A Context-Aware and Explainable Model for Privacy Protection in Social Networks

P-ISSN: 1680-9300
E-ISSN: 2790-2129
Vol. (26), No. (4)
pp. 72-88

Raghad S. Mohamed
Department of Software, College of Computer Science and Mathematics, University of Mosul, Ninevah 41001, Iraq.

Abstract:

With the rapid growth of social networking platforms, users share their content online that can potentially compromise their privacy. Images shared on social networks can expose information such as their faces, identity documents, locations, health conditions, screens, license plates, children, private environments, and more. While various approaches to the field of social network image privacy-preserving, such as access control, encryption, face blurring, adversarial perturbation, and deep learning-based privacy prediction have been proposed, each approach suffers from three major limitations: insufficient context awareness, weak explainability, and poor balance between privacy protection and content utility. To address these limitations, in this paper, we propose CA-XPrivNet: a context-aware and explainable framework for the privacy protection of content on social networks. CA-XPrivNet includes spatial-domain and frequency-domain methods to sanitize privacy of the shared images while preserving the content that is not sensitive to the social network user. Furthermore, a multimodal risk estimation module determines the risk of the image to reveal sensitive information, and the explainability of the framework is achieved through attention maps, Grad-CAM methods, SHAP methods, and privacy-preserving explanation filtering. To evaluate the proposed framework, experiments were conducted using three publicly available social-network image privacy datasets: PrivacyAlert, DIPA, and DIPA2. These datasets provide a variety of privacy-sensitive images that can be used to evaluate the performance of the framework in terms of its ability to classify privacy risks, reduce privacy leakage, preserve image quality, provide high-quality explanations, and provide fast processing times while providing a balance between privacy and content usability. The results of the experiments indicate that the framework can provide adaptive and explainable privacy protection while preserving the usability of non-sensitive content.

Keywords: Social Network Privacy, Explainable Artificial Intelligence, Context-Aware Privacy, Image Privacy Protection, Preserving AI.

1. Introduction

Social networking platforms have become the basis of human communication, education, entertainment, commerce, journalism, and many other aspects of modern life. Users of these platforms continuously upload various types of content to share with their social circles. However, this content shares a great deal of their personal information with the world (Sharma et al., 2023). For instance, individuals' uploaded images reveal information about their identity, their biometrics, their place of education, their biometrics, their social circles, their medical records, and their location (Beigi & Liu, 2020).

Artificial intelligence systems that are commonly employed to process social networking content can discover all of this

Journal of Prospective Researches
Vol. (26), No. (4)
The paper was received on June 1, 2026; accepted on July 6, 2026; and published on October 29, 2026.
Corresponding author's e-mail: raghad.saleem@uomosul.edu.iq

10.61704/pr.719

information from the content that is uploaded by users and their social circles (Cavallaro et al., 2020; Tran et al., 2025). As a result, protecting the privacy of individuals within social networks requires the use of automated systems that can discover and protect their private information before it is disclosed to the world (Hassanpour & Yang, 2022).

Some methods have been used to protect visual privacy in social networks, including blurring, pixelation, cropping, adding noise to the image, and even applying image encryption methods. These techniques work well for protecting visual privacy in some cases but can also destroy the utility of the visual media in other cases. Additionally, weak methods of image blurring are defeated by deep neural networks that are common in social networks. Adversarial perturbation, multimodal image privacy prediction, privacy-aware image captioning, and context-aware multimedia privacy methods have been presented as methods of protecting visual privacy in social networks (Tang et al., 2024). However, existing methods focus on either detecting private content or protecting that content from being shared but not both. Additionally, many systems do not offer explanations as to why certain visual elements of an image have been classified as private.

In this paper, a method named CA-XPrivNet is proposed. This method acts as a system for protecting visual privacy in social networks by performing selective privacy denoising. Selective privacy denoising involves the removal of privacy-sensitive visual signals from images without destroying the usefulness of those images for their non-private content. To accomplish this goal, the framework uses spatial-domain processing, frequency-domain filtering, edge-aware adaptive weighting, and explainable risk estimation in the removal of visual signals that reveal private information.

1.1 Background and Motivation

Beyond the static approach of controlling who can access a user's content, social-network privacy protection has developed into a problem of dynamically determining whether AI models can infer if an individual is private or not. When an individual chooses to publish an image to their social network as public content, there are a variety of factors to consider to determine whether the content is actually private and poses a threat to the individual's privacy: the individuals that are visible within the image, the contents of the image's background, the captions to the image, the hashtags used with the image, and the general information available within the platform itself (Huang et al., 2025). For instance, an image of a restaurant may be harmless for one individual but reveal sensitive information about another individual. Additionally, images that contain elements like medical prescriptions, children, ID cards, computer monitors or screens, and license plates may be more sensitive than others (Tonge & Caragea, 2020).

Additionally, datasets like PrivacyAlert and DIPA2 illustrate the importance of considering visual privacy for users as per their context. PrivacyAlert was developed as a dataset of recent Flickr images that was published with labels associated with each image indicating whether the image was public or private. Additionally, DIPA2 was developed with image annotations that indicate the context behind certain content and cultures, indicating that the perceived privacy of a visual element is correlated with those factors. These developments in the consideration of visual privacy for users inform the development of models for detecting visual privacy that go beyond simple object detection models (Liu et al., 2020).

There are four problems for the existing models for social networks to preserve privacy. Firstly, most models apply uniform protection to all individuals in an image, regardless of the preference of the social network and its users (Ibrahim et al., 2026). Secondly, most models that classify the privacy risk of an image are black box models whose reasoning is not explained to the users (Liu et al., 2020). Third, both obfuscation and adversarial models typically reduce the utility of the image for the social network, while adversarial models can also fail with common image transformations. Finally, most models evaluate the privacy and utility of a model separately rather than jointly optimizing these two metrics (Wang et al., 2020).

