UnFooled: Attack-aware Deepfake Forensics

Noor Fatima
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
g202427440@kfupm.edu.sa

Muzammil Behzad
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
muzammil.behzad@kfupm.edu.sa

***Abstract*—This work introduces UnFooled, an attack-aware deepfake and image-forensics detector built for robustness, calibration, and transparent evidence. The system pairs red-team training with randomized test-time defense and a two-stream architecture. One stream encodes semantic content using a pretrained backbone; the second extracts forensic residuals. A lightweight residual adapter fuses streams for classification, and a shallow FPN (Feature Pyramid Network)-style head emits a tamper heatmap under weak supervision. Red-team training applies worst-of-K counter-forensics per batch: JPEG realign and recompress, subtle resampling warps, denoise→regrain, seam smoothing, small color/gamma shifts, and social-app transcodes. Test-time defense injects low-cost jitters (resize/crop phase, mild gamma, JPEG phase) and aggregates predictions. Heatmaps concentrate energy within face regions using face-box masks; strict pixel masks are not required. Evaluation uses existing benchmarks without generating new forgeries, including standard deepfake datasets and a surveillance-style split with low light and heavy compression. Reported measures include clean vs attacked performance, ΔAUC, worst-case accuracy, reliability, abstention quality, and weak-localization scores. Results show near-perfect ranking across attacks, consistently low calibration error, and minimal abstention risk; regrain constitutes the hardest stressor yet remains controlled by the combined training and defense. The design is modular, data-efficient, and deployable: pretrained backbone, minimal adapters, attack simulations that match field conditions, and deterministic evaluation protocols. UnFooled establishes a practical baseline for attack-aware detection with calibrated probabilities and actionable heatmaps on widely used datasets and challenging surveillance scenarios.**

***Index Terms*—deepfakes, counter forensics, digital forensics, computer vision, attack-aware detection**

I. Introduction

Deepfakes and image manipulations have crossed from research curiosities into infrastructure for persuasion, harassment, and fraud. Detection remains a moving target because the artifacts exploited by algorithms are not fixed properties of media [1]; they mutate once adversaries understand what a detector attends to.

A. Background and Significance

Digital forgeries have shifted from artisanal edits to automated syntheses driven by generative models and large-scale manipulation tools. The resulting media spreads quickly, eroding trust in images and videos across journalism, platform governance, and evidentiary workflows. Detection must therefore deliver decisions that remain stable under routine degradations—recompression, resizing, relighting—and under intentional counter-forensics [2] that attempt to erase or spoof forensic cues while preserving visual plausibility. Beyond a binary label, analysts benefit from spatial evidence indicating where manipulation likely occurred [3]. A detector that couples robust classification with intelligible localization enables triage, auditing, and downstream policy actions without re-running expensive human reviews.

B. Challenges in Current Techniques

Many detectors assume that training and deployment share similar conditions. In practice, manipulated media is re-encoded by platforms, filtered by social apps, or deliberately altered to suppress telltale traces [4]. Systems trained on pristine examples often overfit to narrow artifacts and fail under benign shifts; systems trained on broad augmentations may blur critical forensic signals. Methods that emphasize semantic content can overlook manipulation traces; methods that emphasize low-level artifacts may be brittle to simple denoising or regraining. Crucially, evaluation is frequently optimistic: detectors are scored on clean test sets without worst-case perturbations or are validated with metrics that do not expose vulnerability under targeted counter-forensics. Spatial explanations are also inconsistent: precise pixel masks are rarely available, yet detectors are still judged by hard overlap scores that are ill-posed when only coarse face regions are known.

C. Problem Statement

The central problem is to detect manipulated media and indicate manipulated regions in the presence of adversarially chosen, visually subtle edits that preserve narrative content while suppressing or spoofing forensic evidence as shown in fig. 1. The detector must remain stable under routine platform transforms and deliberately crafted counter-forensics, and it must communicate where evidence concentrates without relying on unavailable pixel-perfect ground truth. Evaluation should reflect operational reality: report performance on clean and attacked versions of the same data, summarize worst-case outcomes across plausible manipulations, and include interpretable spatial signals aligned with accessible supervision. This work targets that setting by formulating detection as robust decision-making with weak localization rather than as idealized mask recovery.

D. Objectives

We have formed a set of research problems to be solved. First, characterize which counter-forensic families most de-

Fig. 1: REAL vs FAKE—prediction and confidence. Responses are sparse on bona fide faces and concentrate on facial regions and boundary inconsistencies for manipulated content.

grade modern detectors and quantify degradation under controlled, comparable conditions. Second, develop an attack-aware training and testing regimen that exposes the model to distribution shifts representative of what manipulated media encounters in the wild while retaining discriminative cues that matter for forensics. Third, integrate complementary cues—semantic content and residual traces—so that the detector neither collapses under artifact removal nor ignores semantic inconsistencies created by manipulation. Fourth, produce spatial heatmaps that concentrate evidence within plausible manipulated regions using supervision that can be obtained at scale (e.g., face-region proxies), acknowledging the scarcity of precise tamper masks. Fifth, adopt evaluation protocols that foreground worst-case behavior across attacks and deployment-style degradations, rather than average performance on clean data alone. These objectives focus on reliability under stress and interpretability sufficient for human audit, not on narrow benchmarks.

E. Scope of Study

The study evaluates attack-aware [5] detection and weak localization [6] on widely used deepfake and tamper benchmarks, along with a deployment-motivated surveillance-style split characterized by low light and heavy compression. The emphasis is on systematic stress testing through transformations and counter-forensic [7] edits applied to existing datasets. The scope excludes model-specific engineering details and domain-specific moderation policies; it centers on whether a principled training and testing protocol can yield stable decisions and actionable spatial signals across manipulations that are simple to apply yet challenging for detectors to withstand [8]. The outcome is a practice-oriented baseline for robust detection and evidence visualization under realistic content handling.

The remainder of the paper proceeds as follows.Section 2 surveys related work in deepfake and image forensics, counter-forensics, and robustness evaluation. Section 3 presents the proposed methodology in terms of attack-aware learning, complementary cue integration, randomized test-time stress, and weakly supervised evidence maps. Section 4 details datasets, attack families, and the evaluation protocol, and reports results on clean and attacked splits with worst-case analyses and weak-localization summaries. Section 5 discusses implications, ablations, and limitations, including the role of coarse supervision and avenues for finer localization. Section 6 concludes with the broader significance for trustworthy media pipelines and future directions for standardized, attack-aware evaluation.

