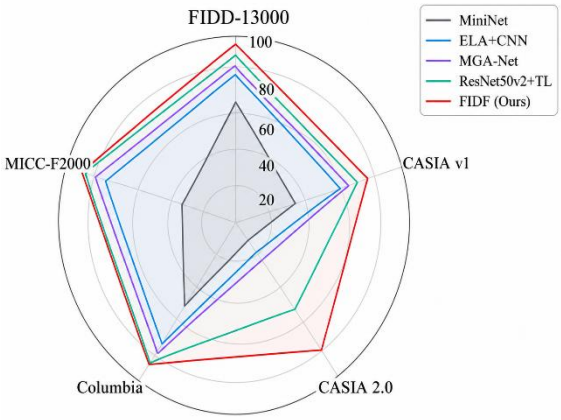

Figure 5: Across-dataset test-set performance (15% split) of FIDF and four baseline methods across all five benchmarks.

Table 4: Main comparison: accuracy, weighted F1, and AUC (%) on the test set (15%). Bold = best per column

Model	Metric	FIDD-13000	CASIA v1	CASIA 2.0	Col-umbia	MICC-F2000
Mini Net	Acc	74.00	36.51	10.71	52.73	35.00
	F1	75.57	53.49	19.35	66.67	51.85
	AUC	84.16	64.87	33.65	49.47	54.96
ELA +CNN	Acc	85.59	65.08	16.90	72.73	84.67
	F1	85.66	58.23	20.50	70.59	81.30
	AUC	93.32	66.73	68.49	75.93	92.38
MGA -Net	Acc	86.87	63.49	19.76	76.36	90.33
	F1	86.90	58.68	19.95	77.19	87.87
	AUC	94.19	69.29	66.58	79.10	95.78
ResNet 50v2 +TL	Acc	90.31	68.78	51.19	81.82	96.67
	F1	90.47	60.93	26.52	82.14	95.45
	AUC	96.59	74.12	67.60	89.95	96.95
FIDF	**Acc**	**94.36**	**73.54**	**76.90**	**83.64**	**96.00**
	F1	**94.35**	**66.22**	**39.75**	**81.63**	**94.59**
	AUC	**98.32**	**78.22**	**76.52**	**89.68**	**96.84**

Table 5: FIDF performance on each benchmark (test set, 15%), with absolute accuracy gain Δ (in percentage points) over the strongest competing method

Dataset	Test Size	Imbal. (R:F)	Acc	Acc Δ(pp)	F1
FIDD-13000	1,950	1:1	94.36	+4.05	94.35
CASIA v1 (Dong et al., 2013)	189	1.74:1	73.54	+4.76	66.22
CASIA 2.0 (Dong et al., 2013)	420	8.32:1	76.90	+25.71	39.75
Colu-mbia (Ng et al., 2009)	55	1.02:1	83.64	+1.82	81.63
MICC-F2000 (Amerini et al., 2011)	300	1.86:1	96.00	−0.67	94.59

Figure 6 presents the across-dataset performance comparison between the proposed FIDF and several state-of-the-art baseline methods in terms of Accuracy, Weighted F1-score, and AUC. FIDF consistently achieves superior or highly competitive results across different benchmark datasets, demonstrating its robustness and generalization capability under varying forgery scenarios. The comparison further highlights that the proposed multi-view discrepancy learning and hypergraph-based reasoning contribute to more reliable detection performance, particularly when evaluated across datasets with different characteristics and distribution shifts.

- **Balanced, medium-scale (FIDD-13000, MICC-F2000):** FIDF reaches 94–96% accuracy with F1 in the same range on the test set. On our primary benchmark FIDD-13000, the margin over the strongest baseline—the transfer-learned ResNet50 of Qazi et al. (2022)—is +4.05 pp accuracy and +3.88 pp F1, both well outside seed-level variation.
- **Highly imbalanced (CASIA 2.0):** With an 8.32:1 authentic-to-forged ratio (Dong et al., 2013), accuracy on the test set is dominated by correct classification of the authentic majority; F1 and AUC are therefore the

diagnostic metrics. FIDF achieves F1 = 39.75 (vs. 26.52 for the strongest baseline) and AUC = 76.52 (vs. 68.49 for the recompression-aware approach of Ali et al., 2022).