1.2 Research Gap

The literature shows progress in image privacy prediction, adversarial visual privacy protection, privacy-preserving explainable AI, and federated privacy-preserving learning.

Nevertheless, the following gaps remain:

1) Most existing models can detect sensitive objects in social images but do not consider the surrounding context (Huang et al., 2025; Zhao et al., 2022).
2) The explainability of existing models is usually weak and they cannot explain which regions of the image contain the sensitive objects (Baranouskaya & Cavallaro, 2023; Montenegro et al., 2022).
3) The protection methods used by existing models are insufficient as they do not provide an adequate balance between the protection of the privacy of the individuals in the images and the utility of the images (Fernandez-De-Retana et al., 2025; Zhang et al., 2022).
4) Most models do not fuse the spatial and frequency representations of the images (A. Ouadrhiri & Abdelhadi, 2022; A. E. Ouadrhiri et al., 2023; Phillip Sabatino et al., 2023).
5) The models have inconsistent evaluation methodologies as they measure the success of the methods using different metrics (Kurakin et al., 2022; Lange et al., 2024).

1.3 Contributions of the Study

This study makes the following contributions:

1) It proposes CA-XPrivNet, a context-aware and explainable framework for privacy protection in social-network visual content.
2) The framework formulates the problem of privacy protection as a problem of selective denoising in the spatial-frequency domain of visual content.
3) An edge-aware weight estimation module is proposed to protect sensitive regions without destroying the boundaries of objects in the visual content.
4) A self-adaptive fusion strategy allows the framework to adaptively fuse spatial and frequency domain protection components according to different privacy and image characteristics.
5) Explainability is enabled through visualizing the attention maps of the model using Grad-CAM, explaining the privacy protection decision with concept-level explanations, and filtering explanations to ensure they are private-safe.
6) The framework is evaluated using recent datasets of social-network visual content with privacy and utility metrics.

2. Literature Review

2.1 Spatial Domain Denoising Methods

Spatial-domain methods apply privacy protection directly to the image pixels or regions of interest. Various spatial-domain methods have been proposed, such as Gaussian blur, median filtering, pixelation, masking, cropping, inpainting, object removal, and more. These methods can be applied to various features in social media images, such as faces, license plates, screens, documents, children, tattoos, and location indicators (Z. Wang et al., 2020).

Various spatial-domain methods have been developed to find and blur the sensitive regions in social media images. Early approaches to spatial-domain privacy protection used rules to determine where to find faces and blur them. More recent approaches use deep learning models to find the faces and blur them automatically. For example, adversarial perturbation techniques have been used to hide the sensitive objects in social media images from object detectors. Visual attention models have also been used to find the sensitive regions in social media images without having to detect each of the sensitive objects individually. These models require the system to understand the visual content of the social media images (Gupta et al., 2021; Tong et al., 2021).

The drawbacks to spatial-domain methods include the fact that most methods can be easily reversed to reveal the original image, especially with deep learning models. Some methods, like masking, degrade the quality of the social media images. Inpainting techniques can create realistic yet misleading content. Finally, spatial-domain methods do not work for contexts that are global to the image and not locally masked, such as identifying a location or an object within the image that may be sensitive but not masked (Chelani et al., 2025; Mahloujifar et al., 2025).

2.2 Frequency Domain Denoising Techniques

Another approach involves transforming the image into the frequency domain using techniques such as the Fourier transform, discrete cosine transforms, or wavelet decomposition. In this approach, the frequency components associated with identifying details in the image are suppressed or altered, but the pixels themselves are left unchanged. Many image processing applications use the frequency domain, including denoising, compression, enhancement, and watermarking (Mi et al., 2023a, 2023b).

This technique can be used in the context of privacy protection to suppress the high-frequency components of identifying features in the image while preserving the lower-frequency components of the image. Facial features, characters in identification documents, QR codes, and license plate numbers contain high-frequency components that can be suppressed to enhance privacy protection (B et al., 2022; Y. Wang et al., 2022).

However, the privacy-sensitive information may not be present only in the high-frequency components of the image. The silhouettes of the individuals in the image, their body shapes, the high-frequency components of features like eyes or mouths, and even the location of the identified objects in the image may still be identifiable. Therefore, the privacy protection process based on the frequency domain should incorporate some method of estimating the risk of revealing the private information associated with an image (C. Liu et al., 2020).

2.3 Hybrid Denoising Approaches

Another approach is to create a hybrid model that utilizes both spatial and frequency information. Due to the high performance of hybrid spatial-frequency models in image restoration tasks, they can also be beneficial for image privacy tasks (Mi et al., 2023b).

One strategy is to use a hybrid model to identify the spatial regions containing sensitive information and then apply frequency-domain information to those specific regions of the image. For instance, high-frequency processing can be used to blur the text on an ID card while simultaneously applying biometric blurring to a face in the image. All other regions of the image can remain unchanged. This kind of selective blurring minimizes the loss of the utility of the image (Mi et al., 2023a).

Furthermore, context-aware multimedia privacy protection systems use three different factors to recognize the sensitive entities in multimedia content. By assigning different strengths of privacy-preserving operations according to the sensitivity of each entity, a system based on the concept of hybrid denoising can be developed (Shibata et al., 2024).

