# PRIVACY-PRESERVING DEEPFAKE DETECTION AND FINE-GRAINED ARTIFACT LOCALIZATION WITH FORENSIC-AWARE IMAGE HIDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Remote face-forgery analysis typically requires transmitting identifiable facial content to remote services, creating privacy risks during transmission and storage. Encryption can conceal image content, but an image-like ciphertext may reveal the presence of protected communication and facilitate traffic-level discovery or tracking. Moreover, a conventional decrypt-then-infer pipeline exposes plaintext at the endpoint that performs inference. We propose VeilTrace, a recoverable image-hiding framework that represents a face as a cover-like protected image and reconstructs it only at an authorized forensic endpoint for downstream detection and fine-grained localization. VeilTrace addresses two key challenges. First, high RGB reconstruction fidelity does not guarantee preservation of forensic signals. We therefore introduce a forensic-response consistency objective based on fixed forensic proxy responses. Second, residual misalignment between real/fake pairs can contaminate pixel-level DSSIM supervision. We therefore apply full DSSIM only to geometrically reliable pairs and use a mask-supported fallback for unreliable pairs when a usable manipulation mask is available, yielding our Hybrid-DSSIM strategy. Extensive experiments show that VeilTrace reduces direct exposure of facial plaintext under the evaluated threat model while retaining high forgery-detection accuracy, with a rounded mean frame-level AUC of 88.98% in the FA+SBI setting and 49.76 dB recovery PSNR. Hybrid-DSSIM yields the strongest artifact-map agreement in the controlled ablation.

## 1 INTRODUCTION

Digital-media forensics seeks to assess visual authenticity, but facial evidence can reveal a person's identity. Many face-forgery detectors assume direct access to the input image. When inference is outsourced, this assumption exposes facial content to a remote service during both processing and storage. Reliable forensic analysis, however, typically requires not only an image-level authenticity decision but also spatial localization of manipulated regions (Yu et al., 2025; Peng et al., 2026). The challenge is therefore to limit the exposure of facial content while preserving the fine-grained, localized cues needed for accurate detection and localization.

Existing privacy-preserving approaches target different stages of remote face-forgery analysis, yet each suffers from stage-specific limitations. Encryption hides image content in transit and at rest, but its image-like ciphertext can be conspicuous, leaking the presence of protected communication to traffic observers. If the service decrypts before inference, the inference endpoint is additionally exposed to plaintext. SecDFDNet avoids plaintext reconstruction by any individual outsourced server, but requires a specialized secure-computation protocol (Chen et al., 2023). StegaFFD avoids image recovery by detecting forgeries directly in the stego domain, but requires protected-domain inference (Ma et al., 2026). Overall, existing designs trade off communication-level detectability, endpoint exposure, protocol complexity, and model compatibility, as illustrated in Figure 1. A recoverable image-hiding alternative introduces two additional challenges. Conventional reconstruction metrics such as PSNR, SSIM, and LPIPS do not directly constrain the heterogeneous manipulation cues used by downstream forensic models (Shiohara & Yamasaki, 2022; Nguyen et al., 2024; Li et al., 2024b; Cui et al., 2025). Moreover, pixel-wise localization targets can be corrupted by



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

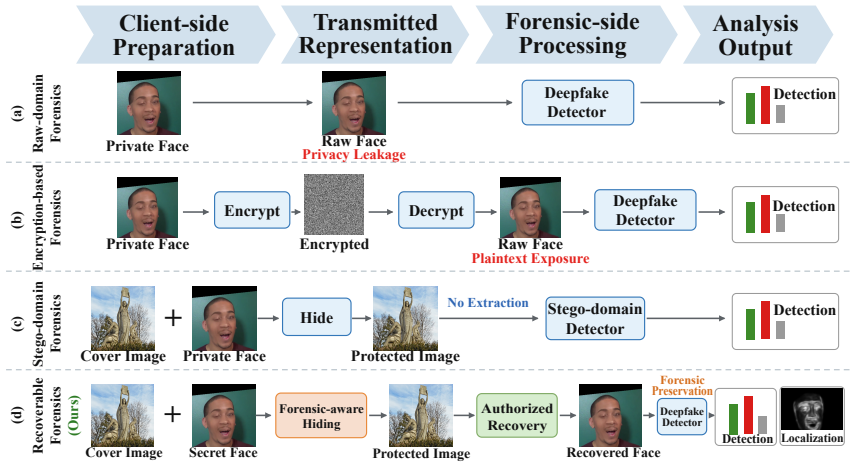


Figure 1: Remote face-forgery analysis settings. (a) **Raw-domain forensics** transmits the face directly for inference. (b) **Encryption-based forensics** decrypts the input at the inference endpoint. (c) **Stego-domain forensics** performs inference on the protected image without face recovery. (d) **Recoverable forensics (ours)** transmits a cover-like image and recovers the face only at an authorized endpoint for detection and localization.

residual misalignment in cropping, pose, or facial landmarks, producing spurious responses outside manipulated regions.

To address these limitations, we propose VeilTrace, a recoverable forensic framework evaluated under a nominal channel without unmodeled lossy or geometric transformations. VeilTrace transmits a cover-like protected image and reconstructs the face only at an authorized endpoint, thereby reducing plaintext exposure under the evaluated threat model while maintaining compatibility with conventional detectors and localizers. To preserve manipulation-sensitive evidence, we enforce consistency between fixed forensic proxy responses of the original and recovered images during training. This objective provides a preservation surrogate rather than guaranteeing retention of every detector feature. For localization, Hybrid-DSSIM applies full DSSIM supervision to geometrically reliable pairs and a dilated, blurred mask-supported target to unreliable pairs with a usable manipulation mask. Pairs without a usable target are omitted from the localization loss. The pristine counterpart is used only to construct training targets and is unavailable at inference time. Our main contributions are summarized as follows:

1. We propose VeilTrace, a recoverable face-hiding framework that transmits a cover-like protected image and reconstructs the face only at an authorized endpoint, thereby reducing plaintext exposure under the evaluated threat model while remaining compatible with conventional deepfake detectors and fine-grained localizers after authorized recovery.
2. We introduce a forensic-response consistency objective that aligns fixed forensic proxy responses between original and recovered images, serving as a preservation surrogate for manipulation-sensitive evidence. We further propose *Hybrid-DSSIM*, a geometry-aware localization supervision strategy that applies full DSSIM to geometrically reliable pairs and a dilated, blurred mask-supported target to unreliable pairs with a usable manipulation mask, while omitting pairs without a usable target from the localization loss.

## 2 Related Work

**Face Forgery Detection and Localization.** Generalizable detectors aim to learn cues that transfer across manipulation pipelines. Face X-Ray focuses on blending boundaries (Li et al., 2020b), SBI synthesizes self-blended pseudo-forgeries (Shiohara & Yamasaki, 2022), LAA-Net emphasizes artifact-prone facial regions (Nguyen et al., 2024), and FreqBlender introduces frequency-aware training (Li et al., 2024b). Forensics Adapter and DFD-HR use pretrained representations or hierarchical routing for cross-domain detection (Cui et al., 2025; Sun et al., 2026a). LRL and LiSiam

study local consistency and localization-invariant representations (Chen et al., 2021; Wang et al., 2022). For localization, Delocate and KFD provide spatial evidence beyond image-level classification (Hu et al., 2024; Yu et al., 2025), while DiffusionFF estimates fine-grained DSSIM maps from detector features (Peng et al., 2026). TruFor considers pixel-level localization in a broader image-forensics setting (Guillaro et al., 2023). Most of these methods assume direct access to the image under analysis. In contrast, we evaluate detection and localization after a privacy-preserving hide–recover step.

