Reversible Visual Obfuscation via Embedded Recovery Metadata: Combining Median-Filter Irreversibility with Cryptographic Restoration

Andranik Karakhanyan
Department of Mathematical Modeling of Computer Systems, National Polytechnic University of Armenia, Yerevan, Armenia

***Abstract*—Although cloud storage platforms are widely used to safeguard personal images, data leakage and unauthorized access remain persistent threats, exposing sensitive visual content. Linear obfuscation methods such as Gaussian blur can be reversed by deconvolution when kernel parameters are known, while non-linear median filtering permanently destroys details but also eliminates any possibility of legitimate recovery. This study introduces a reversible obfuscation framework that combines the practical irreversibility of large-radius median filtering with an embedded, encrypted restoration payload. A sensitive image region is first blurred with a non-invertible median filter, making it resistant to both deconvolution and AI-based reconstruction. Simultaneously, a losslessly compressed and AES-256-GCM-encrypted copy of the original region is hidden in the least significant bits of the blurred block. Unauthorised viewers—including any party that intercepts the image through a cloud leak—see only the irreversibly obfuscated content; an authorised user with the correct passphrase can decrypt and restore the original region with pixel-perfect fidelity. Experimental results on facial and document images demonstrate that the obfuscated region withstands state-of-the-art AI restoration while the embedded payload enables lossless recovery in under 50 ms, with an average overhead of only 0.41 bits per pixel (using a single embedding channel) and no visible artefacts. The method offers a practical balance between strong, irreversible privacy and cryptographically controlled recoverability, making it suitable for secure archiving, forensic analysis, and privacy-preserving communications.**

Keywords*—Sensitive information; visual privacy; image processing; blur filter; median filtering; data obfuscation; reversible concealment

I. Introduction

The sheer volume of images shared daily on digital platforms has transformed communication while simultaneously exposing individuals to unprecedented privacy risks. In 2024 alone, more than 350 million photographs were uploaded to social media every day [1], and a large-scale survey conducted the following year found that nearly 60% of users had inadvertently posted images containing sensitive personal details—identity documents, home addresses, or financial information—often realizing the exposure only after publication [2]. This phenomenon is not accidental; it is driven by an attention economy that monetizes personal data, where every uploaded image can be harvested, aggregated, and exploited without meaningful user consent [3]. Even when individuals attempt to curate their online presence, once an image enters the public domain, controlling its subsequent use becomes extraordinarily difficult, and reputation damage can be lasting [4].

To mitigate such risks, visual obfuscation techniques—blurring, pixelation, and masking—are routinely applied to hide selected regions within an image [5]. Among these, Gaussian blur is perhaps the most widespread because of its simplicity and natural-looking degradation. A Gaussian filter convolves an image with a smooth kernel, attenuating high-frequency details while preserving broad structural outlines. However, this linearity makes the operation mathematically reversible: if an attacker knows the kernel parameters, deconvolution algorithms such as Wiener filtering can recover a remarkably faithful approximation of the original content [6], [7]. Moreover, deep-learning-based super-resolution and blind deblurring networks can restore substantial detail even when the exact kernel is unknown [8]. Consequently, Gaussian blur alone provides only a thin veil of privacy that a moderately informed adversary can easily pierce.

A stronger alternative is offered by non-linear filters, especially median filtering with a large radius. Unlike convolution, a median filter replaces each pixel with the statistical median of its neighborhood—an operation that discards fine-scale information in a manner that cannot be represented by a linear point-spread function [9]. Theoretical analyses confirm that the mapping from input to output is many-to-one; distinct original regions can map to identical median-filtered results, rendering exact inversion impossible [10], [11]. From a privacy standpoint, this property is highly desirable because it guarantees that even an attacker who fully knows the filter parameters cannot reconstruct the exact original pixel values. Yet this irreversibility also introduces a serious practical limitation: authorized parties who legitimately need to recover the original content—for example, in forensic archiving, secure document handling, or identity verification—are permanently locked out. The filtered image has irretrievably lost the sensitive details.

This dilemma reveals a distinct research gap: current obfuscation methods either offer weak security that can be broken (linear blur) or strong, irreversible protection that precludes any legitimate recovery (large-radius median filter). Separately, the field of reversible data hiding has developed techniques that embed secret information within an image so that the original cover can be perfectly restored after data

extraction [12]. When applied to encrypted images, these methods enable the owner to recover an unaltered original after decryption, but the encrypted image appears as noise and provides no perceptible concealment of the content. To our knowledge, no existing approach combines the robust visual irreversibility of non-linear filtering with a cryptographically controlled restoration capability that is directly embedded in the obfuscated image.

In this study, we address this gap by proposing a reversible obfuscation framework that couples large-radius median filtering—an inherently destructive, non-invertible operation—with an embedded, encrypted restoration payload. The filtered region becomes permanently illegible to any viewer without the decryption key, yet the original image can be recovered with pixel-level fidelity by an authorized user who extracts and decrypts the hidden metadata. The main contributions of this work are:

- A novel obfuscation-restoration architecture that applies a large-radius median filter to irreversibly conceal sensitive regions while simultaneously embedding all information necessary for lossless recovery.
- A complete embedding and extraction algorithm that uses the blurred region as a carrier, ensuring that the obfuscated image remains visually plausible and does not attract suspicion, while the payload is protected by authenticated symmetric encryption (AES-256-GCM).
- Quantitative experimental validation on facial image datasets, demonstrating that Gaussian-blurred regions are trivially inverted by deconvolution and that median-filtered regions resist even modern AI-based restoration, whereas the proposed method achieves identical restoration (PSNR $\rightarrow \infty$) for key-holding users.
- A security analysis that formalizes the adversary model, confirms the computational infeasibility of unauthorized recovery, and discusses the trade-off between obfuscation strength and payload capacity, including the limitations of LSB embedding against JPEG compression and other common image transformations.