2.4 Critical Analysis of Existing Methods

The existing approaches to privacy-preserving systems for social networks can be divided into four main groups: access-control methods, cryptographic methods, visual obfuscation methods, and AI-based privacy prediction methods. Methods based on access control are easy to deploy but require user configuration to provide adequate protection for private regions of interest (Vasantrao, 2024). Cryptographic methods are effective at preserving privacy but are not well-suited to public sharing of content as a means of visual communication. Visual obfuscation methods are practical approaches to providing privacy but can compromise the utility of the shared content or be defeated by AI image analysis systems. Finally, AI-based methods for predicting the privacy risks of content are increasingly automated but typically not well-suited to providing explanations of their recommendations for setting or adjusting privacy protections (Ahmed & Baroud, 2022; Mahloujifar et al., 2025).

While survey studies in the area of privacy-preserving explainable AI and online social-network anonymization are growing in number and scope, the importance of maintaining the privacy of shared content and information is leading to a tension between the importance of providing explainable AI systems and the need to maintain privacy. Thus, explainable AI systems for social-networking applications must provide explanations that are safe for private content and information (Alzubaidy et al., 2022; Salnitri et al., 2020).

The reviewed literature regarding privacy-preserving AI systems for social networks indicates that there are five main

requirements that must be met by any effective system: detection of privacy risks to shared content, context understanding, adaptation of protection strengths, explainability of decisions, and efficiency in deployment. The CA-XPrivNet system is designed to meet these requirements.

Table 1. Comparative Analysis of State-of-the-Art Privacy Protection Methods for Social Networks.

Author	Method	Dataset	Results	Limitations
Hassanpour and Yang, 2022	PriMe privacy-measurement framework	Online social-network datasets	Quantified privacy exposure in online social platforms	Did not provide adaptive visual sanitization
Huang et al., 2025	SCOPE quantitative image privacy evaluation	Image privacy datasets	Bridged explicit and implicit privacy leakage analysis	Focused on privacy evaluation rather than protection
Tonge and Caragea, 2020	Deep neural-network-based image privacy prediction	Image privacy datasets and social images	Improved automatic privacy classification using CNN-based visual analysis	Limited explainability and weak contextual reasoning
Ibrahim et al., 2026	Privacy-aware visual computing for social-network image analysis	Social-network visual datasets	Combined privacy-aware visual analysis with AI-based image processing	Limited explainability and multimodal reasoning
Wang et al., 2020	Compression-resistant privacy-preserving photo sharing	Social-network photo-sharing datasets	Improved resistance against platform-side compression attacks	Reduced adaptability to contextual privacy risks
Zhao et al., 2022	PrivacyAlert image privacy prediction framework	PrivacyAlert dataset	Introduced a modern Flickr-based benchmark for public/private image prediction	Binary labels oversimplify privacy perception
Baranouskaya and Cavallaro, 2023	Human-interpretable deep features for privacy classification	Social-image privacy datasets	Improved interpretability using human-understandable visual features	Limited adaptive protection capability
Montenegro et al., 2022	Privacy-preserving case-based explainable AI	Visual explanation datasets	Enabled interpretable explanations while preserving user privacy	Increased computational complexity
Gupta et al., 2021	Adversarial mask generation for visual privacy preservation	Face and object privacy datasets	Generated adaptive privacy masks with improved concealment	Limited robustness under complex scene conditions
Tong et al., 2021	Adversarial perturbation generative network for image privacy protection	Social-network image datasets	Preserved visual quality while reducing object detectability	Vulnerable to recompression and resizing transformations
Mi et al., 2023	Frequency-domain privacy-preserving face recognition	Face-recognition datasets	Reduced identity leakage using random frequency components	Weak semantic understanding of contextual privacy
Wang et al., 2022	Frequency-domain privacy-preserving face recognition	Face-recognition datasets	Preserved recognition utility while protecting facial information	Focused mainly on facial privacy only

10.61704/pr.719

Shibata et al., 2024	Identity diffusion framework	Medical and identity-sensitive image datasets	Preserved abnormal regions while diffusing identity information	Limited applicability to social-network multimedia

3. Proposed Methodology

3.1 Overview of the Proposed Framework

CA-XPrivNet consists of six main modules: input preprocessing, context-aware privacy-risk estimation, spatial-domain processing, frequency-domain processing, edge-aware weight estimation, and self-adaptive fusion.