**Privacy-Preserving Visual Analysis.** Prior work protects privacy at different stages of training and inference. FedForgery protects decentralized training data through federated learning (Liu et al., 2023), whereas SecDFDNet performs outsourced face forgery inference with additive secret sharing (Chen et al., 2023). RIC studies recoverable privacy-preserving image classification (Liu et al., 2024). PIRNet and AIRPNet analyze restoration in the steganographic domain (Deng et al., 2023; Gao et al., 2026). HideMIA and StegaVAR extend hidden-domain analysis to medical imaging and video recognition (Lin et al., 2024; Chen et al., 2026). StegaFFD performs face forgery detection directly on stego images (Ma et al., 2026). In our setting, the face is reconstructed only at an authorized forensic endpoint, so the fidelity of hiding and recovery directly affects downstream forensic analysis.

**Recoverable Image Hiding and Recovery.** HiNet formulates concealing and revealing as forward and inverse processes in the wavelet domain (Jing et al., 2021), while DeepMIH extends invertible hiding to higher payload capacities and studies invisibility, security, and recovery accuracy (Guan et al., 2023). ReImage embeds a shuffled representation of the image for neural self-recovery (Kim & Seo, 2026), while RecoverMark embeds facial content into surrounding regions for manipulation localization and recovery (An et al., 2026). These methods primarily target concealment and recovery, or proactive manipulation protection. In contrast, our setting focuses on preserving manipulation-sensitive forensic responses through concealment and authorized recovery.

# 3 Methodology

## 3.1 Overall Framework

Remote face-forgery analysis exposes identity-bearing content when plaintext faces are transmitted to external services. Encryption can conceal image content, yet its image-like ciphertext may render the presence of protected communication conspicuous to traffic observers. A decrypt–then–infer service exposes plaintext at the inference endpoint, whereas stego-domain inference remains incompatible with conventional image models. We therefore conceal the face within a natural cover, reconstruct it only at an authorized endpoint, and require the hide–recover path to preserve forensic responses and support reliable localization under imperfect alignment.

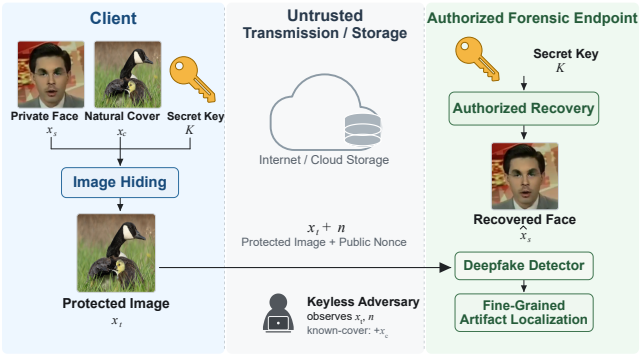


Figure 2: VeilTrace deployment and threat model.

VeilTrace implements this setting through the three stages illustrated in Fig. 3. Given a private face $x_s$, an unrelated cover $x_c$, a secret key $K$, and a public nonce $n$, the client produces a protected image $x_t$. The authorized endpoint uses $(K, n)$ to recover $\hat{x}_s$, which is then fed to the downstream detector and localizer:

$$x_t = \mathcal{H}_\theta(x_s, x_c \mid K, n), \qquad \hat{x}_s = \mathcal{R}_\theta(x_t \mid K, n), \qquad \hat{y} = D_\psi(\hat{x}_s), \quad \hat{M} = G_\phi(\hat{x}_s). \tag{1}$$

The *Forensic-Aware Recoverable Hiding* component uses keyed wavelet lifting to form $x_t$, inverse lifting to recover $\hat{x}_s$, and the fixed forensic-response loss $\mathcal{L}_{\text{for}}$ to preserve manipulation-sensitive

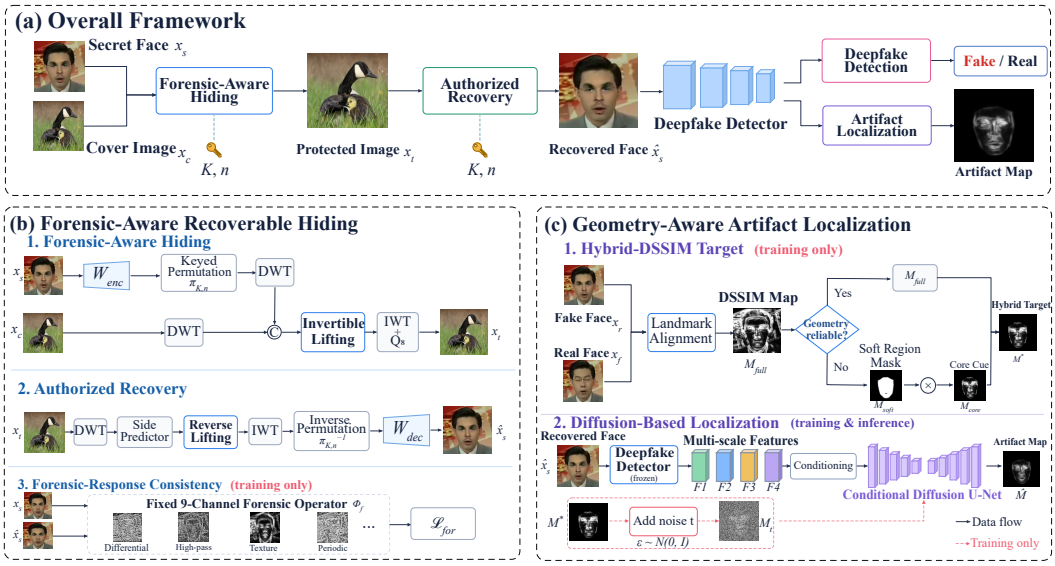


Figure 3: Overview of VeilTrace. (a) The overall framework hides a private face in a natural cover, performs authorized recovery, and then applies deepfake detection and artifact localization. (b) Forensic-aware recoverable hiding combines keyed hiding, authorized recovery, and a training-only forensic-response consistency loss. (c) Geometry-aware artifact localization constructs Hybrid-DSSIM targets and uses a conditional diffusion localizer. Dashed paths indicate training-only supervision.

responses. The *Geometry-Aware Artifact Localization* component constructs a continuous Hybrid-DSSIM target from each real–fake pair: full DSSIM supervision is applied to geometrically reliable pairs, whereas a region-restricted target is used for geometrically unreliable pairs with a usable manipulation mask. A conditional diffusion localizer is trained on these targets. The hiding, detection, and localization stages are trained sequentially, as described below. The pristine counterpart and manipulation mask are used only to construct localization supervision. At inference, the authorized endpoint first recovers the face from $x_t$. The reconstruction is then processed by the detector, and the detector features condition the localizer. Consequently, neither downstream model operates on the protected image prior to recovery.

### 3.2 Forensic-Aware Recoverable Hiding

**Keyed recoverable hiding.** The hide–recover backbone adopts the invertible wavelet-lifting design used in privacy-preserving restoration (Deng et al., 2023; Gao et al., 2026). Before lifting, a lightweight transform gives $z_s = \mathcal{W}_\omega^{\text{enc}}(x_s)$. We then partition $z_s$ into non-overlapping $16 \times 16$ patches and apply

$$\pi_{K,n} = \text{Perm}(K, n), \qquad \bar{z}_s = \mathcal{S}_{\pi_{K,n}}(z_s). \tag{2}$$

The permutation seed is deterministically produced by HMAC-SHA256$(K, n)$. Different nonces therefore induce distinct permutations for the same key, whereas the matching $(K, n)$ pair reproduces the inverse mapping during authorized recovery. The key is kept external to the network parameters.