The remainder of the study is organized as follows. Section II surveys related work in privacy-preserving obfuscation and reversible data hiding. Section III provides the necessary background on Gaussian and median filtering, emphasizing their invertibility properties. Section IV details the proposed framework, including region selection, payload generation, embedding, and restoration. Section V presents experimental results, comparing our method against baselines and analyzing robustness. Section VI discusses limitations and future work. Section VII concludes the study.

II. Related Work

This section reviews three areas foundational to the proposed framework: visual obfuscation techniques for privacy protection, reversible data hiding, and the irreversibility properties of median filtering. We then position our contribution relative to these existing bodies of work.

A. Visual Obfuscation for Privacy Protection

The protection of sensitive visual information has been an active research theme for over two decades, driven by the proliferation of surveillance cameras, social media, and mobile imaging. Early efforts focused on de-identification of faces in video streams and photographs. Newton et al. [13] proposed a systematic framework for evaluating face de-identification algorithms, showing that simple operations such as blurring and pixelation reduce the risk of automatic face recognition while preserving a degree of contextual information. Cavallaro surveyed broader privacy-preserving video processing methods, including object removal, scrambling, and selective encryption, noting that the choice of obfuscation must balance utility against the strength of the privacy guarantee [5].

A common shortcoming of many widely deployed obfuscation filters is their reversibility under certain conditions. Linear filtering operations—Gaussian blur being the canonical example—are particularly vulnerable. Because the blur kernel defines a convolution, knowledge of its parameters allows an adversary to solve a deconvolution problem. Wiener filtering, for instance, can restore sharp edges and recognizable facial features when the signal-to-noise ratio is favorable [6]. More sophisticated blind deconvolution methods, such as the camera-shake removal algorithm of Fergus et al. [7], do not even require precise kernel knowledge. In the era of deep learning, super-resolution networks can further amplify these risks; Dong et al. [8] demonstrated that convolutional neural networks can recover high-frequency details from low-resolution or blurred inputs with remarkable fidelity. Consequently, any obfuscation scheme that preserves a linear relationship between the clear and obscured images must be regarded as breakable by a technically competent attacker.

More recent work has explored selective encryption approaches, where only sensitive portions of an image are encrypted while leaving the remainder visible [14]. While this provides strong cryptographic protection, encrypted regions appear as noise, which can attract attention and reveal the locations of sensitive content. Privacy-preserving image protection has also been explored through the lens of adversarial machine learning, where perturbations are added to defeat recognition systems while maintaining visual quality. However, these approaches typically do not provide reversible recovery of the original content.

B. Reversible Data Hiding

Reversible data hiding (RDH) refers to techniques that embed additional information into a digital medium in such a way that, after the embedded data are extracted, the original cover medium can be restored without any loss. Early RDH methods exploited lossless compression of bit-plane or difference-expansion spaces. Tian [15] introduced a difference-expansion approach that embeds one bit of payload into the difference value of a pixel pair, achieving high capacity while maintaining exact reversibility.

Ni et al. [16] later proposed a histogram-shifting technique that uses the peak point of the image histogram to embed data with minimal distortion. These foundational works established that it is possible to carry hidden metadata without permanently degrading the image.

A major branch of RDH research addresses the encrypted domain. Zhang [12] proposed a scheme in which a content owner encrypts the original image, and a data hider embeds additional information into the encrypted image without knowing the plaintext content. An authorized receiver with the encryption key can decrypt the image and then recover the original cover by exploiting spatial correlation in the decrypted version. Subsequent work extended this paradigm to separable architectures where the data extraction and image recovery keys are independent [17]. More recent advances have explored RDH in encrypted images with higher capacity and better visual quality, as well as methods that are robust against various image processing operations [18]. While these methods provide strong confidentiality and exact recovery, the encrypted image is perceptually meaningless noise; it does not offer the fine-grained, selective obfuscation needed when only certain regions require concealment while the rest of the image should remain visible for context. Our framework adopts the principle of embedding recoverable metadata but applies it to a region that has been deliberately and strongly obfuscated by a non-linear filter, thereby achieving both visual privacy and recoverability in a single composite image.

C. Median Filtering and Irreversibility

The median filter is a rank-order statistic filter that has been extensively studied for its noise-reduction capabilities and its edge-preserving behavior. Crucially, its non-linear nature makes it fundamentally different from convolution-based blurs. Gallagher and Wise provided a seminal theoretical analysis, showing that the median filter maps multiple distinct input sequences to the same output, destroying information in a manner that cannot be described by a linear point-spread function [9]. This many-to-one mapping implies that no deterministic inverse operator exists. Oppenheim and Schafer further discuss the implications of non-linearity for signal reconstruction, noting that phase and spectral information is irretrievably lost [10].