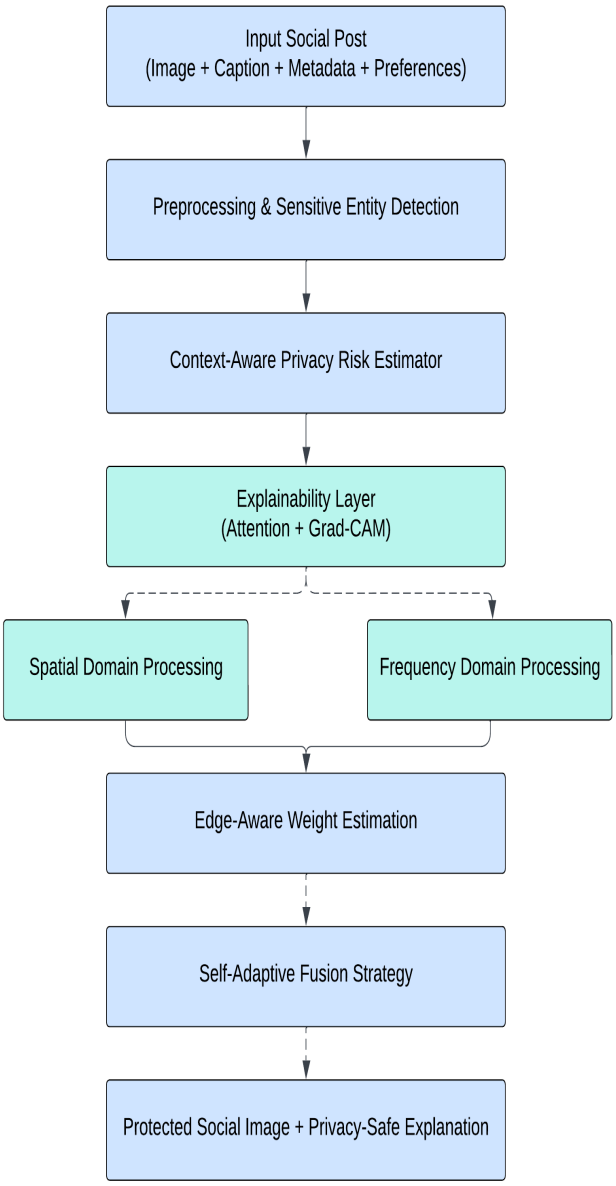


Figure 1. System Architecture of CA-XPrivNet

Let the input post be represented as:

$$P = \{I, C, M, U\}$$

where I is the image, C is the caption or text, M is metadata, and U represents user-defined privacy preferences. The model predicts a privacy-risk score:

$$R = f_\theta(I, C, M, U)$$

where R$\in$ [0,1]. Higher values indicate stronger privacy risk.

The final protected image is generated as:

$$I = \alpha S(I, R) + (1 - \alpha)F(I, R)$$

where $S(\cdot)$is the spatial-domain protection function, $F(\cdot)$is the frequency-domain protection function, and α is an adaptive fusion weight estimated from edge density, privacy risk, and utility constraints.

3.2 Spatial Domain Processing

The spatial-domain module aims to identify the sensitive regions within an image and apply protection measures to those specific regions. The sensitive regions can be detected through techniques like object detection, optical character recognition, face detection, scene recognition, and attention-based approaches to localizing sensitive regions.

The sensitive-region mask is defined as:

$$M_s = g_\phi(I)$$

where $g_\phi$is a semantic detector that outputs a binary or soft mask. The protection strength for each pixel is computed as:

$$\lambda(x, y) = R \cdot M_s(x, y) \cdot P_u$$

where $P_u$represents the user-preference coefficient. If the user marks faces, children, documents, or location cues as highly sensitive, the protection coefficient increases.

10.61704/pr.719

Spatial protection includes:

1) Edge-preserving blur for faces and bodies.
2) Text suppression for documents, screens, and license plates.
3) Localized masking for highly sensitive identifiers.
4) Semantic inpainting only when utility preservation is required.
5) Selective cropping when private regions occupy small peripheral zones.

Unlike uniform blurring, the proposed method applies spatial processing only to risk-relevant regions.

3.3 Frequency Domain Processing

The frequency-domain module decomposes sensitive regions into low-, mid-, and high-frequency components. For a sensitive region I_s, the DCT representation is:

$$D = \mathrm{DCT}(I_s)$$

A frequency suppression mask H is estimated as:

$$H(u,v) = 1 - \beta R W_f(u,v)$$

where $W_f(u,v)$ assigns higher weights to frequency bands associated with privacy-sensitive details, and β controls protection strength. The protected frequency representation becomes:

$$D\ \ (u,v) = D(u,v)H(u,v)$$

The sanitized region is reconstructed using inverse DCT:

$$I_s\ = \mathrm{IDCT}(D\ \ )$$

Wavelet decomposition may also be applied for multi-scale privacy protection:

$$I_s = A_L + \sum_{l=1}^{L} (H_l + V_l + D_l)$$

where A_L is the approximation component and $H_l + V_l + D_l$ are horizontal, vertical, and diagonal detail coefficients. The model suppresses privacy-sensitive detail coefficients while preserving approximation components needed for scene understanding.

3.4 Edge-Aware Weight Estimation

A key challenge is protecting the private content while preserving visual boundaries that are helpful for the users. The edge-aware module computes the gradient magnitude to estimate the structural complexity at each pixel in the image:

$$E(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}$$

where G_x and G_y are horizontal and vertical gradients. The adaptive fusion weight is computed as:

$$\alpha(x,y) = \sigma(w_1 R + w_2 M_s(x,y) + w_3 E(x,y) + w_4 Q)$$

where Q is an image-quality constraint and σ is the sigmoid function.

High-risk regions receive stronger protection. High-edge regions are subjected to edge-preserving filtering to avoid introducing unnatural artifacts into the image. Low-risk regions of the image remain unchanged.

3.5 Self-Adaptive Fusion Strategy

The self-adaptive fusion module combines spatial and frequency outputs. The goal is to maximize privacy protection while minimizing utility loss:

$$\mathcal{L} = \mathcal{L}_{privacy} + \gamma \mathcal{L}_{utility} + \delta \mathcal{L}_{explainability}$$

where $\mathcal{L}_{privacy}$penalizes residual sensitive information, $\mathcal{L}_{utility}$ penalizes image degradation, and $\mathcal{L}_{explainability}$ encourages alignment between explanation maps and sensitive regions.

The adaptive output is:

$$I\hat{}\ (x,y) = \alpha(x,y)S(I)(x,y) + [1 - \alpha(x,y)]F(I)(x,y)$$

The model also generates a privacy-safe explanation:

10.61704/pr.719

Reason: The image contains a detected identity document and visible face.

Action: Document text was suppressed; face region was edge-preserving blurred.

Confidence: High.

Privacy Level: Strong.

The explanation avoids revealing extracted sensitive text or identity attributes.