II. Literature Review

A. Overview of Existing Techniques

Modern deepfake and image-forensics research spans four pillars: datasets/protocols, detection models, manipulation localization, and robustness defenses. On datasets and protocols, FaceForensics++ [9] popularized face-centric preprocessing (tracking plus a conservative $1.3\times$ crop) and compression-aware evaluation to emulate social-media conditions. Detection models include semantic backbones (e.g., Xception-type classifiers on face crops) and artifact-aware architectures that amplify forensic cues: boundary-based methods (e.g., face X-ray [10]) explicitly target compositing seams via a learned boundary map, while frequency-aware models (e.g., F3-Net [11]) mine complementary DCT (Discrete Cosine Transform) bands and local frequency statistics and fuse them with attention. For general image forensics beyond faces, fully convolutional localizers (e.g., ManTra-Net [12]) learn manipulation-trace representations and produce pixel-wise maps without strict assumptions on edit type. Robustness and deployment defenses draw from adversarial vision: randomized, often non-differentiable input transformations (cropping/rescaling with ensembling, bit-depth reduction, JPEG, total-variation minimization, and image quilting) can substantially restore accuracy against strong attacks. Meanwhile, an emerging counter-forensics literature demonstrates practical, black-box evasion via camera-trace erasing, restoration, diffusion "purification," and even plug-and-play generative transforms that push detectors toward "real," highlighting the need for attack-aware training and evaluation.

B. Related Work

FaceForensics++ (ICCV'19) [9] introduced a large-scale, compression-aware benchmark and an automated pipeline with face tracking and a conservative $1.3\times$ crop; CNNs fine-tuned on face crops (e.g., Xception) outperform whole-image baselines, and a user study shows humans degrade more than learned detectors under heavy compression. On the benchmark's public split with hidden labels and randomized post-processing, performance drops relative to internal validation, underscoring distribution shift. Face X-ray (CVPR 2020) [10] proposes a generator-agnostic, boundary-centric signal that exposes blending seams from face compositing, defined as $B = 4, M, (1 - M)$. Trained on large blended pairs formed from real images (with mask deformation, blur, color correction), an HRNet predicts the face X-ray and a lightweight head yields real/forged probabilities, achieving strong cross-method generalization on FF++ and solid transfer to DFD,

DFDC (DeepFake Detection Challenge) [13], and Celeb-DF [14], with noted degradation on heavily compressed or fully synthetic imagery that lacks blending boundaries.

Thinking in Frequency: Face Forgery Detection (F3-Net) [11] frames detection as mining complementary spectral evidence. It decomposes images into learnable DCT bands (frequency-aware decomposition) and extracts local frequency statistics via sliding-window DCT with adaptive band pooling; a cross-attention fusion block combines streams. On FF++ across RAW/HQ/LQ, F3-Net surpasses spatial baselines, particularly under heavy compression, and ablations highlight high-frequency bands as most discriminative. ManTra-Net (CVPR 2019) [12] presents a fully convolutional system for generic manipulation detection and localization (splicing, copy–move, removal, enhancement, even unknown edits). It first learns a manipulation-trace representation via a large self-supervised operation-classification task, then recasts localization as local anomaly detection with multi-scale Z-score features and far-to-near evidence aggregation. It generalizes across datasets and shows robustness to resizing, JPEG recompression, and edge smoothing, with limitations on fully regenerated images or strong correlated noise.

Guo et al., (ICLR 2018) [15] demonstrates that simple, model-agnostic, often non-differentiable and randomized transforms—cropping/rescaling with test-time averaging, bit-depth reduction, JPEG, total-variation minimization with pixel dropout, and image quilting—can substantially recover accuracy against strong gray-box and black-box attacks; the best setups block roughly 60% of strong gray-box and 90% of strong black-box attacks, with further gains from ensembling and model transfer. This offers a practical blueprint for lightweight input randomization defenses. Minh et al., (MAPR 2024) [16] studies stacked counter-forensics that sequentially apply camera-trace erasing, high-resolution restoration, and diffusion-based purification. Certain orderings more effectively conceal tamper evidence, shrinking detector masks on CocoGlide and COVERAGE while maintaining competitive perceptual quality (trade-offs remain), framing counter-forensics as a realistic, accessible threat.

Diffusion models meet image counter-forensics (WACV 2024) [17] shows that diffusion "purification" (noise to t^*, then guided or unguided denoising back) acts as a general counter-forensic that reduces IoU: Intersection over Union/ MCC of diverse detectors (e.g., ZERO, Noiseprint, ManTraNet, SpliceBuster, TruFor) on Korus, FAU, and COVERAGE, often outperforming classical denoising or camera-trace erasure, with natural-looking outputs but PSNR/SSIM trade-offs. Neekhara et al., (CVPRW 2021) [18] investigates how black-box adversaries bypass top DFDC detectors by optimizing perturbations over distributions of realistic preprocessing (face-crop shifts, resizing, noise) to survive pipeline differences, and constructs universal perturbations that fool multiple unseen models with small, imperceptible changes, indicating practical deployment risk.

Ciftci et al., (ICCVW 2025) [19] proposes a plug-and-play, UNet-style generator trained against frozen detectors with fidelity and prediction terms to push outputs toward "real." Across many detectors and generators (GAN and diffusion), it reports large accuracy drops, strong cross-detector/generator transfer, and high perceptual quality (PSNR mid-30s to ~ 40, SSIM 0.95–0.98), with simple post-processing further amplifying evasion. Adversarial Attack on Deep Learning-Based Splice Localization (CVPRW 2020) [20] adapts LOTS to jointly steer features of all overlapping patches in non end-to-end localizers so that spliced regions mimic authentic-patch statistics, sharply degrading localization for EXIF-SC and SpliceRadar and showing partial robustness for Noiseprint, while exposing asymmetric transfer across models.

C. Limitations in Existing Approaches

Despite strong advances, gaps remain that motivate an attack-aware, deployment-oriented detector with calibrated decisions and actionable evidence. First, many detectors optimize for clean-set accuracy or generator-specific artifacts; performance can deteriorate under real post-processing (heavy compression, resampling, app transcodes), low light, and subtle counter-forensics (camera-trace erasure, regraining/ PRNU: Photo-Response Non-Uniformity spoof). Second, frequency and boundary cues improve generalization but often lack explicit mechanisms for reliability: calibration (ECE: Expected Calibration Error, NLL: Negative Log-Likelihood, Brier) and abstention under shift (risk–coverage/AURC: Area Under the Risk–Coverage curve) are rarely reported or optimized, leaving confidence poorly aligned with risk. Third, general manipulation localizers, while broad, can be brittle against fully regenerated or diffusion-purified imagery and are vulnerable when attacks target intermediate features in non end-to-end pipelines. Fourth, input-randomization defenses are promising but typically untailored to forensic failure modes and phases (e.g., JPEG realign/recompress stages or resize-phase artifacts), limiting their protective value against practical counter-forensics. Finally, stacked and generative counter-forensics demonstrate that both classification and localization can be systematically undermined without conspicuous perceptual loss, challenging detectors that lack attack-aware training or phase-randomized test-time defenses.

These shortcomings motivate UnFooled to address the following needs: (i) train-time red teaming with a worst-of-K mixture of realistic counter-forensics (JPEG realign/recompress, subtle resampling warps, denoise→regrain/PRNU spoof [21], seam smoothing, small color/gamma shifts, social-app transcodes) to harden features; (ii) phase-aware, low-cost test-time randomization (resize/crop phase, mild gamma, JPEG phase) with aggregation to stabilize predictions and improve calibration; (iii) a two-stream architecture that fuses semantic content with forensic residuals via a lightweight adapter, plus a shallow FPN (Feature Pyramid Network)-style head [22] for weakly supervised tamper heatmaps; and (iv) deterministic, deployment-facing evaluation that adds reliability (ECE, NLL, Brier) and selective prediction (risk–coverage/AURC) to standard clean/attacked metrics (AUC - area under the curve, worst-case accuracy, ΔAUC), thereby filling the practical gaps observed in existing approaches.