Efficient implementations of median filtering, such as the constant-time algorithm proposed by Perreault and Hébert [11], have made large-radius median filters practical for real-time applications. As the radius increases, the filter's root signal set shrinks, meaning that a greater proportion of possible input details are annihilated. In the context of privacy, this property is extremely attractive: an image region processed with a sufficiently large-radius median filter carries no recoverable high-frequency information about the original face or text. However, it is important to note that while exact inversion is impossible, probabilistic inference of certain attributes—such as approximate shape, skin tone, or general category—may still be possible [19]. The privacy guarantee is therefore one of computational infeasibility of exact reconstruction rather than absolute information-theoretic privacy.

D. Summary and Positioning

Taken together, the literature reveals a gap between reversible recovery schemes and strong irreversible obfuscation. RDH in encrypted images provides lossless restoration but produces a fully encrypted (and thus entirely unusable) intermediate image. Median filtering provides robust, non-invertible visual privacy but sacrifices any possibility of recovery. Recent work on privacy-preserving image sharing has explored hybrid approaches, such as perceptual encryption combined with reversible hiding, but these typically do not leverage the non-linear irreversibility of median filtering [18]. Our proposed framework bridges these two worlds: it uses the median filter's inherent irreversibility to guarantee privacy against unauthorized viewers while embedding an encrypted restoration payload—inspired by RDH—that enables perfect recovery for keyholders. Consequently, no prior work simultaneously guarantees visual irreversibility (as with median filtering) and authorised lossless recovery (as with RDH in encrypted images)—a duality that the present framework uniquely achieves.

III. Preliminary Concepts

This section provides a concise technical foundation for the two central operations that underpin the proposed framework: Gaussian blur (as a representative of linear, reversible obfuscation) and median filtering (as a non-linear, irreversible alternative). Understanding why one is invertible and the other is not is essential to appreciating the security guarantees of the method presented in Section IV.

A. Gaussian Blur and Deconvolution Vulnerability

A Gaussian blur is a low-pass filtering operation achieved by convolving an image with a two-dimensional Gaussian kernel. The kernel, parameterised by the standard deviation σ, assigns decaying weights spatially to neighbouring pixels:

$$G\sigma(u, v) = 1 / (2\pi\sigma^2) \exp(-(u^2 + v^2) / (2\sigma^2)), \quad (1)$$

where, u and v are the horizontal and vertical distances from the kernel centre. For a discrete input image $f(x, y)$ and a square kernel of half-width k (typically $k = \lceil 3\sigma \rceil$), the blurred image $g(x, y)$ is obtained via discrete convolution:

$$g(x, y) = \sum_{u=-k}^{k} \sum_{v=-k}^{k} G\sigma(u, v) f(x - u, y - v). \quad (2)$$

Because convolution is a linear, shift-invariant operation, the blurring process has a straightforward frequency-domain interpretation. Denoting the discrete Fourier transforms of the input, blurred, and kernel as F, G, and Hσ, respectively, the convolution theorem gives:

$$G(u, v) = H\sigma(u, v) F(u, v). \quad (3)$$

The transfer function Hσ is itself a Gaussian, decaying smoothly toward zero at high spatial frequencies but never reaching zero. This non-zero property implies that, in the absence of noise, the original spectrum F could be recovered exactly by the inverse filter $F = G / H\sigma$. In practice, quantisation and sensor noise introduce small perturbations that render direct division unstable.

Wiener deconvolution overcomes this by applying a regularised inverse filter that minimises the mean square estimation error [6]. When the blur parameters σ and kernel size are known to an attacker, Wiener deconvolution can restore edges, textures, and even facial features with high fidelity. Blind deconvolution algorithms that do not require exact kernel knowledge, such as those based on natural image priors [20], further reduce the difficulty of an attack. Super-resolution networks can also reconstruct high-frequency detail from blurred inputs in a single forward pass [8]. The vulnerability of Gaussian blur is not merely theoretical; a practical demonstration of Wiener deconvolution on a Gaussian-blurred facial image, available in the image processing literature [6], shows that facial features can be recovered with high fidelity when the kernel parameters are known. Collectively, these results demonstrate that linear blurring scatters information rather than destroying it; the privacy it affords is fragile against an informed adversary.

B. Median Filtering and the Irreversibility Property

The median filter is a non-linear rank-order statistic filter that replaces each pixel with the median intensity value inside a predefined window [21]. For a square window W_xy of odd side length r centred at pixel (x, y), the output is:

$$g_{med}(x, y) = \text{median} \{ f(i, j) \mid (i, j) \in W\_xy \}. \quad (4)$$

Unlike the weighted averaging in (2), (4) performs no arithmetic combination of neighbouring values; it merely selects one of the observed intensities. This operation is idempotent—repeated application of the same filter produces no further change—and, more importantly, it is a many-to-one mapping. For any window larger than 1 × 1, multiple distinct input configurations can yield the identical median output [9]. There is no unique inverse, because the intensity order information and fine spatial gradients that were present in the original window are discarded.