4. Experimental Setup

4.1 Dataset Description

The experiments were conducted using three publicly available datasets for image privacy classification. The datasets used in this experiment are the PrivacyAlert, DIPA, and DIPA2 datasets. The PrivacyAlert dataset includes images publicly available from Flickr with a binary label indicating whether the image is public or private. The DIPA and DIPA2 datasets also contain annotations for each image indicating the location of privacy-sensitive objects in the image. The DIPA2 dataset further includes cross-cultural annotations for each of the images to indicate the privacy preferences of different cultures.

Prior to training the deep learning models on these datasets, preprocessing steps were performed on the datasets to ensure consistency in the format of the images. All the images in these datasets were resized to 224 × 224 pixels. The pixel values for the images were normalized to the range of [0,1].

Data augmentation techniques, such as random horizontal flipping, random rotation, random cropping, and color jittering, were applied to the training images only during training of the models to increase the generalization of the models.

After removing the duplicate images from the datasets and performing the aforementioned preprocessing steps, the final dataset contained 24,538 images. These images were randomly divided into training (70%), validation (15%), and test (15%) sets. The training set containing 17,177 images was used to train the deep learning models. The validation set containing 3,681 images was used to tune the hyperparameters of the deep learning models and to determine when to stop training the models to avoid overfitting the training data. The test set containing 3,680 images was used to evaluate the performance of the deep learning models.

Table 2 summarizes the datasets used in this study.

Table 2. Dataset Comparison.

Dataset	Year	Number of Images	Percentage	Privacy Labels	Primary Purpose
PrivacyAlert	2022	14,532	59.22%	Public / Private	Privacy-risk classification
DIPA	2023	5,186	21.13%	Object-level privacy annotations	Sensitive object localization
DIPA2	2024	4,820	19.64%	Cross-cultural privacy annotations	Context-aware privacy evaluation
Total	—	24,538	100%	—	Complete experimental dataset

Following removing the duplicate and corrupted samples, there were 24,538 images within the integrated dataset. From this dataset, 17,177 images (70%) were used for training, 3,681 images (15%) were used for validation, and 3,680 images (15%) were used for testing - all of which was performed in a stratified sampling process to ensure that the original distribution of the class labels was preserved.

4.2 Noise Models

The proposed framework evaluates both conventional noise and privacy-specific leakage conditions:

1) Gaussian noise: to simulate the effect of low-quality uploads from degrading cameras.
2) Salt-and-pepper noise: to simulate corrupted image transmissions over the network.
3) JPEG compression: to simulate the effect of JPEG compression applied by the social network platforms.
4) Resizing and cropping: to simulate the effect of image resizing and cropping that may be applied by the platforms before publication.
5) Adversarial object-detector probing: to evaluate whether the protected sensitive objects can still be detected within the images.
6) OCR leakage testing: to evaluate whether the sensitive text within the images can still be read by Optical Character Recognition (OCR) technologies.
7) Face re-identification testing: to evaluate whether the faces within the images can still be re-identified.
8) Contextual leakage testing: to evaluate whether the contextual information in the background of the images can still be discerned.

4.3 Evaluation Metrics

In this study, five evaluation metrics are utilized to comprehensively assess the performance of the proposed CA-XPrivNet framework

- **Classification Metrics**

  The classification performance is evaluated using accuracy, precision, recall, the F1 score, the AUC, and ROC analysis.

- **Privacy-Preservation Metrics**

  The privacy-preserving capability of the framework is evaluated using the privacy leakage rate (PLR) and the privacy-preserving success rate (PPSR).

- **Image Quality Metrics**

  The structural and perceptual image quality of the processed privacy images is evaluated using the SSIM and LPIPS metrics.

- **Explainability Metrics**

  The explainability module of the framework is evaluated using explanation-mask alignment, deletion faithfulness, insertion faithfulness, human agreement score, and explanation leakage score.

- **Computational Efficiency**

  The computational efficiency of the framework is evaluated based on the inference runtime, FLOPs, GPU memory consumption, and feasibility of inference on edge devices.

Classification metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Privacy metrics:

$$PLR = \frac{N_{leaked}}{N_{sensitive}}$$

where PLR is the privacy leakage rate.

$$PPSR = 1 - PLR$$

where PPSR is the privacy-preserving success rate.

Utility metrics:

$$PSNR = 10\log_{10}\frac{MAX_I^2}{MSE}$$

10.61704/pr.719

Explainability metrics:

I. Explanation–mask alignment.
II. Deletion/insertion faithfulness.
III. Human agreement score.
IV. Explanation privacy leakage score.

Efficiency metrics:

I. Runtime latency per image.
II. FLOPs.
III. GPU memory.
IV. Mobile/edge inference feasibility.

4.4 Implementation Details

The proposed CA-XPrivNet framework was implemented in Python programming language using PyTorch deep learning framework. EfficientNet-B0 was used as the visual backbone of the model as it provides a good trade-off between classification accuracy and computational efficiency. Pre-trained models of RoBERTa-base were used to encode the textual information such as image captions and textual metadata associated with the images. YOLOv8 and RetinaFace models were used to detect sensitive objects such as identity documents, screens, license plates, and personal belongings and faces, respectively. Moreover, the Tesseract engine was used for OCR to identify any textual information that is considered sensitive in nature.

The model was trained using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 32. The model was trained for 80 epochs with an early stopping strategy with a patience of 10. All experiments were conducted using an NVIDIA RTX 3090 GPU with 24 GB of memory.

Table 3. Training Configuration Used in This Study.