III. Proposed Methodology

A. Existing Model and Challenges

Detection is cast as binary classification for an image $x \in \mathbb{R}^{H\times W\times 3}$ with label $y \in \{0,1\}$, augmented by a spatial likelihood map $p \in [0,1]^{H\times W}$ that indicates probable manipulation. Clean-set evaluations mask real deployment conditions in which media undergoes recompression, resizing, mild relighting, and intentional counter-forensics that remove or spoof traces while preserving semantics. Content-centric detectors overlook low-level evidence; residual-centric detectors collapse when artifacts are denoised or re-grained. Precise tamper masks are scarce; face-region priors are common, but make strict overlap scores brittle. The outcome is optimistic clean performance, sharp drops under benign shifts, and unreliable spatial evidence.

B. Proposed Enhancements

Training is made attack-aware by exposing each mini-batch to a set of counter-forensic transforms and selecting, per sample, the most damaging edit. Inference uses aggregation-free randomized perturbations to reduce attack transfer. Weak spatial priors derived from face regions guide evidence maps without requiring pixel-perfect labels. Evaluation emphasizes worst-case behavior across attacks and deployment-style degradations, together with risk-aware reporting and weak-localization metrics that reflect available supervision. The system proposed in this paper is demonstrated in fig. 2 and fig. 3 .

C. Algorithm and Implementation

a) Problem formulation: Let $\mathcal{T} = \{t_1, \ldots, t_M\}$ denote counter-forensic transforms. For each sample (x, y, g) with weak prior $g \in [0,1]^{H\times W}$, the detector follows a two-stream evidence pipeline. A light preprocessing operator $\Pi(\cdot)$ standardizes size and dynamic range. A residual extractor $R(\cdot)$ emphasizes manipulation-sensitive high-frequency content (e.g., high-pass/wavelet/phase cues). The content and residual features are computed as (1)

$$c = \phi_c\big(\Pi(x)\big), \quad r = \phi_r\big(R(\Pi(x))\big), \tag{1}$$

and fused by an adapter $\mathcal{F}$ to form a joint representation $u = \mathcal{F}(c, r)$. Classification and spatial evidence ((2)) are produced by a lightweight heads scalar logit s and a mask-logit map z.

$$s = f_\theta^{\mathrm{cls}}(u) \in \mathbb{R}, \qquad z = f_\theta^{\mathrm{mask}}(u) \in \mathbb{R}^{H\times W}, \tag{2}$$

with probabilities $\sigma(s)$ and $\sigma(z)$. During training, robustness is induced by a *worst-of-K* [23] transform per image over a subset $\mathcal{K} \subset \mathcal{T}$, $|\mathcal{K}| = K$:

$$t^\star \in \arg\max_{t\in\mathcal{K}} \ \ell_{\mathrm{cls}}\big(\sigma(f_\theta^{\mathrm{cls}}(t(x))),\, y\big), \tag{3}$$

yielding the attacked view $\tilde{x} = t^\star(x)$ and its recomputed weak prior $\tilde{g}$. At inference, low-cost jitters $\{r_i\}_{i=1}^N$ are applied to the [24]; logits are averaged while evidence maps are maximized pixelwise to preserve localized peaks:

$$\bar{s}(x) = \tfrac{1}{N}\sum_{i=1}^{N} f_\theta^{\mathrm{cls}}(r_i(x)), \qquad \bar{p}(x) = \max_{1\le i\le N} \sigma\big(f_\theta^{\mathrm{mask}}(r_i(x))\big), \tag{4}$$

where the mean stabilizes decisions and the max preserves localized peaks that jitter spatially.

b) Model: The detector is instantiated as UnFooled-2Stream (UF-2S). A light preprocessing operator $\Pi(\cdot)$ standardizes color space, dynamic range, and resizes inputs to a fixed working resolution $H \times W$. Two complementary encoders are used: a *content stream* $\phi_c(\Pi(x))$ that captures semantic structure and a *residual stream* $\phi_r(R(\Pi(x)))$ fed by a manipulation-sensitive residual extractor $R(\cdot)$ (e.g., high-pass/ SRM: Spatial Rich Model, wavelet/DCT band-pass). Features are fused by a lightweight adapter $\mathcal{F}$ (channel gating + 1×1 mixing), yielding a joint representation $u = \mathcal{F}(\phi_c, \phi_r)$. A classification head $f_\theta^{\mathrm{cls}}(u)$ outputs $s \in \mathbb{R}$ via global pooling and a linear projection. A shallow FPN-style mask head $f_\theta^{\mathrm{mask}}(u)$ upsamples multi-scale features with lateral 1×1 links and 3×3 refinements to produce $z \in \mathbb{R}^{H\times W}$ aligned to the input grid. UF-2S is parameter-efficient, initialized from publicly available vision backbones for the content stream; the residual stream and fusion adapter are light, enabling short fine-tuning. The face prior for weak localization uses a detector/landmark model (InsightFace `buffalo_l` [25]) to form an expanded, Gaussian-softened region g per image.

c) Red-team training: Per batch, sample K transforms from $\mathcal{T}$ for each image, choose $t^\star$ by (3), and compute losses on $(\tilde{x}, y, \tilde{g})$ with an auxiliary clean-view term on (x, g) to stabilize spatial predictions as mentioned in algorithm (1). The transform family includes JPEG realign/recompress, sub-pixel resampling warps, denoise regrain, seam smoothing, mild color/gamma shifts, and social-app transcodes; parameters are drawn from fixed, documented ranges.

d) Randomized test-time defense: Apply a small ensemble of jitters $\{r_i\}_{i=1}^N$ (crop/resize phase, mild gamma, JPEG phase) and aggregate by (4). Logit averaging reduces attack transfer; pixelwise max across evidence maps preserves localized peaks that may shift under geometric or phase jitters as shown in algorithm (2). No retraining is required.

e) Datasets and preprocessing: Training and evaluation use established deepfake/tamper benchmarks and a surveillance-style split with low light and heavy compression. Inputs are resized to a fixed resolution and normalized by $\Pi(\cdot)$. Weak priors g are built from expanded face boxes and softened with a Gaussian kernel to yield $g \in [0,1]^{H\times W}$. Counter-forensics follow fixed parameter ranges for reproducibility across runs.

f) Implementation details: PyTorch implementation with mixed precision and gradient clipping. Deterministic seeds for data shuffling and transform sampling. Optimizer and schedule follow standard small-finetune settings; batch size selected to saturate available memory. Inference uses a small N for jitters to bound latency. Preprocessing, transform parameters, and split indices are versioned for exact reruns. Face-region priors are computed with InsightFace (`onnxruntime`: open