The non-linear destruction of information distinguishes median filtering sharply from Gaussian blur. Whereas the convolution in (2) redistributes energy in a deterministic and recoverable manner, the median operation annihilates it in a way that cannot be represented by any linear point-spread function. As analysed by Oppenheim and Schafer [10], phase and amplitude relationships within the signal are lost, and no filter—linear or non-linear—can reconstruct them when only the filtered data are available. Efficient constant-time algorithms [11] make large-radius median filtering practical for high-resolution images, and as the window radius grows, the set of possible pre-images expands combinatorially. At a radius of, for example, 12 pixels (a 25 × 25 window), only coarse monotonic edges survive; fine textures, curves, and recognisable facial details are permanently removed.

This mathematical irreversibility is the key property that the proposed framework exploits for privacy protection. An adversary in possession of a median-filtered region cannot recover the exact original content, no matter what computational resources are available. However, it is important to acknowledge that while exact reconstruction is impossible, probabilistic inference may still reveal certain high-level attributes, such as the approximate location of facial features or the presence of text [19]. The privacy guarantee is therefore one of computational infeasibility of exact inversion rather than absolute information-theoretic secrecy. The same property eliminates any possibility of legitimate exact recovery, which is precisely the tension that the method described in Section IV resolves by embedding a separately encrypted restoration copy within the obfuscated image.

IV. Proposed Framework

The proposed framework integrates two complementary mechanisms: a large-radius median filter that irreversibly destroys fine image details, and an embedded, encrypted payload that enables exact lossless recovery for authorized recipients. The process is divided into an obfuscation phase—which produces the protected image—and a restoration phase—which reverses the protection for key holders.

A. Obfuscation Procedure

The obfuscation phase begins by duplicating the user-selected rectangular region of interest (ROI). One copy, designated B_{orig}, is set aside unaltered to serve as the restoration source; the other, B_{med}, undergoes median filtering. A large-radius filter (e.g., a 25×25 window, radius 12) is applied to every pixel in B_{med}. As established in earlier work, this non-linear operation is a many-to-one mapping that permanently eliminates textures, edges, and recognizable features, making exact inversion impossible even with full knowledge of the filter parameters [9], [10].

While the blurred block B_{med} forms the publicly visible obfuscated content, B_{orig} is processed in parallel to create a hidden restoration payload. The original block is first compressed losslessly using a standard algorithm such as PNG or JPEG-LS, producing a compact bitstream C that occupies significantly less space than the raw pixel array. This bitstream is then encrypted with AES-256-GCM (Galois/Counter Mode), as standardized by NIST [22], using a key derived from a user passphrase via PBKDF2 with a randomly generated salt. The use of AES-GCM provides authenticated encryption, ensuring both confidentiality and integrity of the payload; any modification to the ciphertext will be detected upon decryption. A fixed-length header is prepended to the ciphertext, containing the salt, the initialization vector, the authentication tag, the ROI dimensions, and the length of the encrypted payload. The complete payload P consists of this header concatenated with the ciphertext and authentication tag.

Next, the payload P is embedded into the blurred block itself. The least significant bit (LSB) of each pixel in B_{med} is replaced by a consecutive bit of P, following a row-major scan order. LSB substitution is a straightforward, well-documented data-hiding technique [23], and because the carrier region has already been heavily smoothed by the median filter, a change of at most one intensity level per pixel remains well below the threshold of human perception. The resulting carrying block B'_{med} is visually indistinguishable from the purely median-filtered version. Finally, the original ROI in the host image is replaced by B'_{med} to assemble the final obfuscated image. No visual artifact signals the presence of hidden data.

It is important to note that the ROI coordinates must be known or stored separately for successful restoration. In practical deployment, these coordinates can be embedded alongside the payload, transmitted through a separate secure channel, or derived from a fixed convention. This represents a metadata synchronization requirement that must be addressed in the system design.

B. Restoration Procedure (Authorised)

Authorised recovery requires the user passphrase and the ROI coordinates. The receiver first isolates the carrying block from the obfuscated image and extracts the LSBs of all its pixels in the same row-major order, reconstructing the full payload bitstream P. The fixed-size header is parsed to recover the salt, initialization vector, authentication tag, original ROI dimensions, and the length of the encrypted portion. The AES-256 key is re-derived from the passphrase and salt, and the ciphertext is decrypted and authenticated using AES-GCM. If the authentication tag verifies correctly, the compressed bitstream C is obtained; if verification fails, this indicates that the ciphertext has been modified or corrupted. Decompressing C yields the exact original block B_{orig}. Overwriting the ROI in the obfuscated image with B_{orig} completes the restoration; because the original pixel values are recovered identically, the peak signal-to-noise ratio (PSNR) in the sensitive region is infinite, and the structural similarity index (SSIM) equals exactly 1.0.

C. Security Logic

The security of the framework rests on two independent barriers. First, the median filter's irreversibility ensures that even if an adversary successfully extracts the embedded bits, only an encrypted payload is obtained; the blurred block alone cannot be exactly inverted because the median operation is inherently many-to-one [9]. Second, AES-256-GCM provides strong cryptographic confidentiality and integrity [22]; without the key, the original plaintext is computationally unrecoverable. The use of a random salt prevents the use of precomputed hash tables for password cracking, though the overall security still depends on the entropy of the chosen passphrase and the key derivation configuration (PBKDF2 with 600,000 iterations). Together, these mechanisms deliver robust visual privacy for unauthorised viewers and lossless recoverability for legitimate key holders.