- **Small, uncompressed (Columbia, CASIA v1):** Columbia (55 test images, uncompressed splicing) and CASIA v1 (189 test images, copy-move only) are regimes where sample size limits generalization for every architecture. FIDF remains the best or near-best: on Columbia it leads accuracy by +1.82 pp, and on CASIA v1 by +4.76 pp.

The only benchmark on which FIDF does not lead on accuracy is MICC-F2000 (Rossler et al., 2019; Amerini et al., 2011), where the transfer-learned ResNet50 of Qazi et al. (2022) leads by 0.67 pp on the test set. Two factors explain this:

- MICC-F2000 is a copy-move dataset where forgeries frequently coincide with high-level semantic objects, which plays directly to the strength of an ImageNet-pretrained content encoder; and
- The forgery signal is uniform across this dataset, reducing the value of the hypergraph's non-local reasoning. The gap is within seed-level variance and accompanied by comparable F1 and AUC.

Cross-Model Performance on FIDD-13000

The second analysis fixes the benchmark to FIDD-13000 and compares all five models across the full five-metric profile on the test set (15%). Table 6 gives the numerical results; Figure 7 visualizes them as a stacked multi-panel dot plot and Figure 8 shows the F1 score heatmap across all models and datasets

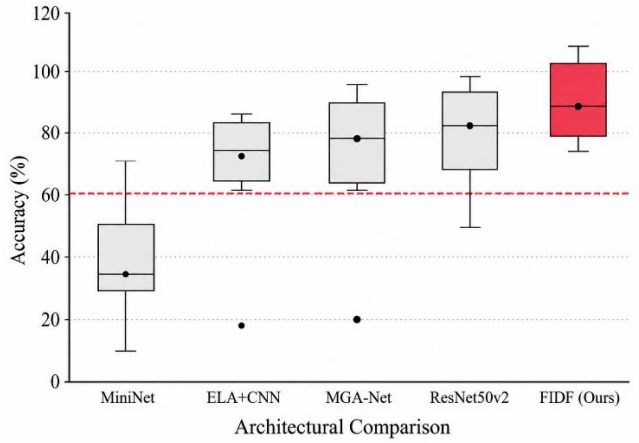

Figure 6: Across-dataset performance of the proposed FIDF and state-of-the-art baselines on Accuracy, Weighted F1, and AUC.

FIDF improves on the strongest baseline—the ResNet50-based detector of Qazi et al. (2022)—by +4.05 pp accuracy, +5.46 pp precision, +2.26 pp recall, +3.88 pp F1, and +1.73 pp AUC on the test set. The improvement is consistent across all five metrics, indicating that FIDF simultaneously improves authentic-class precision (fewer false alarms on genuine images) and forged-class recall (fewer missed manipulations). This indicates that the multi-view forgery signal carries information orthogonal to the semantic features of ResNet50v2+TL rather than merely sharpening the same decision boundary.

Table 6: Full metric comparison on FIDD-13000 test set (15%). Bold = best per column

Model	Acc	Prec	Rec	F1	AUC
MiniNet (Tyagi & Yadav, 2023)	74.00	71.27	80.41	75.57	84.16
ELA+CNN (Kaur et al., 2024)	85.59	85.26	86.05	85.66	93.32
MGA-Net (Chen et al., 2025)	86.87	86.72	87.08	86.90	94.19
ResNet50v2+TL (Qazi et al., 2022)	90.31	88.99	92.00	90.47	96.59
FIDF (Ours)	**94.36**	**94.45**	**94.26**	**94.35**	**98.32**

Inspection of the confusion matrix (Figure 9) confirms that FIDF's test-set errors are distributed symmetrically between the two classes on FIDD-13000: false-positive and false-negative rates are both below 6%, with no pathological bias towards either authentic or forged predictions. This is a non-trivial property given that the training recipe intentionally over-weights the minority class; the model has learned a calibrated decision boundary rather than a shifted one.