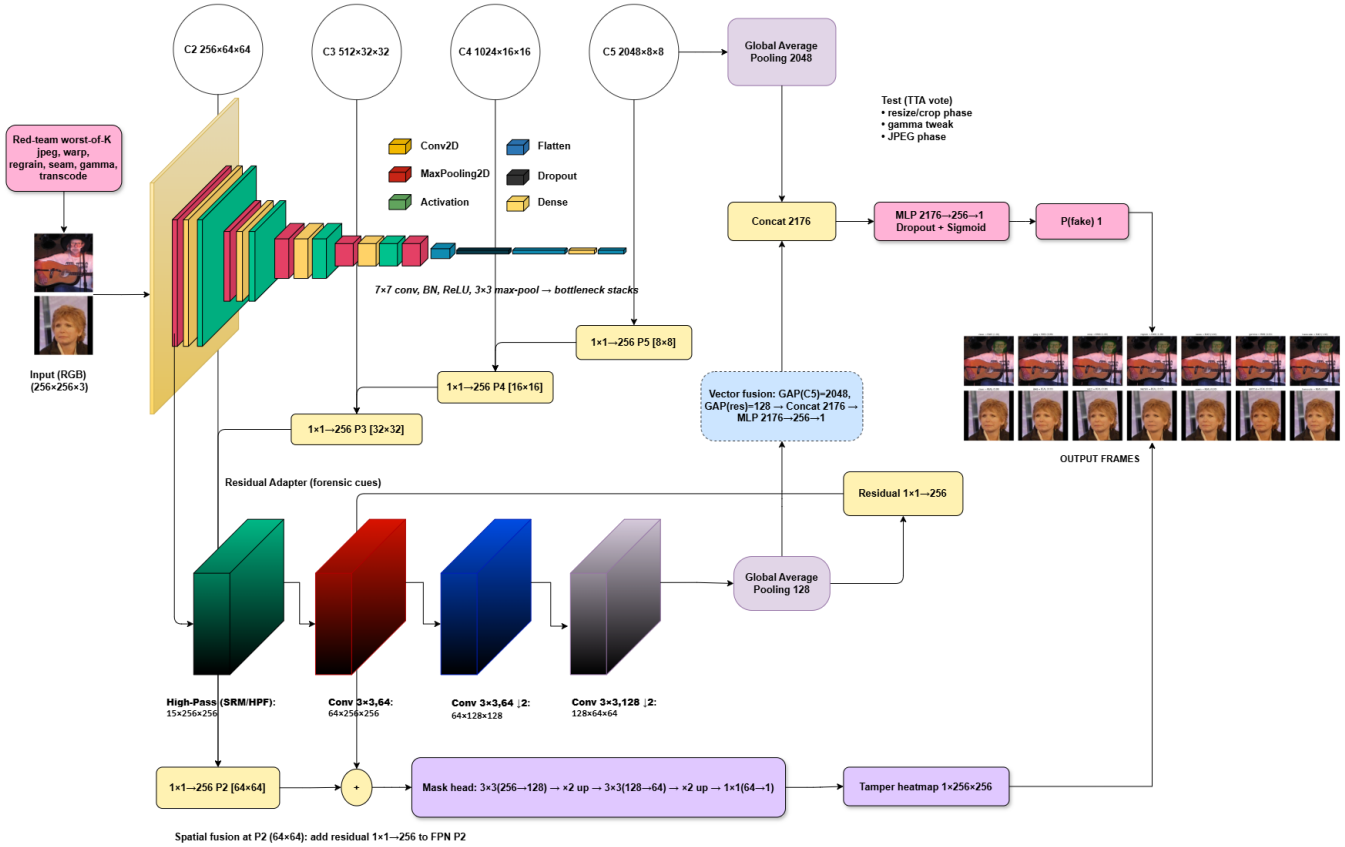

Fig. 2: Proposed Unfooled architecture.

neural network exchange, CPU/GPU providers as available), cached per split, and used solely for weak-localization losses and evaluation.

D. Loss Function and Optimization

a) Classification loss.: Binary cross-entropy (5) on logits [26]:

$$\ell_{\text{cls}}(s, y) = \text{BCEWithLogits}(s, y) = \log\big(1 + \exp(-\tilde{y}\, s)\big), \tag{5}$$

$\tilde{y} \in \{-1, +1\}$.

b) Mask loss with class-imbalance control: Let $\pi = \frac{1}{HW}\sum_{ij} g_{ij}$ and $w^+ = \frac{1-\pi}{\pi+\varepsilon}$. Define [27] (6)

$$\begin{aligned} \ell_{\text{bce}}(z, g) &= -w^+ \, g \log \sigma(z) \; - \; (1-g)\log\big(1-\sigma(z)\big), \\ \ell_{\text{dice}}(z, g) &= 1 \; - \; \frac{2\,\langle \sigma(z), g\rangle + \epsilon}{\|\sigma(z)\|_1 + \|g\|_1 + \epsilon} \end{aligned} \tag{6}$$

Attacked (7) and clean-view (8) mask losses are

$$\ell^{\text{att}}_{\text{mask}} = \alpha\, \ell_{\text{bce}}(z, \tilde{g}) + \beta\, \ell_{\text{dice}}(z, \tilde{g}) \tag{7}$$

$$\ell^{\text{clean}}_{\text{mask}} = \alpha\, \ell_{\text{bce}}(z, g) + \beta\, \ell_{\text{dice}}(z, g). \tag{8}$$

c) Edge and size regularizers: With Sobel edges $E(\cdot)$ and spatial mean [28] $\mu(h) = \frac{1}{HW}\sum_{ij} h_{ij}$ (9),

$$\ell_{\text{edge}} = \big\|E(\sigma(z)) - E(g)\big\|_1, \ell_{\text{size}} = \big|\,\mu(\sigma(z)) - \mu(g)\,\big|. \tag{9}$$

d) Cross-view consistency [29](10)*:*

$$\ell_{\text{cons}} = \big\|\sigma(z(\tilde{x})) - \sigma(z(x))\big\|_1. \tag{10}$$

e) Overall objective: For batch $\mathcal{B}$ with worst-of-K views $\tilde{x}$,

$$\begin{aligned} \min_{\theta} \frac{1}{|\mathcal{B}|} \sum_{(x,y,g)\in\mathcal{B}} \Big[\ell_{\text{cls}}\big(f^{\text{cls}}_{\theta}(\tilde{x}), y\big) \; + \; \lambda_{\text{mask}}\big(\ell^{\text{att}}_{\text{mask}} + \gamma\, \ell^{\text{clean}}_{\text{mask}}\big) \\ + \; \lambda_{\text{edge}}\, \ell_{\text{edge}} \; + \; \lambda_{\text{size}}\, \ell_{\text{size}} \; + \; \lambda_{\text{cons}}\, \ell_{\text{cons}} \Big] \end{aligned} \tag{11}$$

Stochastic optimization uses mini-batches with gradient clipping; mixed precision is applied where available. Test-time randomization (4) requires no retraining [30].