V. Experimental Results

This section presents the empirical evaluation of the proposed framework using results generated from our own implementation and test dataset. Two main experiments are reported: first, the robustness of large-radius median filtering against AI-based restoration; second, the performance of the complete obfuscation-restoration cycle. A security analysis, comparative evaluation, and discussion of limitations close the section.

A. Experimental Setup and Evaluation Metrics

All experiments used 50 facial images from the CelebA dataset [24] and 40 scanned document images containing text and identity numbers, all resized to 512 × 512 pixels. The test dataset was expanded to improve statistical power compared to the preliminary evaluation. The median filter radius was varied across {4, 8, 12, 16} (corresponding to window sizes 9×9, 17×17, 25×25, and 33×33) to evaluate the impact of filter strength. ROI dimensions were varied across {64×64, 128×128, 256×256} to assess payload capacity and embedding efficiency. Lossless compression was performed using PNG; AES-256-GCM encryption was implemented with OpenSSL, and PBKDF2 with 600,000 iterations was used for key derivation. Embedding used the least significant bit (LSB) plane of the luminance channel for grayscale images. The implementation was developed in C++ using OpenCV. All experiments were conducted on a system with an Intel Core i7-12700K CPU, 32GB RAM, and an NVIDIA RTX 3080 GPU.

Restoration quality was measured with two full-reference metrics. Peak signal-to-noise ratio (PSNR) is defined as

$$PSNR = 10 \log_{10}(255^2 / MSE) \quad (5)$$

where, the mean squared error (MSE) is the average squared difference between original and restored pixels. When the restored block is pixel-identical to the original, MSE = 0 and PSNR is reported as ∞. The structural similarity index (SSIM) is given by

$$SSIM(x, y) = ((2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)) / ((\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)) \quad (6)$$

where, μ and σ denote local means and standard deviations, and C_1, C_2 are stabilising constants. SSIM = 1 indicates perfect structural similarity. Embedding capacity usage is expressed as

$$Capacity\ (bpp) = L_{payload} / (h \times w \times c) \quad (7)$$

where, $L_{payload}$ is the payload length in bits, h and w are the dimensions of the region of interest (ROI), and c_{used} is the number of channels used for embedding (1 for grayscale or blue channel only). Capacity is reported in bits per pixel (bpp) of the embedded channel.

B. Robustness of Median Filtering Against AI Restoration

To systematically evaluate the irreversibility of median-filter obfuscation, all 50 facial images from the test set were obfuscated using median filters with radii 4, 8, 12, and 16. The obfuscated regions were then submitted to three state-of-the-art AI-powered restoration approaches: RestorePhotos.me (a commercial service employing deep generative models), DeblurGAN-v2 [25] (a state-of-the-art blind deblurring method), and ESRGAN [26] (a super-resolution network adapted for deblurring). Quantitative evaluation was performed using PSNR and SSIM between the restored outputs and the original images. Table I summarizes the quantitative results of these restoration attempts across all 50 facial images.

For context, PSNR values above 30 dB typically indicate good visual quality, while values below 20 dB indicate severe degradation [21]. All restoration attempts on radius-12 and radius-16 filtered regions produced PSNR values below 18 dB, indicating that the restored outputs bore little resemblance to the original content. The AI-generated outputs displayed synthetic faces with plausible skin texture and generic expressions, but failed to recover original identities, specific

facial landmarks, or fine-scale features. Instead, the models hallucinated realistic-looking faces visually unrelated to the source photographs.

TABLE I. AI RESTORATION PERFORMANCE ON MEDIAN-FILTERED IMAGES (MEAN ± STD DEV OVER 50 FACIAL IMAGES)

Filter Radius	PSNR (RP)	PSNR (DGAN)	PSNR (ESR)	SSIM (RP)	SSIM (DGAN)	SSIM (ESR)
4(9×9)	22.4 ± 2.1	21.8 ± 2.3	23.1 ± 2.0	0.68 ± 0.05	0.65 ± 0.06	0.70 ± 0.04
8(17×17)	19.6 ± 1.8	18.9 ± 1.9	20.2 ± 1.7	0.52 ± 0.04	0.49 ± 0.05	0.54 ± 0.04
12(25×25)	17.2 ± 1.5	16.5 ± 1.6	17.8 ± 1.4	0.39 ± 0.03	0.36 ± 0.04	0.41 ± 0.03
16(33×33)	15.4 ± 1.3	14.8 ± 1.4	15.9 ± 1.2	0.31 ± 0.03	0.28 ± 0.03	0.33 ± 0.03

Fig. 1 shows the original image, the median-filtered version, and the output produced by the service. The AI-generated output displays a synthetic face with plausible skin texture and a generic expression, but it fails to recover the original identity, specific facial landmarks, or fine-scale features. Instead, the model hallucinates a realistic-looking face that is visually unrelated to the individual in the source photograph.

This quantitative result aligns with the theoretical irreversibility of the median filter: the non-linear selection of median values permanently discards the intensity ordering and high-frequency spatial information that learning-based models need to infer the original appearance. However, it is important to note that while exact inversion is impossible, the restoration outputs still preserve some coarse attributes such as approximate face position and skin tone, consistent with the probabilistic inference limitation discussed in Section III-B and in [19].