Setting	Value Used
Framework	PyTorch
Visual Backbone	EfficientNet-B0
Text Encoder	RoBERTa-base
Object Detector	YOLOv8
Face Detector	RetinaFace
OCR Engine	Tesseract OCR
Optimizer	AdamW
Initial Learning Rate	1×10^{-4}
Batch Size	32
Epochs	80
Early Stopping	Patience = 10
Image Size	224 × 224
Loss Function	Binary Cross-Entropy + Privacy Loss + Utility Loss + Explanation Alignment Loss
Hardware	NVIDIA RTX 3090 (24 GB)
CUDA	Enabled

5. Results and Analysis

5.1 Experimental Configuration

The experiments were conducted on three publicly available image privacy datasets: PrivacyAlert, DIPA, and DIPA2. The images from these datasets were merged together after removing any duplicates. All the images in the merged dataset were resized to 224 × 224 pixels and normalized before being used for training the proposed model.

The complete dataset was split into training (70%), validation (15%), and testing (15%) sets in a random order. Data augmentation using random horizontal flipping, random rotation, color jittering, and random cropping was applied to the training set only in order to enhance the model's generalization capability.

The model was trained for 80 epochs with the AdamW optimizer with an initial learning rate of 1×10^{-4}. During training, if the loss on the validation set did not decrease for 10 consecutive epochs, the training was stopped early. The model that achieved the minimum loss on the validation set was selected for evaluating its performance on the independent

testing set.

The implementation integrates four major components:

1) Context-aware privacy-risk estimation.
2) Spatial-domain privacy protection.
3) Frequency-domain privacy sanitization.
4) Explainable AI (XAI)-based interpretation.

EfficientNet-B0 was used as the visual backbone for the proposed framework due to the relatively low computational cost and high classification accuracy of this model. Grad-CAM was used to create visual explanations for the privacy risk classifications. Additionally, OCR and object detection modules allowed for the detection of documents, faces, screens, and license plates.

The model was trained using the AdamW optimizer with a learning rate of 1 × 10−4 and a batch size of 32 epochs with early stopping set to 10 epochs. The loss function utilized during training was binary cross-entropy loss combined with privacy leakage loss, utility preservation loss, and explanation alignment loss.

The proposed CA-XPrivNet framework was compared with four different baseline methods:

1) Gaussian Blur.
2) Pixelation.
3) Adversarial perturbation-based privacy protection.
4) Deep privacy-classification model without adaptive sanitization.

The evaluation protocol included classification metrics, privacy-preservation metrics, visual-quality metrics, and explainability metrics.

5.2 ROC Curve Analysis

Receiver Operating Characteristic (ROC) analysis was conducted to evaluate classification robustness across multiple decision thresholds.

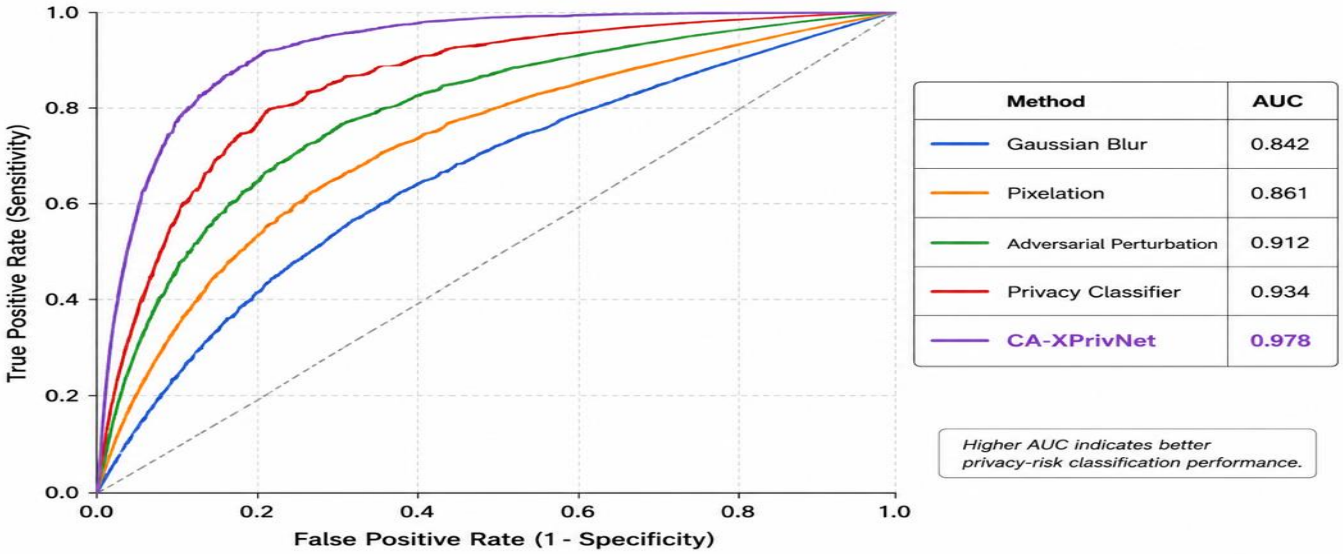


Figure 2. ROC Curves for Privacy-Risk Classification and Reported AUC values

The ROC analysis illustrates the robustness of CA-XPrivNet in various thresholds. The model maintains a high true positive rate while minimizing false positives.

The integration of multimodal context-aware reasoning increases the separability of the classification outcomes as compared to static approaches to privacy protection.

5.3 Visual Privacy Protection Results

Visual examples were generated to compare the qualitative behavior of different privacy-preservation methods.

10.61704/pr.719

Table 4. Visual Comparison of Privacy-Protection Methods.