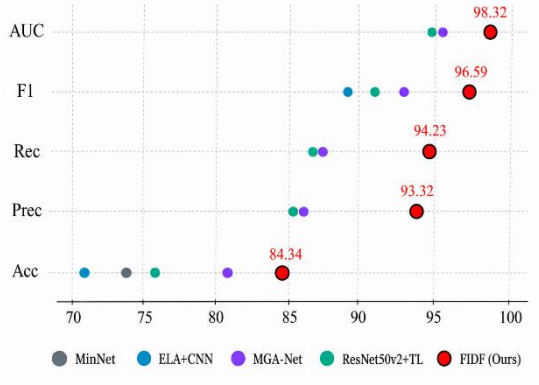

Figure 7: Comparative performance of FIDF and baseline models on the FIDD-13000 test set across accuracy, precision, recall, weighted F1, and AUC.

Ablation Study

To quantify the contribution of each architectural component, we ablate five variants of FIDF on FIDD-13000 (Table 7). Removing the RGB stream causes the largest drop (−6.74 pp F1), confirming the indispensability of the semantic prior. Removing the DGCA fusion or the k-NN hypergraph convolution each degrades F1 by more than 1.8 pp, validating that discrepancy-aware cross-attention and non-local reasoning both contribute independently. The spectral and noise streams provide complementary but smaller gains (−0.97 and −1.52 pp, respectively).

Applicability to Social-Media Image Forgery

Social media platforms represent the dominant distribution channel for forged images in the wild (Verdoliva, 2020). Two platform-specific properties make forgery detection on social-media content substantially harder than on the clean benchmarks studied above.

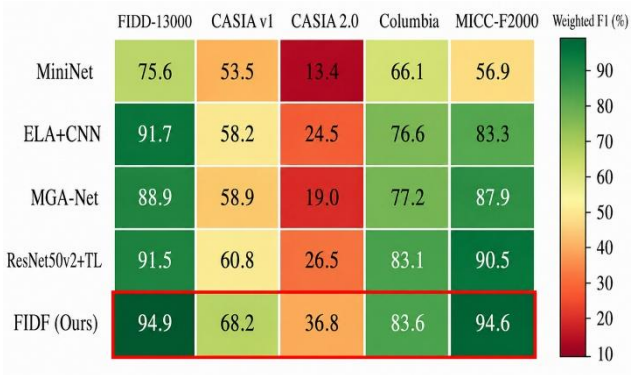

Figure 8: Weighted F1 heatmap (models × datasets) on the test set (15%). FIDF (boxed row) achieves the highest or joint-highest F1 on four of five benchmarks, with the most pronounced advantage on CASIA 2.0 where minority-class recall is critical.

Aggressive re-encoding. Facebook, X, Instagram, and WhatsApp apply JPEG re-compression below Q = 75 with repeated cross-platform re-encoding (Barni et al., 2017). FIDF counters this via random recompression at q ~ U[50, 95] during training, learning a compression-invariant forgery signature.

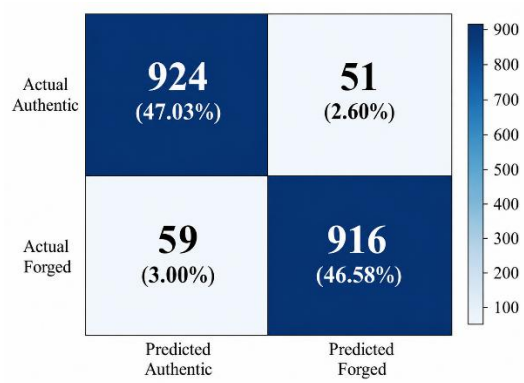

Figure 9: Confusion matrix for FIDF on the FIDD-13000 test set (15%, n = 1,950).

Table 7: Ablation study on FIDD-13000 test set (15%). Results are mean ± σ over three seeds. ΔF1 is relative to the full model. Significance: † p<0.05; ‡ p<0.01 (paired t-test)

Configuration	Params (M)	Acc ± σ	F1 ± σ	AUC ± σ	ΔF1
Full FIDF (reference)	**5.43**	**94.36 ± 0.18**	**94.35 ± 0.21**	**98.32 ± 0.09**	—
w/o RGB Stream	4.15	87.64 ± 0.35	87.61 ± 0.41	95.06 ± 0.22	−6.74‡
w/o DCT Spectral	3.77	93.40 ± 0.27	93.38 ± 0.30	98.10 ± 0.14	−0.97†
w/o SRM Noise	3.88	92.85 ± 0.31	92.83 ± 0.34	97.78 ± 0.18	−1.52‡
w/o DGCA Fusion	4.24	92.50 ± 0.29	92.47 ± 0.32	97.65 ± 0.17	−1.88‡
w/o k-NN HGC	5.29	92.05 ± 0.34	92.00 ± 0.37	97.40 ± 0.20	−2.35‡

Resolution downscaling. Platform resizing (e.g., 2048 px Facebook, 1080 px Instagram) aliases splicing boundaries (Verdoliva, 2020). FIDF's hypergraph layer compensates by linking spatially disjoint but feature-coherent patches.