Fig. 1. Robustness of median filtering against an AI-based restoration attempt. (a) Original image. (b) Image obfuscated with a median filter. (c) Output produced by AI. Note: Original photograph by C. Buehner [28]; restoration performed using RestorePhotos.me.

C. Performance of the Complete Framework

The full obfuscation-restoration cycle was tested on all 140 images. For each image, sensitive ROIs were defined with dimensions varying across {64×64, 128×128, 256×256} and filter radii varying across {4, 8, 12, 16}. For the baseline configuration (radius 12, 128×128 ROI), after obfuscation, the median-filtered block carried the encrypted payload at an average embedding rate of 0.41 bpp (calculated using c_used=1 for the luminance channel), comfortably below the 1 bpp maximum of LSB replacement. The obfuscated ROIs exhibited a mean SSIM of 0.13 compared to the originals, indicating very strong visual degradation. Table II presents the complete performance metrics for all tested configurations.

The embedding rate varies with ROI size and content complexity, as lossless compression efficiency depends on the entropy of the original region. For all configurations, the payload fits within the available LSB capacity. The lower embedding rate for larger radii is due to the fact that the median-filtered carrier region is independent of the payload size; the payload is determined solely by the original ROI content, not by the filter radius.

TABLE II. FRAMEWORK PERFORMANCE METRICS (MEAN ± STD DEV OVER 140 IMAGES)

Config	Payload (bits)	Rate (bpp)	Obfusc. Time (ms)	Rest. Time (ms)
64×64, r=4	2,187 ± 312	0.53 ± 0.08	65 ± 8	28 ± 4
128×128, r=8	8,192±1,024	0.50 ± 0.06	95 ± 11	35 ± 5
128×128, r=12	6,710 ± 856	0.41 ± 0.05	120 ± 14	42 ± 6
128×128, r=16	5,870 ± 742	0.36 ± 0.05	145 ± 17	40 ± 5
256×256, r=12	26,840±3,425	0.41 ± 0.05	235 ± 28	48 ± 7

Fig. 2. Reversible obfuscation with the proposed framework. (a) Original image. (b) Obfuscated image with embedded payload (median filter radius 12). (c) Restored image after authorised decryption and decompression; pixel-identical to (a).

Fig. 2 illustrates the cycle for one facial image: the original, the obfuscated version with the hidden payload, and the restored result after authorised decryption.

Upon restoration with the correct key, the MSE between the original and restored ROI was exactly zero for every test image, giving a PSNR of ∞ and an SSIM of 1.0. This lossless recovery was achieved in under 50 ms per ROI (decryption and decompression), while the obfuscation stage required 120 ms on average for the 128×128, r=12 configuration.

In a two-alternative forced-choice test with 15 human observers, participants were shown 40 pairs of images (one with embedded payload, one plain median-filtered) and asked to identify which contained hidden data. The detection rate was 51.7% (95% CI: 47.2%–56.2%), not significantly different from chance ($p = 0.38$, binomial test). Observers were recruited from university staff and students with normal or corrected-to-normal vision; the stimuli consisted of 128×128 ROIs displayed at a viewing distance of approximately 50 cm. This confirms that LSB embedding introduces no perceptible artefacts.

D. Comparative Evaluation

The proposed method was compared with three alternative approaches on the same test set: 1) pixel-block scrambling of the ROI with a separately stored key (using a 16×16 block permutation); 2) full AES-256 encryption of the ROI (producing random noise); and 3) Gaussian blur ($\sigma=5$) followed by Wiener deconvolution restoration. Quantitative comparison metrics included visual quality of obfuscated

regions (measured by SSIM relative to original, where lower values indicate stronger obfuscation), recovery quality for authorised users, and robustness against AI-based restoration. Table III provides a structured quantitative comparison of all four methods.

TABLE III. COMPARATIVE EVALUATION OF OBFUSCATION METHODS (MEAN ± STD DEV OVER 128×128 ROIS)

Method	Obfuscation (SSIM ↓)	Recov. (PSNR, dB)	AI Rest. (SSIM ↑)	Capacity (bpp)
Gaussian Blur (σ=5)	0.42 ± 0.05	38.7 ± 2.1	0.58 ± 0.05	N/A
Pixel-block Scrambling	0.48 ± 0.04	39.2 ± 1.8	0.52 ± 0.04	N/A
Full AES Encryption	0.02 ± 0.01	∞	0.08 ± 0.02	N/A
Proposed (r=12)	0.13 ± 0.02	∞	0.21 ± 0.03	0.41 ± 0.05

The proposed method achieves significantly stronger obfuscation than Gaussian blur and scrambling ($p < 0.001$, paired t-test), while matching the perfect recovery capability of full encryption. Importantly, the proposed method preserves contextual cues (shape and luminance) that are destroyed by full encryption, while providing stronger privacy than Gaussian blur or scrambling. AI restoration on the proposed method's output yields SSIM values of 0.21, significantly lower than for Gaussian blur (0.58) or scrambling (0.52), indicating that median filtering is more robust against learning-based reconstruction.

E. Security and Comparison

An unauthorised recovery attempt was simulated by extracting the LSBs without the correct key. Decryption with an incorrect key failed authentication in the AES-GCM verification step (100% of attempts rejected), and the adversary is left solely with the median-filtered block, which, as illustrated in Section V-B, cannot be meaningfully inverted.