The shuffled payload and cover are transformed by Haar DWT. Following PIRNet and AIRPNet, the $i$-th lifting block updates the payload and carrier streams as

$$p_i = p_{i-1} - P_i(c_{i-1}), \qquad c_i = c_{i-1} + U_i(p_i), \tag{3}$$

where $P_i$ and $U_i$ denote the learned prediction and update subnetworks, respectively. The transmitted image is $x_t = Q_8(\text{IWT}(c_L))$, where $Q_8(x) = \text{round}(255\ \text{clip}(x, 0, 1))/255$. The same 8-bit quantization is included during training with a straight-through gradient. During recovery, a side predictor estimates the untransmitted lifting stream from $\text{DWT}(x_t)$ before reverse lifting, IWT, inverse permutation, and $\mathcal{W}_\omega^{\text{dec}}$ recover $\hat{x}_s$.

**Downstream forensic training.** Once the hiding model has converged, we freeze its parameters and use the same hide–recover pipeline employed at inference to generate authorized reconstructions for detector training. Each real or forged training sample is passed through this pipeline, and the detector $D_\psi$ is optimized with binary cross-entropy on the resulting reconstructions. Afterward, both the detector and its feature encoder are frozen prior to training the conditional diffusion localizer.

### 3.3 Geometry-Aware Artifact Localization

**Hybrid-DSSIM target.** Pixel-level DSSIM supervision requires not only a nominal pristine counterpart but also sufficient geometric alignment between the real–fake pair. The localization target is therefore constructed from the original real–fake pair, whereas the localizer operates on the authorized reconstruction of the forged image. Given a forged face $x_f$ and its pristine counterpart $x_r$, we estimate a conservative integer translation from stable landmarks,

$$\Delta = \mathrm{round}\left(\mathrm{median}_j[l_f^j - l_r^j]\right), \qquad M_{\mathrm{full}} = O \odot [1 - \mathrm{SSIM}(x_f, A_\Delta(x_r))]\,. \tag{4}$$

where $A_\Delta$ denotes the integer translation operator, $O$ is the valid overlap, and SSIM follows Wang et al. (2004). Throughout, we adopt the unscaled DSSIM convention $\mathrm{DSSIM} = 1 - \mathrm{SSIM}$. A pair is accepted for full DSSIM supervision only when the shift, residual landmark error, and static-region DSSIM are all small:

$$g = \mathbf{1}\left[\|\Delta\|_\infty \le \tau_s, \quad e_{\mathrm{lm}} \le \tau_l, \quad e_{\mathrm{static}} \le \tau_d\right]. \tag{5}$$

here $e_{\mathrm{lm}}$ is the median residual landmark error after translation, normalized by image size, and $e_{\mathrm{static}}$ is the mean DSSIM response inside the valid non-manipulated region.

If $g = 0$ but a manipulation mask is available, a dilated and blurred support $M_{\mathrm{soft}}$ gives $M_{\mathrm{core}} = M_{\mathrm{full}} \odot M_{\mathrm{soft}}$. The target is

$$M^* = \begin{cases} 0, & y = 0, \\ M_{\mathrm{full}}, & y = 1,\ g = 1, \\ M_{\mathrm{core}}, & y = 1,\ g = 0 \text{ and mask available.} \end{cases} \tag{6}$$

Pairs that lack both reliable geometry and a usable manipulation mask are excluded from the localization loss. By contrast, SBI real/fake pairs come from the same source face and provide additional aligned full-DSSIM supervision.

**Diffusion localizer.** The target $M^*$ is a continuous artifact map rather than a binary mask. We use the detector-conditioned diffusion formulation of DiffusionFF (Peng et al., 2026) as the localization architecture and modify only the target construction in our controlled ablation. Once the detector has been trained and frozen, it extracts multi-scale features $F^{1:4}$ from the authorized reconstruction $\hat{x}_s$. These features condition a U-Net denoiser supervised by $M^*$. The forward noising process and the masked denoising objective are detailed in the Loss Functions subsection. At inference, reverse diffusion yields a single-channel artifact map.

### 3.4 Loss Functions

Reconstructing RGB pixels alone does not explicitly constrain the local forensic responses that a forgery detector may exploit. To address this, we introduce a fixed nine-channel forensic operator $\Phi_f$ comprising two Laplacian responses, horizontal and vertical Sobel responses, a diagonal-difference response, the mean response of two fixed $5 \times 5$ high-pass filters, a multi-scale local-texture response, a local chromatic residual, and a short-range periodic-difference response. The operator outputs are deterministically normalized before loss computation. $\Phi_f$ contains no learnable parameters, and its output is not provided to the detector. The forensic-response consistency loss is defined as

$$\mathcal{L}_{\mathrm{for}} = \left\|\Phi_f(\hat{x}_s) - \mathrm{sg}[\Phi_f(x_s)]\right\|_1, \tag{7}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. The response of the original face is a fixed target, so gradients propagate only through the recovered branch.

The hiding network is optimized with

$$\mathcal{L}_{\mathrm{hide}} = \lambda_c \mathcal{L}_{\mathrm{cover}} + \lambda_r \mathcal{L}_{\mathrm{rec}} + \lambda_p \mathcal{L}_{\mathrm{payload}} + \lambda_f \mathcal{L}_{\mathrm{for}} + \lambda_a \mathcal{L}_{\mathrm{aux}}. \tag{8}$$

The loss $\mathcal{L}_{\text{cover}}$ contains protected-to-cover reconstruction terms, the cover-PSNR constraint, low-frequency wavelet consistency, and an out-of-range penalty. $\mathcal{L}_{\text{rec}}$ combines RGB reconstruction losses with SSIM for authorized recovery. $\mathcal{L}_{\text{payload}}$ constrains recovery of the shuffled payload and its low-frequency component. $\mathcal{L}_{\text{aux}}$ contains side-stream prediction, wavelet-guidance, and total-variation terms. The term $\mathcal{L}_{\text{for}}$ is specific to preserving the fixed forensic responses. For localization, the target $M^*$ is diffused according to the DDPM process (Ho et al., 2020):

$$M_t = \sqrt{\bar{\alpha}_t} M^* + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \qquad \mathcal{L}_{\text{diff}} = \frac{\sum_i v_i \left\| \epsilon_i - \epsilon_\phi(M_{t,i}, t, F_i^{1:4}) \right\|_2^2}{\sum_i v_i}. \tag{9}$$

Here $\epsilon \sim \mathcal{N}(0, I)$ and $v_i \in \{0, 1\}$ indicates whether a valid localization target is available. The diffusion localizer is optimized with this masked denoising loss after the privacy model and detector feature encoder have been frozen.

# 4 Experiments and Analysis

## 4.1 Experimental Settings

**Datasets.** We train the forensic models on FaceForensics++ (FF++) C23, which contains DeepFakes, Face2Face, FaceSwap, and NeuralTextures (Rössler et al., 2019). For cross-dataset detection, we evaluate on Celeb-DF-v1/2 (Li et al., 2020c), DeepFakeDetection (DFD) (Dufour et al., 2019), DFDC (Dolhansky et al., 2020), DFDCP (Dolhansky et al., 2019), FaceShifter (Li et al., 2020a), and UADFV (Yang et al., 2019). Hiding, authorized recovery, and the main security tests use a fixed 200-sample FF++ held-out set. Fine-grained localization is evaluated on FF++ with continuous DSSIM targets. For reference, we also include image-payload results collected by SADBench on DIV2K (Sun et al., 2026b). These results follow a different protocol from our FF++ evaluation.