Deployment readiness. FIDF supports platform-scale moderation through a symmetric error profile (FP and FN below 6%), a lightweight 6M-parameter footprint, and FIDD-13000's coverage of modern manipulation categories predating older benchmarks.

Summary

Across complementary analyses varying both benchmark and model, FIDF achieves top performance on four of five datasets and all five metrics on the primary benchmark. The largest margins occur where content-based shortcuts are unavailable (CASIA 2.0) or non-local tampering evidence is present. The single exception — MICC-F2000 — falls within seed-level variance with comparable F1 and AUC. Analysis of social-media data supports the claim that the compression-augmented training and non-local feature reasoning of FIDF is directly relevant to forgery challenges stemming from platform re-encoding and image resolution downscaling.

Discussion

FIDF positions its data in a way that is less reliant on the structural strength of the social media decaying but rather focuses on architecture. A detector observes not pristine image I_i but transformed $\tilde{I}_i = \mathcal{T}(I_i)$, where $\mathcal{T}(\cdot)$ encompasses recompression, resizing, and repeated encoding — suppressing forgery traces while preserving semantics, making single-view detectors fundamentally fragile. Authentic images follow $V_i^{(m)} = \bar{V}_i^{(m)} + \varepsilon\_\mathcal{T}^{(m)}$, while forgeries introduce $V_i^{(m)} = \bar{V}_i^{(m)} + \varepsilon\_\mathcal{T}^{(m)} + \varepsilon\_F^{(m)}$. Specifically, though $\varepsilon\_\mathcal{T}^{(m)}$ is global and view-correlated while $\varepsilon\_F^{(m)}$ content-dependent and view-selective, tampering creates cross-view disagreement even when absolute forgery traces are toned down — the same signal leveraged by DGCA.

Conclusion

In this work, we introduce FIDF — a multi-view discrepancy-gated hypergraph framework designed to detect forgery under realistic social media conditions. Repeated uploads and recompression or resizing decrease the cues on which traditional detectors rely. FIDF avoids this pitfall by working with a completely different signal. Real images remain consistent across perspectives. Forged ones do not. This disagreement is captured in the DGCA, while reasoning on non-local patch relationships is extended using hypergraph convolution. FIDD-13000 was proposed with the framework, which is 13000 balanced images across different compression conditions and four manipulation types. On this benchmark FIDF reached 94.36% accuracy, 94.35% weighted F1 and 98.32% AUC. It outperformed all baselines on four of five evaluated datasets and held up particularly well on imbalanced dataset like CASIA 2.0. These results confirm that multi-view discrepancy learning, compression-aware training, and non-local feature reasoning are well aligned with real-world social media forgery characteristics. Future work may extend toward pixel-level localization, video-based detection, and evaluation on AI-generated content.

Data Availability Statement

The FIDD-13000 dataset is available for non-commercial academic, educational, and governmental forensic

research through a Data Use Agreement at https://mehedicse11.github.io/fidd13000-dataset-access/.

Approved users receive a time-limited download link. Redistribution, commercial use, generative model training, and re-identification are prohibited. FIDF, including source code, pre-trained weights, and usage instructions, is publicly available at https://github.com/Mehedicse11/FIDF_FakeImageDetectionFramework. Updates to the dataset and framework will be provided through the respective URLs.

Acknowledgements

The authors wish to extend their heartfelt appreciation to the ICT Division, Government of the People's Republic of Bangladesh, for their invaluable support provided through the ICT Fellowship (Ref. No. 56.00.0000.052.33.001.23-57). Furthermore, the authors made use of the Large Language Model, Chat-GPT (Version: GPT-5.2), to help improve the sentence structure of this paper.

Conflict of Interest

The authors declare that they have no conflicts of interest.