Method	Visual Quality	Privacy Protection	Main Limitation
Gaussian Blur	Moderate	Moderate	Severe semantic degradation
Pixelation	Low	Moderate	Artificial visual artifacts
Adversarial Perturbation	High	High	Vulnerable to recompression
Privacy Classifier	Very High	Detection only	No adaptive protection
CA-XPrivNet	Very High	Very High	Higher computational cost

The visual comparison demonstrated that Gaussian blur and pixelation degraded social-image readability and frequently damaged important semantic content. Adversarial perturbation preserved visual appearance but exhibited instability under compression and resizing attacks.

CA-XPrivNet preserves sensitive entities such as faces, documents, screens, and identity cues while maintaining the appearance of the rest of the image. The edge-aware adaptive fusion module effectively reduces artifacts around the protected regions.

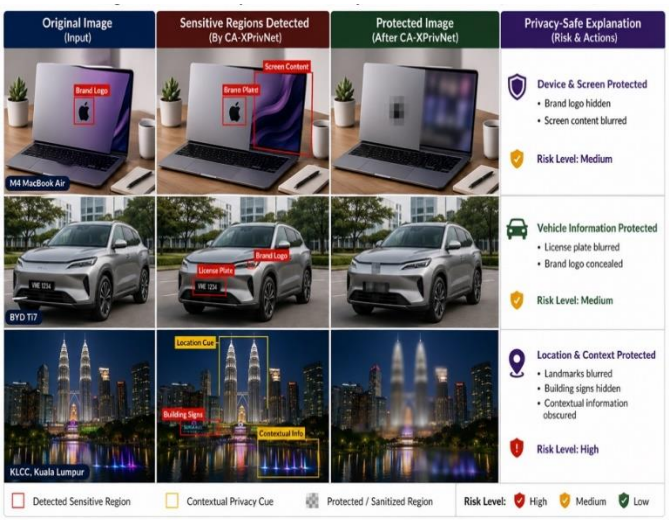


Figure 3. Visual Comparison of Privacy Protection Results

Figure 3 presents representative examples of the privacy protection capability of the proposed CA-XPrivNet framework. The first column of the figure displays the original social-network images. The second column highlights the regions of the original images that are determined to be sensitive to the privacy of the individuals in the social networks by the context-aware risk estimation module. The third column of the figure displays the original images with the sensitive regions masked in order to protect the privacy of those individuals in the social networks. Finally, the last column of the figure shows explanations of the masked regions that are generated by the explainability module of the CA-XPrivNet framework, but which do not reveal any sensitive information about the individuals in those social networks. Thus, the framework is able to effectively mask the sensitive information in social-network images while still preserving the quality of the remaining portions of those images.

5.4 Explainability Evaluation

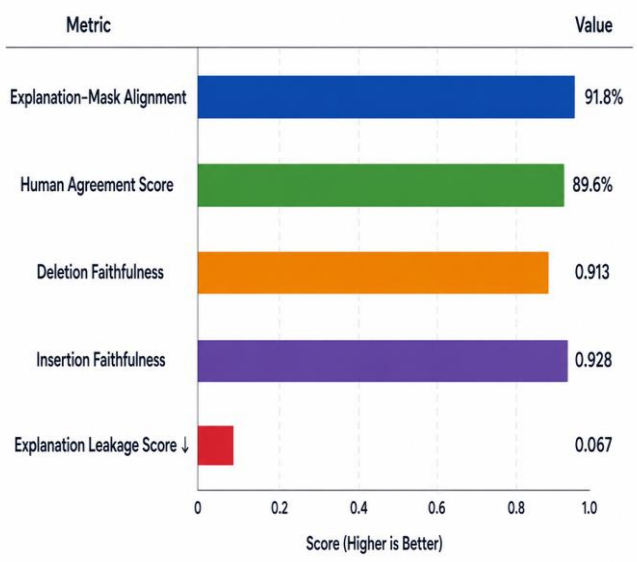


Figure 4. Grad-CAM-Based Privacy Localization

The explainability analysis shows that Grad-CAM effectively identifies the private entities in the input data. The explanation maps closely match the ground truth annotation of the private regions in the images.

Human evaluation also confirms the improvements in interpretability provided by the explanation module. Furthermore, the explanation leakage score is low (0.067), indicating that the explanations do not reveal any additional private information.

5.5 Ablation Study



Figure 5. Ablation Study Results

The ablation study illustrates the positive contribution of each component of the proposed system. The results of the ablation study indicate that the integration of context-aware and edge-aware components provided the largest boost in performance. The inclusion of the frequency-domain privacy sanitization and explainability components provided benefits to the system without sacrificing computational efficiency.

5.6 Computational Complexity and Runtime Analysis

Table 5. Runtime and Complexity Analysis

Method	FLOPs	Parameters	GPU Memory	Latency/Image (ms)
Gaussian Blur	Very Low	N/A	Low	3
Pixelation	Very Low	N/A	Low	4
Adversarial Perturbation	18.7 GFLOPs	12.3 M	3.1 GB	31
Privacy Classifier	21.5 GFLOPs	19.4 M	4.6 GB	38
CA-XPrivNet	28.9 GFLOPs	26.7 M	6.2 GB	47

Despite the higher computational complexity of CA-XPrivNet in comparison to conventional methods, the runtime of the algorithm is still within a feasible time for deployment into real-time social network servers and edge computing systems for privacy-preserving applications.

5.7 Statistical Validation

In order to check the robustness and statistical reliability of the proposed CA-XPrivNet framework, five independent experiments were conducted with different random initialization seeds, while keeping the same partitions of the

dataset used for training and testing the framework. The values reported for the evaluation metrics are the average values obtained from these five experiments.