**Metrics and evaluation protocols.** For detection, we sample at most 32 frames per test video and report frame-level AUC. We do not compute video-level AUC or aggregate scores per video, and we use no target-domain fine-tuning. Each controlled comparison uses identical preprocessing and evaluation code. Localization is evaluated using PSNR, SSIM, LPIPS (Zhang et al., 2018), and FID (Heusel et al., 2017) under the protocols of the corresponding references. During localization, the localizer receives only the authorized reconstruction, while the original real–fake pair is used solely to construct the training target. Hiding and recovery are evaluated with PSNR, SSIM, LPIPS, and forensic-response correlation. Unauthorized recovery is evaluated by reconstruction PSNR, SSIM, RGB correlation, and face-detection rate.

**Detectors and comparison methods.** Forensics Adapter (FA) (Cui et al., 2025) serves as our main detector and is trained with SBI pseudo-forgeries (Shiohara & Yamasaki, 2022). We evaluate the same trained model on plaintext faces and authorized VeilTrace reconstructions. ConvNeXt-B (Liu et al., 2022) provides an additional controlled comparison, while Xception (Chollet, 2017) supports comparison with StegaFFD. We also include HiNet+Xception, FreqDebias, and CRDA as reported references. Results taken from prior work follow their original protocols and are therefore not directly comparable to our controlled evaluations.

**Implementation and robustness details.** The hiding model is frozen prior to detector training. Native FF++ samples and SBI pseudo-forgeries are passed through the same hide–recover pipeline, and the detector is optimized with binary cross-entropy. The hiding model and the detector feature encoder are then frozen for localization training. For learned unauthorized recovery, identity-disjoint splits contain 96 attacker-training and 80 evaluation samples, with 24 overlapping rows removed. A U-Net attacker is trained for 1,500 AdamW updates with learning rate $10^{-3}$, weight decay $10^{-4}$, and loss $\ell_1 + 0.2\,\ell_2^2$. It receives either $x_t$ or $[x_t, x_c, x_t - x_c]$. Wrong-key and wrong-nonce tests alter only the corresponding decoding input. Carrier sensitivity uses ten fixed-cover reassignments per held-out secret without retraining. For robustness, JPEG quality 70, resizing, a 95% crop, an RGB–YCbCr–RGB round trip, and Gaussian noise with $\sigma = 5/255$ are applied to the protected image before recovery. These tests characterize the evaluated attacks and channels, not cryptographic secrecy.

Table 1: Hiding and authorized recovery. Best and second-best values are bold and underlined. † denotes SADBench results on DIV2K (Sun et al., 2026b). ‡ marks PUSNet/WIN: Rec. PSNR/SSIM follow WIN (Huang & Huang, 2026); LPIPS/Corr./AUC are our measurements.

(a) Hiding and authorized recovery

| Method | Venue | Stego PSNR↑ | Stego SSIM↑ | Rec. PSNR↑ | Rec. SSIM↑ | LPIPS↓ |
|---|---|---|---|---|---|---|
| HiNet† (Jing et al., 2021) | ICCV'21 | 30.24 | 0.884 | 36.83 | 0.969 | 0.009 |
| DeepMIH† (Guan et al., 2023) | TPAMI'23 | **43.96** | **0.994** | <u>47.19</u> | <u>0.997</u> | **0.000** |
| PRIS† (Yang et al., 2024) | EAAI'24 | 30.75 | 0.867 | 29.86 | 0.925 | 0.045 |
| Ours | Ours | <u>38.02</u> | <u>0.975</u> | **49.76** | **0.9974** | <u>0.00072</u> |

(b) Forensic-response consistency ablation

| Variant | Venue | Rec. PSNR↑ | Rec. SSIM↑ | LPIPS↓ | Forensic Corr.↑ | Mean AUC↑ |
|---|---|---|---|---|---|---|
| PUSNet‡ (Li et al., 2024a) | CVPR'24 | 26.94 | 0.8284 | 0.120 | 0.522 | 63.53 |
| WIN‡ (Huang & Huang, 2026) | TPAMI'26 | <u>49.46</u> | <u>0.9969</u> | <u>0.00085</u> | <u>0.935</u> | <u>80.40</u> |
| w/o $\mathcal{L}_{\text{for}}$ | Ours | 49.06 | <u>0.9969</u> | 0.00094 | 0.934 | 80.26 |
| Full model | Ours | **49.76** | **0.9974** | **0.00072** | **0.953** | **81.58** |

Table 2: Cross-dataset frame-level AUC (%). Best and second-best directly compared results are bold and underlined. † and ‡ denote reported results under the source protocols. (Raw) denotes detection on plaintext faces, whereas StegaFFD and Ours are privacy-preserving detection pipelines.

| Method | Venue | CDF1 | CDF2 | DFD | DFDC | DFDCP | FSh | UADFV | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| RECCE (Raw)‡ (Cao et al., 2022) | CVPR'22 | 76.80 | 73.20 | 81.20 | 71.30 | 73.40 | – | – | 75.18 |
| UCF (Raw)‡ (Yan et al., 2023) | ICCV'23 | 77.90 | 75.30 | 80.70 | 71.90 | 75.90 | – | – | 76.34 |
| LSDA (Raw)‡ (Yan et al., 2024) | CVPR'24 | 86.70 | 83.00 | 88.00 | 73.60 | 81.50 | – | – | 82.56 |
| ProDet (Raw)‡ (Cheng et al., 2024) | NeurIPS'24 | 90.94 | 84.48 | – | 72.40 | 81.16 | – | – | 82.25 |
| FreqDebias (Raw)‡ (Kashiani et al., 2025) | CVPR'25 | 87.50 | 83.60 | 86.80 | 74.10 | 82.40 | – | – | 82.88 |
| CRDA (Raw)‡ (Chou et al., 2026) | AAAI'26 | 90.10 | 85.36 | – | 74.29 | 79.73 | – | 95.70 | 85.04 |
| Xception (Raw)† | CVPR'17 | 74.85 | **75.97** | **82.11** | **69.19** | 68.11 | <u>59.48</u> | **88.01** | **73.96** |
| HiNet + Xception (Raw)† | ICCV'21 | <u>74.94</u> | 66.94 | 70.40 | 63.85 | 64.26 | 47.01 | 80.49 | 66.84 |
| StegaFFD† (Ma et al., 2026) | MIR'26 | **81.22** | <u>72.43</u> | <u>80.64</u> | 66.40 | <u>71.03</u> | 58.68 | 73.63 | 72.00 |
| Ours (Xception) | Ours | 73.23 | 70.82 | 75.05 | <u>67.17</u> | **71.92** | **65.16** | <u>85.35</u> | <u>72.67</u> |
| ConvNeXt-B (Raw) (Liu et al., 2022) | CVPR'22 | **81.54** | <u>81.16</u> | **87.33** | **76.03** | **84.21** | **72.48** | **94.21** | **82.42** |
| Ours (ConvNeXt) | Ours | <u>80.67</u> | **81.57** | <u>86.87</u> | <u>75.48</u> | <u>84.10</u> | <u>72.07</u> | <u>90.27</u> | <u>81.58</u> |
| FA (Raw) (Cui et al., 2025) | CVPR'25 | 85.25 | 86.46 | 91.88 | 81.96 | 83.41 | **81.47** | **98.12** | <u>86.94</u> |
| FA+SBI (Raw) | CVPR'25 | <u>87.90</u> | **88.48** | <u>93.48</u> | <u>85.73</u> | **90.99** | <u>80.97</u> | 95.33 | **88.98** |
| Ours (FA) | Ours | **88.03** | <u>87.68</u> | **93.77** | **85.89** | <u>89.99</u> | 79.40 | <u>98.08</u> | **88.98** |

### 4.2 Image Hiding and Extraction Performance

Recoverable forensics must conceal identity-bearing content without degrading the manipulation-sensitive evidence required after authorized recovery. Table 1 therefore evaluates protected-image fidelity, recovery quality, and forensic-response preservation. The SADBench values are included as cross-protocol references because they were measured on DIV2K rather than our FF++ split.