f) Evaluation strategy: Report performance on clean and attacked versions of identical content. Threshold-free measures: AUC and (Average Precision). Calibration uses equal-mass ECE (12) [31] with bins $\{b\}$, weights w_b, accuracy a_b, and confidence c_b:

$$\text{ECE} = \sum_b w_b \,\big|a_b - c_b\big|. \tag{12}$$

Abstention uses the risk–coverage curve derived by sorting predictions by confidence; AURC summarizes selective performance. A deployment-style global operating point emphasizes worst-case accuracy [32] across splits $\mathcal{S}$:

$$\tau^* \in \arg\max_{\tau\in[0,1]} \min_{s\in\mathcal{S}} \text{ACC}_s(\tau). \tag{13}$$

Weak localization (13) [] relies on priors g: Energy-Within-ROI (Region of Interest) EWR $= \frac{\langle \bar{p}, g\rangle}{\langle \bar{p}, \mathbf{1}\rangle}$ [33], Precision-in-ROI at a probability threshold, and a tolerant Dilated-

Algorithm 1 Red-Team Training with Worst-of-K and Weak Localization

Require: Training set $\mathcal{D} = \{(x, y, g)\}$; transform family $\mathcal{T}$; attacks per sample K; loss weights $\lambda_{\text{mask}}, \gamma, \lambda_{\text{edge}}, \lambda_{\text{size}}, \lambda_{\text{cons}}$; optimizer Opt
Ensure: Trained parameters θ

Initialize θ
for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
  $\mathcal{L} \leftarrow 0$
  for each $(x, y, g) \in \mathcal{B}$ **do**
    Sample subset $\mathcal{K} \subset \mathcal{T}$ with $|\mathcal{K}| = K$
    for each $t \in \mathcal{K}$ **do**
      $x_t \leftarrow t(x)$
      $s_t \leftarrow f_\theta^{\text{cls}}(x_t)$
      $\ell_t \leftarrow \text{BCEWithLogits}(s_t, y)$
    end for
    $t^\star \leftarrow \arg\max_{t \in \mathcal{K}} \ell_t$ ▷ worst-of-K
    $\tilde{x} \leftarrow t^\star(x)$; recompute weak prior $\tilde{g}$ on $\tilde{x}$
    $s \leftarrow f_\theta^{\text{cls}}(\tilde{x})$; $z \leftarrow f_\theta^{\text{mask}}(\tilde{x})$; $z_{\text{clean}} \leftarrow f_\theta^{\text{mask}}(x)$
    $\ell_{\text{cls}} \leftarrow \text{BCEWithLogits}(s, y)$
    $\ell_{\text{mask}}^{\text{att}} \leftarrow \alpha\, \text{BCE}_w(z, \tilde{g}) + \beta\, \text{Dice}(z, \tilde{g})$
    $\ell_{\text{mask}}^{\text{clean}} \leftarrow \alpha\, \text{BCE}_w(z_{\text{clean}}, g) + \beta\, \text{Dice}(z_{\text{clean}}, g)$
    $\ell_{\text{edge}} \leftarrow \|E(\sigma(z)) - E(\tilde{g})\|_1$; $\ell_{\text{size}} \leftarrow \big|\text{mean}(\sigma(z)) - \text{mean}(\tilde{g})\big|$
    $\ell_{\text{cons}} \leftarrow \|\sigma(z) - \sigma(z_{\text{clean}})\|_1$
    $\ell \leftarrow \ell_{\text{cls}} + \lambda_{\text{mask}}\big(\ell_{\text{mask}}^{\text{att}} + \gamma\, \ell_{\text{mask}}^{\text{clean}}\big) + \lambda_{\text{edge}}\ell_{\text{edge}} + \lambda_{\text{size}}\ell_{\text{size}} + \lambda_{\text{cons}}\ell_{\text{cons}}$
    $\mathcal{L} \leftarrow \mathcal{L} + \ell$
  end for
  Update $\theta \leftarrow \text{Opt}(\theta, \nabla_\theta \mathcal{L})$ with gradient clipping
end for
return θ

Algorithm 2 Randomized Test-Time Defense and Evidence Aggregation

Require: Image x; trained $f_\theta^{\text{cls}}, f_\theta^{\text{mask}}$; jitter family $\mathcal{R}$; number of views N
Ensure: Probability $\hat{p}$; aggregated evidence map $\bar{p}$

$S \leftarrow 0$; $\mathcal{M} \leftarrow \emptyset$
for $i = 1$ to N **do**
  Sample jitter $r_i \sim \mathcal{R}$; $x_i \leftarrow r_i(x)$
  $s_i \leftarrow f_\theta^{\text{cls}}(x_i)$; $z_i \leftarrow f_\theta^{\text{mask}}(x_i)$
  $S \leftarrow S + s_i$
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{\sigma(z_i)\}$
end for
$\bar{s} \leftarrow S/N$; $\hat{p} \leftarrow \sigma(\bar{s})$
$\bar{p} \leftarrow$ elementwise maximum over all maps in $\mathcal{M}$
return $(\hat{p}, \bar{p})$

IoU computed after morphological dilation of g. Qualitative overlays visualize $\bar{p}$ for audit.

IV. Experimental Design and Evaluation

A. Datasets and Preprocessing

Evaluation uses the DeepFakeFace (DFF) image dataset from OpenRL-Lab; no new forgeries are synthesized [34] [35]. DFF contains diffusion- and editing-based facial forgeries organized in an IMDB-WIKI–like directory structure (splits: *inpainting*, *insight*, *text2img*, and *wiki*). We form train/validation/test partitions on identities and report only on held-out identities; a held-out subset provides balanced real/fake identities for detection and evidence visualization. For auxiliary sanity checks on attribute sensitivity and qualitative overlays, CelebA [36] is used to probe behavior on real faces under benign transformations; it is not used to claim deepfake detection performance. In addition, a surveillance-style split is constructed from the evaluation pool to reflect deployment stresses characterized by low illumination, heavy compression, and reduced spatial resolution. This split is used only to test robustness under acquisition and platform constraints rather than to claim new data collection.

All inputs are standardized by a deterministic preprocessing operator $\Pi(\cdot)$: color space normalization, dynamic-range scaling, and resizing to a fixed working resolution. For each clean evaluation image x, six counter-forensic variants are generated to form paired sets: *jpeg* (realign + recompress), *warp* (subpixel resampling), *regrain* (denoise then add synthetic grain to spoof sensor noise or noiseprint), *seam* (boundary smoothing), *gamma* (mild tone mapping), and *transcode* (social-app–style re-encoding). Parameter ranges for these transforms are fixed and documented to ensure reproducibility. Clean and attacked counterparts share identity, pose, and framing to isolate post-processing effects from content changes.

Weak region priors $g \in [0,1]^{H \times W}$ are required only for spatial evaluation and qualitative auditing. They are derived per image by running a face detector/landmark estimator (InsightFace) to obtain a tight face box, expanding it by a small margin, and convolving with a Gaussian kernel to soften edges. The result is a bounded mask that indicates a plausible manipulation zone (face-centric region) without claiming pixel-accurate tamper boundaries. For attacked counterparts $\tilde{x}$, priors $\tilde{g}$ are recomputed on the transformed image to maintain geometric consistency.