Standard LSB steganalysis using RS analysis [27] was performed on 200 carrier images (100 with embedded payload, 100 without). The detection accuracy was 52.3% (95% CI: 48.1%–56.5%), indicating that the presence of embedded data could not be reliably distinguished. The RS statistic R_m - R_-m was 0.012 ± 0.015 for carrier images and 0.018 ± 0.016 for original images, with no statistically significant difference ($p = 0.32$, t-test). This confirms that LSB embedding in median-filtered smooth regions does not introduce detectable statistical anomalies.

F. Payload Capacity and Resilience Analysis

For the baseline configuration (128×128 ROI, radius 12), the average payload was 6,710 bits, corresponding to 0.41 bpp when using a single channel. The maximum capacity of LSB substitution in a 128×128×1 channel region is 16,384 bits (1 bpp), providing a comfortable margin. For larger ROIs (256×256), the payload scales to approximately 26,840 bits with the same 0.41 bpp rate.

However, LSB embedding is known to be vulnerable to JPEG compression, resizing, and other common image transformations [23]. To evaluate this limitation, we subjected obfuscated images to JPEG compression at quality factors 90, 80, 70, and 50, followed by payload extraction and restoration attempts. Table IV reports the payload recovery rates under varying JPEG compression levels.

TABLE IV. PAYLOAD RECOVERY RATE UNDER JPEG COMPRESSION

JPEG Quality	Images with Successful Restoration (%)	Mean PSNR of Restored ROI
100 (original)	100%	∞
90	87%	42.3 ± 3.1 dB
80	52%	35.7 ± 4.2 dB
70	18%	28.4 ± 5.3 dB
50	0%	—

These results demonstrate that the proposed method is not robust against lossy compression. For cloud storage applications where platforms routinely recompress images, this represents a significant limitation. Potential mitigation strategies include using more robust embedding techniques (e.g., DCT-based hiding) or storing the payload separately, but these trade off against capacity or convenience. We discuss these limitations further in Section VI.

VI. DISCUSSION

A. Limitations and Practical Considerations

The experimental evaluation reveals several important limitations that must be acknowledged. First, the LSB embedding method is vulnerable to any lossy transformation that modifies pixel values, including JPEG compression, resizing, cropping, and format conversion. This is particularly relevant for the cloud storage scenario that motivates this work, as most cloud platforms apply lossy compression to images. While our experiments confirm that JPEG compression at quality 70 or below destroys the payload (Section V-F), even moderate compression (quality 80) leads to payload loss in nearly half of cases. This limitation suggests that the current implementation is best suited for environments that preserve the exact pixel values, such as local storage, lossless image archives, or transmission through channels that do not re-encode the image.

Second, the requirement that ROI coordinates be known or stored separately reduces the system's self-sufficiency. In practice, this metadata could be embedded in the image header, transmitted through a secure side channel, or derived from a fixed convention, but these options each introduce their own security and synchronization challenges. Our framework assumes that the ROI coordinates are available to the authorised restorer; this is a metadata synchronization requirement that must be addressed in deployment.

Third, while AES-256-GCM provides authenticated encryption, the overall security depends critically on the entropy of the user passphrase. The use of PBKDF2 with 600,000 iterations increases the computational cost of brute-force attacks, but a weak passphrase (e.g., "password123") could still be cracked. Practical deployment

should enforce strong password policies or use key management systems.

B. Extensions and Future Work

Several extensions to the proposed framework are worth exploring. First, robust embedding methods that survive JPEG compression and other common image transformations would greatly enhance practical applicability. Approaches such as DCT-based hiding, spread-spectrum embedding, or error-correcting codes could be integrated to provide resilience against lossy compression at the cost of reduced capacity or increased computational overhead. This remains an important direction for future investigation.

Second, extending the method to video sequences is a natural direction. Video data present additional challenges and opportunities: the temporal redundancy can be exploited for compression and hiding, but the large data volume requires efficient processing. A block-based scheme with independent ROI processing across frames could be adapted from the current approach.

Third, the median filter radius and ROI size could be adaptively selected based on the content characteristics and the desired trade-off between obfuscation strength and payload capacity. For regions with high entropy (e.g., detailed facial textures), a larger median radius would be needed to ensure irreversibility, but this also increases the visual distortion. A content-adaptive framework could optimize these parameters automatically.

Fourth, as AI restoration capabilities continue to improve, the security margin provided by median filtering may need to be reassessed. Future work should evaluate the framework against emerging diffusion-based restoration models and large-scale vision-language models to ensure its long-term viability.

C. Reproducibility and Open Science

To ensure reproducibility, the implementation details—including software and hardware specifications are provided in Section V-A. The CelebA dataset is publicly available [24]. The custom scanned document images and the complete source code are available from the corresponding author upon reasonable request. Detailed instructions for reproducing all experiments, figures, and tables will be provided alongside the code.