VeilTrace achieves the highest reported recovery PSNR and SSIM among the listed methods, while retaining the second-highest protected-image fidelity. Since these external rows use mixed protocols, this comparison is indicative rather than same-protocol. Within our controlled ablation, $\mathcal{L}_{\text{for}}$ improves every metric, increasing forensic correlation from 0.934 to 0.953 and mean frame AUC from 80.26% to 81.58%. This gain despite both variants exceeding 49 dB PSNR shows that RGB fidelity alone does not ensure forensic preservation. The result supports using forensic-response consistency as a complement to pixelwise reconstruction losses because it constrains evidence relevant to downstream analysis that is not reflected by RGB metrics.

### 4.3 Cross-Dataset Detection

Table 2 reports cross-dataset results using FA as the main detector, with ConvNeXt-B and Xception comparisons. StegaFFD detects directly in the stego domain, whereas VeilTrace applies conven-

Table 3: Fine-grained localization. † denotes results reported under the DiffusionFF protocol. Best and second-best values are bold and underlined.

(a) FF++ DSSIM-map evaluation

| Method | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ |
|---|---|---|---|---|
| LRL† (Chen et al., 2021) | 23.608 | 0.595 | 0.326 | 241.625 |
| LiSiam† (Wang et al., 2022) | 20.753 | 0.408 | 0.426 | 292.149 |
| DiffusionFF† (Peng et al., 2026) | <u>26.376</u> | <u>0.718</u> | <u>0.198</u> | <u>43.093</u> |
| Ours | **40.655** | **0.864** | **0.100** | **39.174** |

(b) Geometry-aware target construction

| Supervision | Core PSNR↑ | Core SSIM↑ | Corr.↑ | Real p95↓ |
|---|---|---|---|---|
| Naive DSSIM | 14.09 | <u>0.635</u> | 0.270 | **0.0019** |
| Valid only | <u>15.47</u> | 0.231 | 0.481 | 0.0197 |
| Mask only | 8.83 | 0.517 | <u>0.665</u> | 0.1680 |
| Hybrid-DSSIM | **17.12** | **0.741** | **0.682** | <u>0.0071</u> |

tional detectors after authorized recovery; the other methods operate on plaintext faces. FreqDebias and CRDA are external references evaluated under their source protocols. After recovery, VeilTrace matches the 88.98% rounded mean AUC of FA+SBI, exceeds the reproduced FA baseline by 2.04 points, and performs best on three of seven target datasets. In the controlled ConvNeXt-B comparison, recovery reduces mean AUC by 0.84 points, while $\mathcal{L}_{\text{for}}$ improves it by 1.32 points over the ablation. With Xception, VeilTrace has a higher mean AUC than HiNet+Xception and StegaFFD; however, these reported baselines follow the StegaFFD protocol and do not permit a strict same-protocol ranking.

### 4.4 Fine-Grained Artifact Localization

Fine-grained localization can be corrupted by residual misalignment between real–fake pairs, producing false supervision outside manipulated regions. Table 3 and Fig. 4 therefore evaluate localization and compare geometry-aware Hybrid-DSSIM with alternative target constructions. On a 1,024-pair diagnostic probe, only 13.1% of FF++ pairs satisfy all conditions for full DSSIM. In the selected localization validation set, 160 of 1,453 fake targets use full DSSIM (11.0%), while 1,293 use the region-restricted fallback (89.0%). VeilTrace obtains the best PSNR, SSIM, LPIPS, and FID among the listed localization results, with a 14.279 dB PSNR gain over DiffusionFF. These literature values follow the DiffusionFF protocol and provide context rather than a new same-code benchmark. In the controlled target ablation, Hybrid-DSSIM ranks first in core PSNR, SSIM, and correlation while maintaining a low response on real images. Mask-only supervision produces coarse maps and the largest false response. Although Naive DSSIM has the lowest real p95, 91.1% of its fake predictions collapse to near-black maps. The qualitative results show that Hybrid-DSSIM better preserves graded artifact

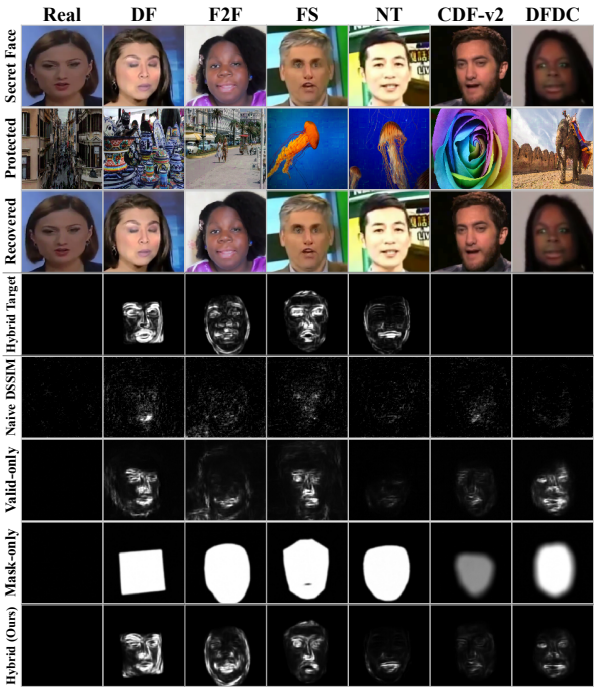


Figure 4: Qualitative hiding, recovery, and localization on FF++ and Celeb-DF-v2/DFDC. Naive DSSIM is displayed at $10\times$, while other maps use the original $[0, 1]$ scale.

responses. This is useful for localization because continuous targets can represent artifact strength, whereas poorly aligned or binary targets may suppress manipulated regions or produce responses on real images.

### 4.5 Sensitivity to Post-Transmission Distortions

Table 4 measures the effects of compression, resampling, cropping, color conversion, and noise on recovery and downstream detection. The color round trip retains 81.32% AUC and 45.56 dB recovery, close to the clean setting. In contrast, JPEG compression, resizing, and cropping disrupt the invertible path, reducing recovery to 12–13 dB and AUC to near chance. Gaussian noise causes intermediate degradation. VeilTrace is therefore robust to the tested color conversion but remains sensitive to lossy and geometric transformations. This result reflects the nominal channel assumption of the method and identifies post-transmission editing as a practical deployment limitation.

Table 4: Robustness to post-transmission distortions. Row-wise best and second-best values are bold and underlined.

| Metric | Clean | JPEG | Resize | Crop | Color | Noise |
|---|---|---|---|---|---|---|
| Raw AUC (%) | 82.42 | 81.10 | 82.40 | **82.88** | 82.53 | 80.53 |
| Ours AUC (%) | **81.58** | 50.15 | 50.87 | 50.88 | 81.32 | 60.43 |
| Rec. PSNR | **49.76** | 13.01 | 12.77 | 12.44 | 45.56 | 23.70 |
| Forensic Corr. | **0.953** | 0.089 | 0.132 | 0.034 | 0.915 | 0.454 |

### 4.6 Security Validation

Privacy protection is weakened if a keyless observer can reconstruct a recognizable face or if recovery changes substantially with the carrier image. Table 5 therefore evaluates unauthorized-recovery attacks and carrier reassignment under the stated threat model. Authorized recovery reaches 49.81 dB, while wrong-key, wrong-nonce, and learned attacks remain at 11.48–11.64 dB with low RGB correlation. The low face- detection rates under incorrect keys and nonces further indicate that these outputs are generally not usable as recognizable faces in the tested attacks. Providing the cover does not improve learned recovery. Across ten carrier reassignments, recovery PSNR, forensic correlation, and held-out AUC remain nearly unchanged, showing limited sensitivity to cover selection. These results support the claimed privacy behavior only for the evaluated threat model and do not imply cryptographic secrecy or zero identity leakage.