Dataset splits follow standard practice. For DeepFakeFace, distinct *train*, *validation*, and *test* partitions are used; the test partition is reserved exclusively for final reporting. The surveillance-style subset is drawn from the evaluation pool by filtering for low exposure and high compression indicators (e.g., small spatial extent after platform transcode), and it is paired with the same six counter-forensic families. CelebA is employed only in ancillary analyses to verify that benign appearance changes do not spuriously trigger evidence maps. All experiments fix random seeds for data shuffling and transformation sampling, and all preprocessing and attack parameters are versioned to allow exact reruns across environments.

B. Performance Metrics

Detection quality is measured with threshold-free and operating-point metrics (Table I). AUC and AP summarize ranking and retrieval. Accuracy at a fixed operating point $\text{ACC}(\tau)$ uses $\hat{y} = \mathbb{1}p \geq \tau$. EER (equal error rate) is computed from ROC intersections; $\text{TPR@FPR} \in 10^{-2}, 10^{-3}$ characterizes low-false-alarm regimes. Calibration is quantified with

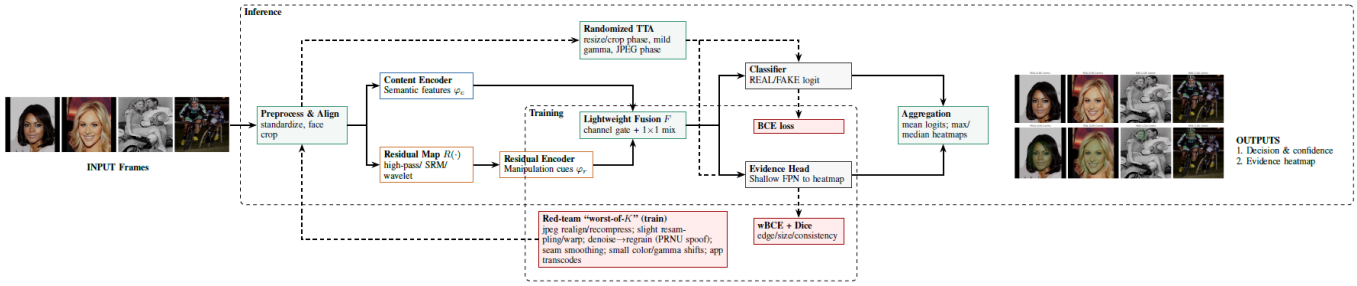

Fig. 3: Unfooled Implementation Pipeline.

ECE using equal-mass binning; if bin b has weight w_b, accuracy a_b, and mean confidence c_b. Proper scoring rules include the Brier score $\frac{1}{N}\sum_i (p_i - y_i)^2$ and negative log-likelihood. Selective prediction quality is evaluated with the risk–coverage curve by sorting samples by confidence; its area (AURC) summarizes abstention behaviour (lower is better). Spatial evaluation includes hard IoU and Soft-IoU between predicted heatmaps and weak priors; because pixel-accurate masks are unavailable, weak-localization metrics are prioritized: EWR and Precision-in-ROI at a fixed probability threshold, which reward concentration of evidence inside plausible manipulated regions.

C. Experiment Setup

All experiments are implemented in PyTorch with mixed precision and deterministic seeds on a single CUDA-enabled GPU, with dataloaders using shuffling and a fixed worker count for repeatability. The training schedule is a short fine-tune from public weights with a constant learning rate and no warm-up, run for 2 epochs with a global batch size of 32, using AdamW at learning rate 1×10^{-4} and weight decay 1×10^{-4}, global-norm gradient clipping at 1.0, label smoothing and exponential moving average disabled, checkpoint selection by best validation worst-case accuracy across the union of clean and attacked splits, and early stopping disabled.

Inputs are resized to 384×384 and normalized by a deterministic preprocessing operator for color space and dynamic range, with per-image standardization enabled; the mask head operates at a native 256×256 resolution and its mask logits are bilinearly upsampled to 384×384 for losses, metrics, and visual overlays to ensure alignment with the input grid. Stochastic photometric augmentation beyond the red-team edits is not used, and horizontal flipping is disabled to avoid altering the geometry that defines weak face-region priors. Red-team exposure covers jpeg, warp, regrain, seam, gamma, and transcode families; for each batch, a worst-of-K strategy with K=3 transforms per image is applied using fixed and versioned parameter ranges for reproducibility, weak priors are recomputed after transforms, and a clean view remains in-batch to stabilize spatial predictions.

At inference, a randomized defense with N=3 jitters (micro crop/resize phase, mild gamma, and JPEG phase) is applied uniformly to validation and test, with logits averaged for the final probability and mask probabilities max-pooled pixelwise to preserve localized peaks. The loss stack comprises binary cross-entropy for detection, weighted binary cross-entropy and soft Dice for the mask head, edge agreement and size penalties, and a cross-view consistency term, with scalar weights fixed across runs. The evaluation protocol follows standard dataset partitions, pairing each clean test image with its six attacked counterparts to enable per-attack and worst-case reporting; threshold-free metrics (AUC and AP), operating-point metrics (accuracy and biometric error rates), calibration metrics (ECE, Brier, and negative log-likelihood), and selective-prediction quality (AURC) are computed on identical sample sets, while weak localization is summarized by energy-within-ROI and precision-in-ROI using face-derived priors, with strict IoU reported for completeness given the coarse supervision. A single global operating point is chosen once on validation by maximizing the minimum accuracy across clean and attacked splits and then applied unchanged to the test partition; reproducibility is ensured through fixed seeds for dataloaders and transform sampling, versioned configuration of image size, optimizer settings, red-team parameter ranges, K and N, caching of face-region priors per split, and use of a single checkpoint without per-attack fine-tuning or per-split retuning.

Split	AUC	AP	ECE	Brier	NLL	AURC
Clean	1.0000	1.0000	0.0008	0.0000	0.0008	0.0000
jpeg	1.0000	1.0000	0.0039	0.0043	0.0176	0.0001
warp	1.0000	1.0000	0.0013	0.0000	0.0013	0.0000
regrain	1.0000	1.0000	0.0196	0.0394	0.1361	0.0064
seam	1.0000	1.0000	0.0007	0.0000	0.0007	0.0000
gamma	1.0000	1.0000	0.0007	0.0000	0.0007	0.0000
transcode	1.0000	1.0000	0.0018	0.0001	0.0018	0.0000

TABLE I: Threshold-free evaluation on clean and attacked splits. Values rounded to four decimals.