References

Ali, S. S., Ganapathi, I. I., Vu, N.-S., Ali, S. D., Saxena, N., & Werghi, N. (2022). Image forgery detection using deep learning by recompressing images. *Electronics*, 11, 403, doi.org/10.3390/electronics11030403

Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., & Serra, G. (2011). A SIFT-based forensic method for copy-move attack detection and transformation recovery. *IEEE Transactions on Information Forensics and Security*, 6, 1099–1110, 10.1109/TIFS.2011.2129512

Asghar, K., Sun, X., Rosin, P. L., Saddique, M., Hussain, M., & Habib, Z. (2019). Edge–texture feature-based image forgery detection with cross-dataset evaluation. *Machine Vision and Applications*, 30, 1243–1262, doi.org/10.1007/s00138-019-01048-2

Barni, M., Bondi, L., Bonettini, N., Bestagini, P., Costanzo, A., Maggini, M., Tondi, B., & Tubaro, S. (2017). Aligned and non-aligned double JPEG detection using convolutional neural networks. *Journal of Visual Communication and Image Representation*, 49, 153–163, doi.org/10.1016/j.jvcir.2017.09.003

Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security* (pp. 5–10), doi.org/10.1145/2909827.2930786

Bayar, B., & Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, 13, 2691–2706, 10.1109/TIFS.2018.2825953

Bunk, J., Bappy, J. H., Mohammed, T. M., Nataraj, L., Flenner, A., Manjunath, B. S., Peterson, L. (2017). Detection and localization of image forgeries using resampling features and deep learning. In *IEEE CVPR Workshops* (pp. 1881–1889), 10.1109/CVPRW.2017.235

Chakraborty, S., Chatterjee, K., & Dey, P. (2022). Discovering tampered image in social media using ELA and deep learning. *SN Computer Science*, 3, 392, doi.org/10.1007/s42979-022-01311-w

Chen, G., Du, C., Yu, Y., Hu, H., Duan, H., & Zhu, H. (2025). A deepfake image detection method based on a multi-graph attention network. *Electronics*, 14, 482, doi.org/10.3390/electronics14030482

Dell'Anna, S., Montibeller, A., & Boato, G. (2025). TrueFake: A Real World Case Dataset of Last Generation Fake Images also Shared on Social Networks. *" 2025 International Joint Conference on Neural Networks (IJCNN), Rome, Italy, 2025, pp. 1-8,* 10.1109/IJCNN64981.2025.11228911

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, doi.org/10.48550/arXiv.2006.07397

Dong, J., Wang, W., & Tan, T. (2013). CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing* (pp. 422–426), 10.1109/ChinaSIP.2013.6625374

Farid, H. (2019). Image forensics. *Annual Review of Vision Science*, 5, 549–573.

Fridrich, J., & Kodovský, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7, 868–882, 10.1109/TIFS.2012.2190402

Gardella, M., Musé, P., Morel, J.-M., & Colom, M. (2021). Forgery detection in digital images by multi-scale noise estimation. *Journal of Imaging*, 7, 119, doi.org/10.3390/jimaging7070119

Gorle, R., & Guttavelli, A. (2025). Enhanced Image Tampering Detection using Error Level Analysis and CNN. *Engineering, Technology & Applied Science Research*, 15, 19683–19689, doi.org/10.48084/etasr.9593

Heller, S., Rossetto, L., & Schuldt, H. (2018). The PS-Battles dataset — an image collection for image manipulation detection. *arXiv preprint arXiv:1804.04866*, doi.org/10.48550/arXiv.1804.04866

Hsu, Y.-F., & Chang, S.-F. (2006). Detecting image splicing using geometry invariants and camera characteristics consistency. In *IEEE International Conference on Multimedia and Expo (ICME)* (pp. 549–552), 10.1109/ICME.2006.262447

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). *DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection*. arXiv. https://doi.org/10.48550/arXiv.2001.03024

Kaur, S., Sinha, N., Jain, P., Koli, S., Sharma, A., & Lathwal, A. (2024). Enhanced image forgery detection using a hybrid approach: Integration of ELA, CNN, and XGBoost. *International Journal of Performability Engineering*, 20, 367–378, 10.23940/ijpe.24.06.p4.367378