Table 5: Unauthorized recovery and carrier reassignment. Bold and underline mark the authorized or best result and the strongest attack or second result.

(a) Unauthorized recovery

| Metric | Authorized | Wrong key | Wrong nonce | Learned | + cover |
|---|---|---|---|---|---|
| Rec. PSNR | **49.81** | 11.52 | 11.48 | 11.50 | 11.64 |
| Rec. SSIM | **0.9973** | 0.1398 | 0.1362 | 0.2336 | 0.2370 |
| RGB Corr. | **0.9999** | 0.0540 | 0.0445 | 0.1972 | 0.2071 |
| Face det. (%) | **100.0** | 8.8 | 5.0 | 40.0 | 33.8 |

(b) Carrier sensitivity under reassignment

| Metric | Original mapping | 10 reassignments |
|---|---|---|
| Stego PSNR↑ | 38.02 | $\mathbf{38.06 \pm 0.04}$ |
| Stego SSIM↑ | **0.9750** | $0.9749 \pm 0.0008$ |
| Rec. PSNR↑ | 49.76 | $\mathbf{49.79 \pm 0.02}$ |
| Forensic Corr.↑ | 0.9532 | $\mathbf{0.9535 \pm 0.0002}$ |
| FF++ held-out AUC (%)↑ | **99.73** | $99.70 \pm 0.10$ |

## 5 Conclusion

Remote face-forgery analysis must balance forensic access against the privacy risks of transmitting identity-bearing facial content. Encryption can conceal image content but may expose the presence of protected communication to traffic observers, while decrypt–then–infer services expose plaintext at the inference endpoint. Secure-computation and stego-domain methods avoid some of these exposures but require specialized protocols or limit compatibility with conventional forensic models. In this paper, we propose VeilTrace, a recoverable alternative that transmits a cover-like protected image, reconstructs the face only at an authorized endpoint, and preserves downstream evidence through forensic-response consistency and geometry-aware Hybrid-DSSIM supervision. Experiments verify VeilTrace preserves reconstruction and forgery detection performance after authorized recovery, and that Hybrid-DSSIM improves artifact-map alignment in localization.

## AI USE DISCLOSURE

The authors used large language model (LLM) assistants to polish and restructure the manuscript, retrieve related work, and cross-check references and reported numbers. The authors conceived the threat model, adversarial-carrier analysis, method design, experiments, and interpretation of results. They also conducted all experiments. All LLM-generated text and suggestions were reviewed, verified, and edited by the authors, who take full responsibility for the content of this paper.

## REFERENCES

Haonan An, Xiaohui Ye, Guang Hua, Yihang Tao, Hangcheng Cao, Xiangyu Yu, and Yuguang Fang. RecoverMark: Robust watermarking for localization and recovery of manipulated faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8587–8597, 2026.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4122, 2022.

Beijing Chen, Xin Liu, Zhihua Xia, and Guoying Zhao. Privacy-preserving DeepFake face image detection. *Digital Signal Processing*, 143:104233, 2023.

Lixin Chen, Chaomeng Chen, Jiale Zhou, Zhijian Wu, and Xun Lin. StegaVAR: Privacy-preserving video action recognition via steganographic domain analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(4):2930–2938, 2026.

Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. Local relation learning for face forgery detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1081–1088, 2021.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *Advances in Neural Information Processing Systems*, volume 37, pp. 21979–21998, 2024.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

Yuxuan Chou, Tao Yu, Wen Huang, Yuheng Zhang, Tao Dai, and Shu-Tao Xia. Improving deepfake detection with reinforcement learning-based adaptive data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3381–3389, 2026.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics Adapter: Adapting CLIP for generalizable face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217, 2025.

Xin Deng, Chao Gao, and Mai Xu. PIRNet: Privacy-preserving image restoration network via wavelet lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22368–22377, 2023.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset, 2019. arXiv preprint arXiv:1910.08854.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake detection challenge (DFDC) dataset, 2020. arXiv preprint arXiv:2006.07397.

Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. DeepFakes detection dataset by Google and Jigsaw, 2019.

Fangyuan Gao, Chao Gao, Xin Deng, Chenxiao Zhang, Junjie Huang, and Mai Xu. AIRPNet: Adaptive image restoration with privacy protection in steganographic domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):8148–8165, 2026.

Zhenyu Guan, Junpeng Jing, Xin Deng, Mai Xu, Lai Jiang, Zhou Zhang, and Yipeng Li. DeepMIH: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):372–390, 2023.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20606–20615, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Juan Hu, Xin Liao, Difei Gao, Satoshi Tsutsui, Qian Wang, Zheng Qin, and Mike Zheng Shou. Delocate: Detection and localization for deepfake videos with randomly-located tampered traces. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5862–5871, 2024.

Yuanfei Huang and Hua Huang. Tackling ill-posedness of reversible image conversion with well-posed invertible network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (6):6847–6864, 2026.

Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. HiNet: Deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4733–4742, 2021.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. FreqDebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8775–8785, 2025.

Minyoung Kim and Paul Hongsuck Seo. Robust image self-recovery against tampering using watermark generation with pixel shuffling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8877–8886, 2026.

Guobiao Li, Sheng Li, Zicong Luo, Zhenxing Qian, and Xinpeng Zhang. Purified and unified steganographic network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27569–27578, 2024a.

Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. FreqBlender: Enhancing DeepFake detection by blending frequency knowledge. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44965–44988, 2024b.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5074–5083, 2020a.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-Ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5001–5010, 2020b.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei 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)*, pp. 3207–3216, 2020c.

Xun Lin, Yi Yu, Zitong Yu, Ruohan Meng, Jiale Zhou, Ajian Liu, Yizhong Liu, Shuai Wang, Wenzhong Tang, Zhen Lei, and Alex C. Kot. HideMIA: Hidden wavelet mining for privacy-enhancing medical image analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8110–8119, 2024.

Decheng Liu, Zhan Dang, Chunlei Peng, Yu Zheng, Shuang Li, Nannan Wang, and Xinbo Gao. FedForgery: Generalized face forgery detection with residual federated learning. *IEEE Transactions on Information Forensics and Security*, 18:4272–4284, 2023.

Jun Liu, Jiantao Zhou, Jinyu Tian, and Weiwei Sun. Recoverable privacy-preserving image classification through noise-like adversarial examples. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(7):216:1–216:27, 2024.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022.

Guoqing Ma, Xun Lin, Hui Ma, Ajian Liu, Yizhong Liu, Wenzhong Tang, Shan Yu, Chenqi Kong, and Yi Yu. StegaFFD: Privacy-preserving face forgery detection via fine-grained steganographic domain lifting. *Machine Intelligence Research*, 23(2):409–428, 2026.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17395–17405, 2024.

Siran Peng, Haoyuan Zhang, Li Gao, Tianshuo Zhang, Xiangyu Zhu, Bao Li, Weisong Zhao, and Zhen Lei. DiffusionFF: A diffusion-based framework for joint face forgery detection and fine-grained artifact localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14095–14105, 2026.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, 2022.

Jiamu Sun, Zhiyuan Yan, Ke-Yue Zhang, Taiping Yao, and Shouhong Ding. DFD-HR: Generalizable deepfake detection via hierarchical routing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13984–13995, 2026a.