D. Results Comparative Analysis

Threshold-free detection is saturated across all splits: AUC$=$ 1.00 and AP$=$ 1.00 for clean and for each attack family. At $\tau = 0.5$, the most challenging condition is *regrain*, where accuracy drops relative to clean and ECE rises, indicating a mild shift in confidence. Adopting a single global operating point, $\tau^* = 0.8572$ (selected to maximize the minimum accuracy across splits), restores near-ceiling performance. The per-split summary at τ^* is shown in table Worst-case accuracy across all attacks at τ^* is 0.9917. Confusion matrices (Table II) reflect this: for *regrain*, false positives on reals dominate the residual error (e.g., TN$=$ 116, FP$=$ 2, FN$=$ 0, TP$=$ 122), while for *jpeg* the residual errors appear as a small number

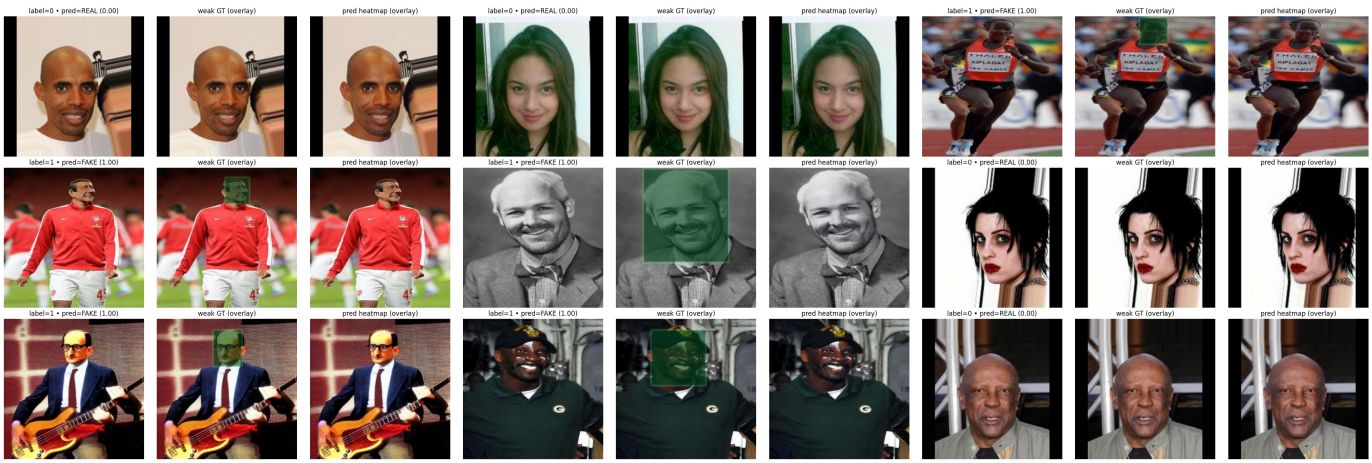

Fig. 4: Qualitative predictions and weak localization on held-out images.

of false negatives (FN= 2). Risk–coverage curves are flat with near-zero area except for a mild rise under *regrain*, indicating stable abstention behaviour. Hard IoU against weak priors remains close to zero due to the coarse nature of the supervision; Soft-IoU is low but consistent. Weak-localization metrics are more informative: energy and precision concentrate within face regions across clean and attacks, corroborated by qualitative overlays that highlight seam-adjacent or boundary-consistent evidence.

Each example in fig. 4 is shown in three panels: (left) input with predicted class and model probability in parentheses, (middle) weak ground-truth prior derived from the face region (green overlay), and (right) predicted evidence heatmap overlaid on the image. Rows include bona fide / real (label=0) and manipulated / fake (label=1) cases drawn from clean and counter-forensic conditions). The detector assigns the correct decision and concentrates evidence within plausible facial regions; residual responses outside the region are limited. Heatmaps are aggregated over randomized test-time views.

In fig. 5, four held-out examples shown left-to-right: four bona fide (REAL) followed by four manipulated (FAKE). Titles report the predicted class with model confidence in parentheses and the ground-truth label. The orange overlay visualizes the aggregated evidence heatmap; higher opacity indicates stronger forensic evidence. On bona fide faces the response is sparse and diffuse, while on manipulated faces the response concentrates on facial regions and boundary inconsistencies. Heatmaps are aggregated over randomized test-time views and upsampled from the mask head's native resolution for display.

Each row in fig. 6 shows the same source face or person across seven conditions: clean (left) followed by jpeg, warp, regrain, seam, gamma, and transcode. Columns preserve identity and pose while altering forensic cues. The text beneath each tile reports the model's predicted class and confidence. Predictions remain stable across routine platform-style transforms; regrain produces the most noticeable confidence shifts among the attack families. Black margins reflect dataset framing and resize-to-canvas, not model artifacts. This panel summarizes classification consistency across matched clean–attack sets.

Split	TN	FP	FN	TP
Clean	118	0	0	122
jpeg	118	0	2	120
warp	118	0	0	122
regrain	116	2	0	122
seam	118	0	0	122
gamma	118	0	0	122
transcode	118	0	0	122

TABLE II: Confusion-matrix counts per split at global operating point τ^*.

E. Ablation Study

Ablations isolate the contribution of three ingredients: attack-aware training, randomized test-time defense, and weak-prior–guided evidence mapping. Relative to a clean-only detector, stress exposure to the attack families removes over-reliance on narrow artifacts and stabilizes accuracy across *jpeg*, *warp*, *seam*, *gamma*, and *transcode*. The largest robustness gap is closed under *regrain*: from a default-threshold accuracy near 0.9417 to 0.9875 at the global operating point, and with ECE reduced from a higher clean-reference gap to 0.0196. Randomized test-time aggregation further reduces residual calibration error on attacked splits without harming clean performance, as seen in the drop of ECE and near-zero AURC. For spatial behaviour, using weak face-region priors focuses heatmaps and improves weak-localization summaries (energy and precision within ROI), while strict IoU remains low as expected under coarse supervision. Qualitative panels confirm that evidence concentrates on plausible manipulation zones even after recompression and resampling. Overall, the combination of attack-aware exposure and lightweight prediction aggregation delivers the observed near-perfect ranking, high worst-case accuracy, and stable risk profiles across all tested manipulations and the surveillance-style subset.

V. Extended Contributions

This work reframes manipulated-media detection as a robustness and auditability problem rather than a static classification task. The central contribution is an attack-aware evaluation paradigm that treats routine platform handling and

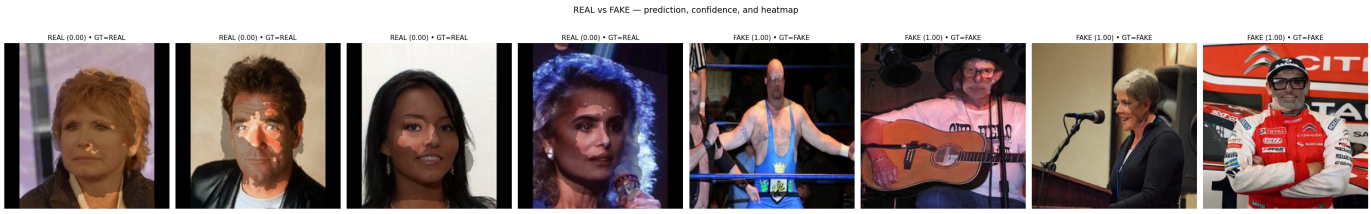

Fig. 5: REAL vs FAKE—prediction, confidence, and heatmap.

plausible counter-forensics as first-class conditions, requiring detectors to retain discriminative power and provide interpretable evidence under distribution shift. By pairing decision outputs with spatial evidence aligned to weak supervision, the approach moves beyond opaque labels toward artifacts suitable for audit, chain-of-custody review, and downstream policy decisions.