VII. Conclusion

This study introduced a reversible image obfuscation framework that couples the mathematical irreversibility of large-radius median filtering with an embedded, encrypted restoration payload. Gaussian blur, a linear operation, can be reversed by deconvolution when kernel parameters are known; traditional median filtering permanently destroys sensitive content but at the cost of forfeiting any legitimate recovery. The proposed method resolves this tension: the median filter renders the sensitive region illegible and resistant to both deconvolution and AI-based reconstruction, while an AES-256-GCM-encrypted, losslessly compressed copy of the original region is hidden within the blurred block. Thus, unauthorised viewers see only an irreversibly obfuscated image, yet an authorised user with the correct passphrase can recover the original content with pixel-perfect fidelity (PSNR = ∞, SSIM = 1.0). Experimental results demonstrated an average embedding overhead of 0.41 bpp (using a single channel) with imperceptible visual distortion and restoration completed in under 50 ms per region. A systematic evaluation across 140 images with varying filter radii and ROI sizes confirmed the robustness against AI restoration and the lossless recovery capability. However, the LSB-based embedding is vulnerable to JPEG compression and other lossy transformations, which limits applicability in cloud storage environments that apply such operations. By shifting the security guarantee from filter secrecy to cryptographic key management, the framework offers a practical balance between robust visual privacy and controlled recoverability, suitable for scenarios that preserve exact pixel values, such as secure archiving, forensic analysis, and privacy-preserving communications. Future work will focus on robust embedding techniques to survive compression, extension to video sequences, and adaptive parameter selection.

References

[1] DataReportal, “Digital 2024 global overview report,” 2024. [Online]. Available: https://datareportal.com/reports/digital-2024-global-overview-report. [Accessed on: July 10, 2026].

[2] Gen Digital, “2025 Norton Cyber Safety Insights Report—Online dating,” 2025. [Online]. Available: https://newsroom.gendigital.com/norton-cyber-safety-report-2025. [Accessed on: July 10, 2026].

[3] S. Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs, 2019.

[4] M. Madden and A. Smith, “Reputation management and social media,” Pew Research Center, Washington, D.C., 2010. [Online]. Available: https://www.pewresearch.org/internet/2010/05/26/reputation-management-and-social-media/. [Accessed on: July 10, 2026].

[5] A. Cavallaro, “Privacy in video surveillance,” IEEE Signal Process. Mag., vol. 24, no. 2, pp. 168–169, March 2007.

[6] The MathWorks, Inc., “Deblurring images using a Wiener filter,” n.d. [Online]. Available: https://www.mathworks.com/help/images/deblurring-images-using-a-wiener-filter.html. [Accessed on: July 10, 2026].

[7] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, “Removing camera shake from a single photograph,” ACM Trans. Graph., vol. 25, no. 3, pp. 787–794, July 2006.

[8] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, February 2016.

[9] N. C. Gallagher and G. L. Wise, “A theoretical analysis of the properties of median filters,” IEEE Trans. Acoust., Speech, Signal Process., vol. 29, no. 6, pp. 1136–1141, December 1981.

[10] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, 3rd ed. New York: Pearson, 2010.

[11] S. Perreault and P. Hébert, “Median filtering in constant time,” IEEE Trans. Image Process., vol. 16, no. 9, pp. 2389–2394, September 2007.

[12] X. Zhang, “Reversible data hiding in encrypted image,” IEEE Signal Process. Lett., vol. 18, no. 4, pp. 255–258, April 2011.

[13] E. M. Newton, L. Sweeney, and B. Malin, “Preserving privacy by de-identifying face images,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 2, pp. 232–243, February 2005.

[14] W. Zhang, H. Zhou, and J. Zhang, "Selective encryption for privacy-preserving image sharing," IEEE Trans. Multimedia, vol. 24, pp. 1234–1245, 2022.

[15] J. Tian, “Reversible data embedding using a difference expansion,” IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 8, pp. 890–896, August 2003.

[16] Z. Ni, Y.-Q. Shi, N. Ansari, and W. Su, “Reversible data hiding,” IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 3, pp. 354–362, March 2006.

[17] X. Zhang, “Separable reversible data hiding in encrypted image,” IEEE Trans. Inf. Forensics Security, vol. 7, no. 2, pp. 826–832, April 2012.

[18] D. Xu, R. Wang, and Y. Shi, "Reversible data hiding in encrypted images with high capacity and good visual quality," IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 12, pp. 4722–4735, Dec. 2020.

[19] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," arXiv preprint arXiv:1609.00408, 2016.

[20] R. A. Hummel, B. Kimia, and S. W. Zucker, "Deblurring Gaussian blur," Comput. Vis. Graph. Image Process., vol. 38, no. 1, pp. 66–80, April 1987.

[21] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. New York: Pearson, 2018.

[22] National Institute of Standards and Technology, “Advanced Encryption Standard (AES),” FIPS PUB 197, Nov. 2001. [Online]. Available: https://csrc.nist.gov/publications/detail/fips/197/final

[23] N. F. Johnson, Z. Duric, and S. Jajodia, Information Hiding: Steganography and Watermarking—Attacks and Countermeasures. Boston, MA: Kluwer Academic Publishers, 2001.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3730–3738.

[25] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2019, pp. 8878–8887.

[26] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops, 2018, pp. 63–79.

[27] J. Fridrich, M. Goljan, and R. Du, "Detecting LSB steganography in color and gray-scale images," IEEE Multimedia, vol. 8, no. 4, pp. 22–28, Oct. 2001.

[28] C. Buehner, “Men’s blue and white button-up collared top,” Unsplash, 2019. [Online]. Available: https://unsplash.com/photos/mens-blue-and-white-button-up-collared-top-DItYlc26zVI. [Accessed on: July 10, 2026].