Zhen Sun, Zongmin Zhang, Leyi Sheng, Yule Liu, Yifan Liao, Ke Li, Xinhu Zheng, Jiaheng Wei, Wenyuan Yang, and Xinlei He. Stego battlefield: Evaluating image steganography attacks and steganalysis defenses, 2026b. arXiv preprint arXiv:2605.05789.

Jian Wang, Yunlian Sun, and Jinhui Tang. LiSiam: Localization invariance siamese network for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 17:2425–2436, 2022.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22412–22423, 2023.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8984–8994, 2024.

Hang Yang, Yitian Xu, Xuhua Liu, and Xiaodong Ma. PRIS: Practical robust invertible network for image steganography. *Engineering Applications of Artificial Intelligence*, 133:108419, 2024.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, 2019.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72925–72943, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

# A Algorithmic Details

The main paper gives the mathematical formulation of recoverable hiding and Hybrid-DSSIM. We provide the corresponding procedures here to make the execution order and branching conditions explicit.

## A.1 Keyed Hiding and Authorized Recovery

**Algorithm 1: Keyed hiding and authorized recovery**

**Input:** secret face $x_s$, cover $x_c$, secret key $K$, public nonce $n$
**Output:** protected image $x_t$, authorized reconstruction $\hat{x}_s$

$z_s \leftarrow W_\omega^{\mathrm{enc}}(x_s)$
Partition $z_s$ into non-overlapping $16 \times 16$ patches
$d \leftarrow \mathrm{HMAC\text{-}SHA256}(K, n)$
Derive deterministic permutation $\pi_{K,n}$ from the first eight bytes of $d$
$\bar{z}_s \leftarrow S_{\pi_{K,n}}(z_s)$
$p_0 \leftarrow \mathrm{DWT}(\bar{z}_s)$; $c_0 \leftarrow \mathrm{DWT}(x_c)$
**for** $i = 1, \ldots, L$ **do**
    $p_i \leftarrow p_{i-1} - P_i(c_{i-1})$
    $c_i \leftarrow c_{i-1} + U_i(p_i)$
**end for**
$x_t \leftarrow Q_8(\mathrm{IWT}(c_L))$
**Authorized recovery:**
$\tilde{c}_L \leftarrow \mathrm{DWT}(x_t)$
$\tilde{p}_L \leftarrow \mathrm{SidePredictor}(\tilde{c}_L)$
**for** $i = L, \ldots, 1$ **do**
    $\tilde{c}_{i-1} \leftarrow \tilde{c}_i - U_i(\tilde{p}_i)$
    $\tilde{p}_{i-1} \leftarrow \tilde{p}_i + P_i(\tilde{c}_{i-1})$
**end for**
$\tilde{z}_s \leftarrow \mathrm{IWT}(\tilde{p}_0)$
$z'_s \leftarrow S^{-1}_{\pi_{K,n}}(\tilde{z}_s)$
$\hat{x}_s \leftarrow W_\omega^{\mathrm{dec}}(z'_s)$
**return** $x_t, \hat{x}_s$

During hiding training, the recovered branch is additionally constrained by the forensic-response consistency loss in Eq. 7. The response computed from the original face is stop-gradient, so gradients from $\mathcal{L}_{\mathrm{for}}$ propagate only through the recovered branch.

## A.2 Hybrid-DSSIM Target Construction

A forged pair with unreliable geometry and no usable manipulation mask is excluded from the localization loss. SBI pairs originate from the same source face and directly provide aligned full-DSSIM supervision.

# B Additional Experiments

## B.1 Multi-Trial Unauthorized Recovery

The single wrong-key experiment in Table 5 evaluates one mismatched credential. To test whether its low recovery quality depends on the particular wrong key selected for evaluation, we perform a finite multi-trial stress test.

For each sample in the identity-disjoint security evaluation set, we generate 32 deterministic wrong keys and decode the same protected image with every candidate. The ground-truth private face is used only to select the reconstruction with the highest PSNR. This is an oracle procedure: a realistic keyless observer does not know the ground-truth private face and therefore cannot perform this selection in practice.

**Algorithm 2: Hybrid-DSSIM target construction**

**Input:** forged face $x_f$, pristine counterpart $x_r$, label $y$, landmarks, optional manipulation mask $M$
**Output:** localization target $M^*$, target-validity indicator $v$

**if** $y = 0$ **then**
    $M^* \leftarrow 0$; $v \leftarrow 1$; **return**
Estimate integer translation $\Delta$ from stable landmarks
Align $x_r$ by $\Delta$ and compute valid overlap $O$
$M_{\text{full}} \leftarrow O \odot [1 - \text{SSIM}(x_f, A_\Delta(x_r))]$
Compute residual landmark error $e_{\text{lm}}$
Compute static-region DSSIM $e_{\text{static}}$
$g \leftarrow \mathbf{1}[\|\Delta\|_\infty \leq 8,\ e_{\text{lm}} \leq 0.012,\ e_{\text{static}} \leq 0.10]$
**if** $g = 1$ **then**
    $M^* \leftarrow M_{\text{full}}$; $v \leftarrow 1$
**else if** a usable manipulation mask is available **then**
    $M_{\text{soft}} \leftarrow \text{Blur}(\text{Dilate}(M))$
    $M^* \leftarrow M_{\text{full}} \odot M_{\text{soft}}$; $v \leftarrow 1$
**else**
    $v \leftarrow 0$
**end if**
**return** $M^*, v$

We repeat the procedure with 32 mismatched nonces. Since the nonce is public in our threat model, this experiment measures sensitivity to nonce mismatch rather than secrecy of the nonce.

Table 6: Multi-trial mismatched-credential stress test. "Best of 32" uses the ground-truth private face only to select the highest-PSNR reconstruction among 32 mismatched candidates for each sample.

| Metric | Authorized | Wrong key | Best of 32 keys | Wrong nonce | Best of 32 nonces |
|---|---|---|---|---|---|
| Rec. PSNR↑ | **49.81** | 11.52 | 11.97 | 11.48 | 11.97 |
| Rec. SSIM↑ | **0.9973** | 0.1398 | 0.1508 | 0.1362 | 0.1531 |
| RGB Corr.↑ | **0.9999** | 0.0540 | 0.1488 | 0.0445 | 0.1491 |
| Face det. (%)↑ | **100.0** | 8.8 | 2.5 | 5.0 | 6.3 |

Selecting the highest-PSNR reconstruction among 32 wrong keys increases recovery PSNR from 11.52 to 11.97 dB. Authorized recovery remains at 49.81 dB, leaving a 37.84 dB gap. RGB correlation increases from 0.0540 to 0.1488 under oracle selection but remains far below the authorized value of 0.9999.

The best-of-32 nonce experiment gives a similar result, with 11.97 dB PSNR and 0.1491 RGB correlation. Within this finite candidate set, the low recovery quality of the single mismatched-credential test is therefore not explained solely by the particular wrong key or nonce selected in that experiment.

Candidate selection is based on reconstruction PSNR rather than face-detection output. Face-detection rate is therefore an auxiliary diagnostic and is not expected to change monotonically with the number of candidates. This experiment is not an exhaustive key search and does not establish cryptographic secrecy or zero identity leakage.

### B.2 Localization Specificity and Sampling Stability

The controlled localization ablation in Table 3 measures agreement between the predicted artifact map and its corresponding target. We further examine whether the predictions retain sample-specific spatial information, whether they are stable across diffusion seeds, and whether they collapse to nearly empty maps.

**Pair-specific correspondence gain.** For fallback samples, we compare correlation with the correct core target against correlation with a different target from the same manipulation method:

$$\Delta_{\text{pair}} = \text{Corr}(\hat{M}_i, M_i) - \text{Corr}(\hat{M}_i, M_j), \tag{10}$$

where $i \neq j$ and samples $i$ and $j$ share the same manipulation method. A positive value indicates sample-specific spatial correspondence beyond a method-level pattern.