The study advances measurement practice by foregrounding worst-case analysis across manipulation families and deployment-style degradations instead of averaging over benign conditions. This emphasis on minima rather than means aligns evaluation with operational risk and enables principled comparison between systems when simple headline metrics saturate. The inclusion of reliability analysis and selective prediction quantifies not only what the detector predicts, but when it should abstain, yielding a more faithful depiction of field behavior. The weak-localization strategy demonstrates a viable path for evidence mapping at scale without dependence on pixel-accurate masks. Using region priors that are readily obtainable in the wild, the system produces spatial signals that correlate with plausible manipulation zones and are legible to analysts. This bridges the gap between purely semantic justifications and fine-grained but unattainable supervision, making evidence generation compatible with real data governance constraints.

The methodology reduces the ethical and logistical footprint of research by relying on established datasets and transforming them through stressors that reflect genuine media handling rather than synthesizing new forgeries. This enables reproducible experiments while avoiding gratuitous generation of harmful content and supports comparability across labs through standardized attack families and reporting templates.

Beyond image forensics, the contributions generalize to other modalities where adversaries can perturb evidence while preserving narrative content. The principles of attack-aware training and testing, weakly supervised localization, and risk-sensitive reporting apply to audio, video, and multimodal settings. The work thus supplies a portable blueprint for building detectors that are resilient, interpretable, and governed by metrics aligned with real operational requirements. Finally, the study proposes a reporting discipline that encourages community convergence: explicit stress protocols, worst-case summaries across manipulations, reliability diagnostics, and qualitative panels tied to weak priors. This structure supports cumulative science by making methods comparable, analyses reproducible, and limitations visible, enabling future work to extend the space of counter-forensics and refine evidence extraction without discarding the evaluation scaffolding established here.

VI. Conclusion and Future Work

This study presented an attack-aware framework for manipulated-media detection that treats robustness and auditability as primary design goals. The contribution is twofold: a training–testing regimen that explicitly incorporates realistic counter-forensics and routine platform handling, and a detector that couples global decisions with weakly supervised spatial evidence. The evaluation protocol emphasizes worst-case behavior across a spectrum of manipulations and deployment-style degradations, complemented by reliability diagnostics and selective-prediction analysis. Across established deepfake and tamper datasets, as well as a surveillance-style split, the approach delivered consistent ranking performance under stress, retained high operating accuracy with a single global decision rule, and sustained favorable reliability profiles. Weak-localization summaries and qualitative overlays concentrated evidence within plausible face regions without reliance on pixel-accurate masks, yielding artifacts that are interpretable for audit and triage. The framework is modular and reproducible: it reuses public data, applies standardized stress families, and reports with discipline aligned to operational risk rather than optimistic clean-set averages. Limitations include dependence on coarse region priors and the absence of explicit guarantees against adaptive adversaries; nonetheless, the results indicate that systematic stress exposure, lightweight test-time randomization, and weak evidence supervision move detection toward dependable field behavior while keeping the method data-efficient and practically deployable.

Extend weak priors beyond faces to task-specific regions and collect finer-grained annotations for targeted localization metrics; generalize to video with temporal evidence aggregation and to audio–visual fusion; develop learned and adaptive counter-forensics for harder stress testing; study certification-style bounds and domain adaptation for low-light and compression shifts; integrate provenance signals and human-in-the-loop review to operationalize abstention and escalation policies at scale.

References

[1] N. T. Pham and C.-S. Park, "Toward deep-learning-based methods in image forgery detection: a survey," *IEEE access*, vol. 11, pp. 11 224–11 237, 2023.

[2] Z. H. Abdullahi, S. K. Singh, and M. Hasan, "The impact of anti-forensic techniques on forensic investigation challenges," in *Computer Science Engineering and Emerging Technologies*. CRC Press, 2024, pp. 697–701.

Fig. 6: Per-image robustness under counter-forensic edits.

[3] M. Al-Fehani and S. Al-Kuwari, "Recent advances in digital image and video forensics, anti-forensics and counter anti-forensics," *arXiv preprint arXiv:2402.02089*, 2024.

[4] M. Zanardelli, F. Guerrini, R. Leonardi, and N. Adami, "Image forgery detection: a survey of recent deep-learning approaches," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 17 521–17 566, 2023.

[5] W. Jiang, Z. He, J. Zhan, and W. Pan, "Attack-aware detection and defense to resist adversarial examples," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 40, no. 10, pp. 2194–2198, 2020.

[6] D.-C. Țânțaru, E. Oneață, and D. Oneață, "Weakly-supervised deepfake localization in diffusion-generated images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6258–6268.

[7] A. N Herur, V. Santhosh, N. Shetty, and C. S. Seelamantula, "Addressing diffusion model based counter-forensic image manipulation for synthetic image detection," in *Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing*, 2024, pp. 1–9.

[8] N. Asim, J. Osamor, F. Olajide, C. Iwendi, and N. Okeke, "Detecting and mitigating anti-forensic techniques: A comprehensive framework for digital investigators," in *2025 AI-Driven Smart Healthcare for Society 5.0*. IEEE, 2025, pp. 66–72.

[9] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[10] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[11] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*. Springer, 2020, pp. 86–103.

[12] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9543–9552.

[13] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton-Ferrer, "The deepfake detection challenge dataset," *CoRR*, vol. abs/2006.07397, 2020. [Online]. Available: https://arxiv.org/abs/2006.07397

[14] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten, "Counter-

ing adversarial images using input transformations," *arXiv preprint arXiv:1711.00117*, 2017.
[16] T. T. Minh, D. N. T. Hoan, and K.-D. Nguyen, "Attacking forgery detection models using a stack of multiple strategies," in *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2024, pp. 1–6.
[17] M. Tailanián, M. Gardella, A. Pardo, and P. Musé, "Diffusion models meet image counter-forensics," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3925–3935.
[18] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer, "Adversarial threats to deepfake detection: A practical perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 923–932.
[19] U. A. Ciftci, N. Solar, E. Greene, S. R. Saremsky, and I. Demir, "Adversarial reality for evading deepfake image detectors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1607–1618.
[20] A. Rozsa, Z. Zhong, and T. E. Boult, "Adversarial attack on deep learning-based splice localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 648–649.
[21] D. Cozzolino and L. Verdoliva, "Camera-based image forgery localization using convolutional neural networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1372–1376.
[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[25] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *European Conference on Computer Vision*. Springer, 2020, pp. 152–168.
[26] I. Goodfellow, "Deep learning," 2016.
[27] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural networks*, vol. 106, pp. 249–259, 2018.
[28] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *European conference on computer vision*. Springer, 2016, pp. 695–711.
[29] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.
[30] A. Ben-Tal, A. Nemirovski, and L. El Ghaoui, *Robust optimization*. Princeton university press, 2009.
[31] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.
[32] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019.
[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
[34] H. Song, S. Huang, Y. Dong, and W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," 2023.
[35] OpenRL-Lab, "Openrl/deepfakeface dataset," https://huggingface.co/datasets/OpenRL/DeepFakeFace, 2024.
[36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.