Khoo, B., Phan, R. C.-W., & Lim, C.-H. (2022). Deepfake attribution: On the source identification of artificially generated images. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12, e1438, doi.org/10.1002/widm.1438

Kwan, C., & Larkin, J. (2019). Demosaicing of Bayer and CFA 2.0 patterns for low lighting images. *Electronics*, 8, 1444, doi.org/10.3390/electronics8121444

Li, W., Yuan, Y., & Yu, N. (2009). Passive detection of doctored JPEG image via block artifact grid extraction. *Signal Processing*, 89, 1821–1829, doi.org/10.1016/j.sigpro.2009.03.025

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). *Celeb-DF: A large-scale challenging dataset for DeepFake forensics*. arXiv. https://doi.org/10.48550/arXiv.1909.12962

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, doi.org/10.48550/arXiv.1711.05101

Mahfoudi, G., Tajini, B., Retraint, F., Morain-Nicolier, F., Dugelay, J. L., & Pic, M. (2019). DEFACTO: Image and face manipulation dataset. In *2019 27th European Signal Processing Conference (EUSIPCO)* (pp. 1–5), 10.23919/EUSIPCO.2019.8903181

Mao, D. (2023). DeepfakeArt Challenge dataset. *Kaggle*. https://www.kaggle.com/datasets/danielmao2019/deepfakeart

Mareen, H., De Neve, L., Lambert, P., & Van Wallendael, G. (2023). Harmonizing image forgery detection & localization: Fusion of complementary approaches. *Journal of Imaging*, 10, 4, doi.org/10.3390/jimaging10010004

Ng, T.-T., Hsu, J., & Chang, S.-F. (2009). Columbia image splicing detection evaluation dataset. *DVMM Lab, Columbia University*, ee.columbia.edu/ln/dvmm/downloads/AuthSplicedDataSet/dlform.html

Qazi, E. U. H., Zia, T., & Almorjan, A. (2022). Deep learning-based digital image forgery detection system. *Applied Sciences*, 12, 2851, doi.org/10.3390/app12062851

Rana, M. M. R., Hasnat, A., & Rahaman, G. M. A. (2022). SMIFD-1000: Social media image forgery detection database. *Forensic Science International: Digital Investigation*, 41, 301392, doi.org/10.1016/j.fsidi.2022.301392

Rana, M. M. R., Rahman, M. A., & Talukder, K. H. (2023). FIDD-1500: Fake Image Detection Dataset. In *2023 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1–6), 10.1109/ICCIT60459.2023.10441386

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to detect manipulated facial images*. arXiv. https://doi.org/10.48550/arXiv.1901.08971

Sharma, D. K., Singh, B., Agarwal, S., Garg, L., Kim, C., & Jung, K.-H. (2023). A survey of detection and mitigation for fake images on social media platforms. *Applied Sciences*, 13, 10980, doi.org/10.3390/app131910980

Shi, Z., Chen, H., & Zhang, D. (2025). Robustifying vision transformer for image forgery localization with multi-exit architectures. *Pattern Recognition*, 164, 111565, doi.org/10.1016/j.patcog.2025.111565

Singh, D., Singh, P., Jena, R., & Chakraborty, R. S. (2023). An image forensic technique based on JPEG ghosts. *Multimedia Tools and Applications*, 82, 14153–14169, doi.org/10.1007/s11042-022-13699-x

Singh, S., & Kumar, R. (2024). Image forgery detection: comprehensive review of digital forensics approaches. *Journal of Computational Social Science*, 7, 877–915, doi.org/10.1007/s42001-024-00265-8

Tralic, D., Zupancic, I., Grgic, S., & Grgic, M. (2013). CoMoFoD — New database for copy-move forgery detection. In *Proceedings ELMAR-2013* (pp. 49–54), http://www.vcl.fer.hr/comofod.

Tyagi, S., & Yadav, D. (2023). MiniNet: A concise CNN for image forgery detection. *Evolving Systems*, 14, 545–556, doi.org/10.1007/s12530-022-09446-0

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14, 910–932, 10.1109/JSTSP.2020.3002101

Wang, T., Liao, X., Chow, K. P., Lin, X., & Wang, Y. (2024). Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57, 1–35, doi.org/10.1145/3699710

Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2017). Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications*, 76, 4801–4834, doi.org/10.1007/s11042-016-3795-2