**Seed-map correlation.** For each input, two artifact maps are generated using deterministic diffusion seeds derived from $s$ and $s + 100000$. Their correlation measures the consistency of the predicted localization map across the two sampling runs.

**Fake-black rate.** We report the fraction of forged predictions whose mean response is below 0.005. This diagnostic identifies near-empty artifact maps.

**Minimum per-method correlation.** Fallback core-target correlation is computed separately for DeepFakes, Face2Face, FaceSwap, and NeuralTextures, and the minimum value is reported. This provides a conservative measure of performance across manipulation types.

Table 7: Localization specificity and sampling stability using the same checkpoints and evaluation split as the controlled target-construction ablation. Best and second-best values are bold and underlined.

| Supervision | Pair gain↑ | Seed Corr.↑ | Black rate↓ | Min. method Corr.↑ |
|---|---|---|---|---|
| Naive DSSIM | 0.030 | 0.431 | 91.1% | 0.149 |
| Valid only | 0.045 | 0.561 | **0.1%** | 0.344 |
| Mask only | <u>0.049</u> | **0.966** | <u>0.2%</u> | <u>0.523</u> |
| Hybrid-DSSIM | **0.120** | <u>0.728</u> | 2.5% | **0.556** |

Hybrid-DSSIM gives the largest pair-specific correspondence gain (0.120), compared with 0.049 for the next-best variant, and the highest minimum per-method correlation (0.556). These results indicate that its fallback predictions retain stronger sample-specific spatial correspondence and that the improvement is not confined to a single manipulation type.

Mask-only gives the highest seed-map correlation (0.966). Seed consistency alone, however, does not measure localization quality: in the controlled main-paper ablation, Mask only has substantially lower core PSNR than Hybrid-DSSIM and produces the largest response on real images. Hybrid-DSSIM gives the second-highest seed correlation (0.728).

Naive DSSIM produces near-black predictions for 91.1% of forged samples. Hybrid-DSSIM reduces this rate to 2.5%. Valid-only and Mask-only have still lower black rates, showing that avoidance of near-empty predictions alone is also insufficient to determine localization quality.

Together with the main localization metrics, these diagnostics show that Hybrid-DSSIM provides the strongest sample-specific correspondence and the highest worst-method correlation while retaining substantial sampling consistency and largely avoiding the collapse observed with Naive DSSIM.

All diagnostics use the same trained checkpoints and evaluation split as the controlled target-construction ablation. No model is retrained for this analysis.

# C Experimental Details

## C.1 Training and Controlled Comparisons

The recoverable hiding model is trained for 30 epochs with batch size 16 using AdamW. The learning rates for the lifting network, secret-representation transform, and side predictor are $2 \times 10^{-5}$, $1 \times 10^{-5}$, and $5 \times 10^{-5}$, respectively. The forensic-response weight is $\lambda_f = 0.40$.

Checkpoint selection for the hiding ablation uses visual reconstruction and cover-quality terms only. Neither forensic-response correlation nor downstream detector accuracy is used to select the full model or the model without $\mathcal{L}_{\text{for}}$.

The main Forensics Adapter detector is trained for five epochs with Adam, learning rate $2 \times 10^{-4}$, batch size 16, and weight decay $5 \times 10^{-4}$. The same detector-training recipe and detector-only augmentations are used in the matched Raw and recovered-domain comparisons.

For the controlled localization ablation, all supervision variants use the same frozen ConvNeXt-B feature encoder and the same diffusion architecture. Artifact maps are generated at $128 \times 128$ resolution using 50 diffusion steps with a linear beta schedule from 0.02 to 0.40. The localization U-Net is optimized with AdamW using learning rate $10^{-4}$. Validation is performed every five epochs. Early stopping is enabled only after at least 30 epochs, and training is capped at 120 epochs.

Thus, the controlled localization variants differ in target construction rather than detector features or localizer architecture.

### C.2 Hybrid-DSSIM and Forensic-Response Settings

Hybrid-DSSIM uses the full-target geometry thresholds

$$\tau_s = 8, \qquad \tau_l = 0.012, \qquad \tau_d = 0.10. \tag{11}$$

The fallback manipulation support uses dilation 7 followed by Gaussian blur with kernel size 9. A pair without reliable geometry and without a usable manipulation mask does not contribute to the localization loss.

The fixed forensic-response operator $\Phi_f$ contains two Laplacian responses, horizontal and vertical Sobel responses, a diagonal-difference response, the mean response of two fixed $5 \times 5$ high-pass filters, a multi-scale local-texture response, a chromatic residual, and a short-range periodic-difference response.

Signed filter responses use local magnitude normalization

$$N_k(r) = \tanh\left(\frac{r}{A_k(|r|) + 10^{-4}}\right), \tag{12}$$

where $A_k(\cdot)$ denotes local averaging. Laplacian, Sobel, diagonal, and high-pass responses use a $9 \times 9$ normalization neighborhood, whereas chromatic and periodic responses use a $7 \times 7$ neighborhood. The texture response combines local statistics at scales 3, 9, and 17.

The subsequent response gating, channel mixing, energy normalization, spatial sign modulation, and response quantization are fixed and non-learnable. The same operator is applied to the original and recovered faces, and its output is used only by the forensic-response consistency loss.

### C.3 Security Protocol

The security experiments use the fixed 200-sample FF++ held-out set.

For learned unauthorized recovery, identities are partitioned before attacker training. The retained identity-disjoint split contains 96 attacker-training samples and 80 evaluation samples; 24 rows crossing the two identity partitions are removed.

The learned attacker is an encoder–decoder U-Net with base width 32. The stego-only attacker receives $x_t$, whereas the known-cover attacker receives

$$[x_t,\ x_c,\ x_t - x_c]. \tag{13}$$

Both attackers are trained for 1,500 AdamW updates using learning rate $10^{-3}$ and

$$\mathcal{L}_{\text{atk}} = \|\hat{x} - x_s\|_1 + 0.2\|\hat{x} - x_s\|_2^2. \tag{14}$$

All learned-attack results are evaluated on the identity-disjoint 80-sample evaluation set.

Wrong-key and wrong-nonce experiments modify only the corresponding recovery input. For the multi-trial experiment in Appendix B.1, wrong-key candidates are generated deterministically from the fixed experiment seed and trial index; mismatched nonces use deterministic trial-specific suffixes.

Carrier sensitivity is evaluated without retraining. Each held-out secret is assigned ten alternative covers while excluding its original cover. Recovery and downstream metrics are recomputed for every reassignment.

Face-detection rate is used only as an auxiliary recognizability diagnostic and is not treated as an identity-recognition or identity-leakage measure.

### C.4 Post-Transmission Distortion Evaluation

The post-transmission experiment in Table 4 evaluates the clean-trained VeilTrace model under fixed transformations; no distortion-aware fine-tuning is used for the reported results.

The evaluated conditions are JPEG quality 70; resizing from $256 \times 256$ to $192 \times 192$ and back to $256 \times 256$; a centered 95% crop followed by resizing to the original resolution; an 8-bit RGB–YCbCr–RGB round trip; and Gaussian noise with $\sigma = 5/255$.

Each transformation is applied to the protected image before authorized recovery. For the Raw baseline, the same transformation type and severity are applied directly to the raw face.

Cross-dataset AUC is the unweighted mean over the same seven target datasets used in the main detection experiment. No target-domain fine-tuning is used. Recovery PSNR and forensic-response correlation are measured on the fixed FF++ hiding evaluation set.