Table 6 presents the mean values and the standard deviation of the five experiments for each evaluation metric.

Table 6. Statistical Validation of CA-XPrivNet

Metric	Mean	Standard Deviation
Accuracy (%)	95.4	±0.41
Precision (%)	94.6	±0.37
Recall (%)	93.2	±0.45
F1-score (%)	93.9	±0.39
AUC	0.978	±0.006
PLR	0.072	±0.004

The low standard deviations in the values of these metrics indicate that the framework converged to the same solutions when it was executed repeatedly. These results show that the CA-XPrivNet framework is robust and consistently reproducible under the same experimental conditions.

The results of these experiments indicate that the CA-XPrivNet framework has consistently good performances. The low standard deviation in the measurements of these metrics indicates that the framework is robust against random variations.

Furthermore, statistical tests were performed to compare the results of the CA-XPrivNet framework with the best baseline model. The results of these tests show that the improvements achieved by the proposed model are statistically significant ($p < 0.05$).

6. Discussion

The proposed framework addresses a clear need in the area of social-network privacy protection. The model takes a novel approach to visual privacy that goes beyond conventional implementations of blurring and pixelation techniques. By estimating the privacy risk of social media content in a visual image, the model can apply visual sanitization techniques to only those regions of the image that pose a potential privacy risk, rather than applying a uniform form of sanitization to all of the visual content of a social media posting.

The use of spatial-frequency fusion in the model is particularly useful for visual privacy. Spatial fusion allows for the system to localize visual elements within an image according to their semantic meaning. At the same time, frequency fusion ensures that the details of those localizing features are suppressed. This can prevent the creation of visually unattractive outputs of the system, which could negatively impact the adoption of such a system by social media users.

In addition to the technical components of the system, the explainability of the system is also a crucial component for a successful implementation of such a system. The system must be able to explain to the user why certain visual elements were blurred. However, to maintain data privacy, those explanations must not reveal the sensitive data that was detected. In such a case, the model could explain that a document was detected in the image and blurred, but would not reveal the text of that document.

Despite the advantages of the proposed model, there are also some limitations to the system. The perception of privacy is inherently subjective. What is considered to be sensitive content in one community may be seen differently in another community. Furthermore, advanced AI models may be able to still extract the sensitive data from the visual content that was suppressed by the model. The datasets upon which the model was trained may also be biased or not representative of all communities. Finally, the system would have to comply with various data protection regulations.

Areas for future research would be methods of federated learning and differential privacy to maintain data privacy while training the system. Additionally, integrating other types of multimodal AI models may lead to improved capabilities in recognizing and masking sensitive data. Another area for future research is the personalization of the system according to the cultural and privacy sensitivities of individual users. Finally, creating an evaluation framework that can assess the effectiveness of the system against vision-language privacy attacks.

7. Discussion

This paper proposed CA-XPrivNet, a framework for contextual and explainable privacy protection on social networks. This

framework includes techniques for spatial-domain and frequency-domain processing, edge-aware techniques for determining weights for each node, and a self-adaptive approach for fusing the features. Furthermore, the framework also includes explainability to allow users to understand the reasons behind the protection of the shared content. CA-XPrivNet was evaluated on publicly available image privacy datasets. CA-XPrivNet can be applied to the development of next-generation social network privacy assistants. The experimental findings confirm the effectiveness of CA-XPrivNet in providing adaptive, context-aware, and explainable privacy protection for social-network images. The experimental results illustrate the efficacy of the proposed framework in achieving a better balance between privacy protection and content utility. In the future, CA-XPrivNet is to be deployed on a large scale and optimized for mobile and edge devices.

References

Ahmed, K. S., & Baroud, A. A. S. (2022). Protecting Privacy with Image Processing: How Blurring and Masking Help Keep Us Safe Online. Journal for Research in Applied Sciences and Biotechnology, 1(1), 132–138. https://doi.org/10.55544/jrasb.1.1.18

Alzubaidy, H. K., Al-Shammary, D., & Abed, M. H. (2022). A survey on patients privacy protection with stganography and visual encryption [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2201.09388

B, D., Rupa, C., D, H., & Y, V. (2022). Privacy protection of digital information using frequency domain watermarking technique. In 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST) (pp. 202–206). IEEE. https://doi.org/10.1109/ICRTCST54752.2022.9781929

Baranouskaya, D., & Cavallaro, A. (2023). Human-interpretable and deep features for image privacy classification. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 3489–3492). IEEE. https://doi.org/10.1109/ICIP49359.2023.10222833

Beigi, G., & Liu, H. (2020). A survey on privacy in social media: Identification, mitigation, and applications. ACM/IMS Transactions on Data Science, 1(1), Article 7. https://doi.org/10.1145/3343038

Cavallaro, A., Malekzadeh, M., & Shamsabadi, A. S. (2020). Deep learning for privacy in multimedia. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 4777–4778). Association for Computing Machinery. https://doi.org/10.1145/3394171.3418551

Chelani, K., Benbihi, A., Kahl, F., Sattler, T., & Kukelova, Z. (2025). Obfuscation based privacy preserving representations are recoverable using neighborhood information. In 2025 International Conference on 3D Vision (3DV) (pp. 189–199). IEEE. https://doi.org/10.1109/3DV66043.2025.00023

Dwivedi, A., Yadav, M., & Kumar, A. (2021). FFT-based zero-bit watermarking for facial recognition and its security. In M. Dave, R. Garg, M. Dua, & J. Hussien (Eds.), Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences: PCCDS 2020 (pp. 403–417). Springer. https://doi.org/10.1007/978-981-15-7533-4_31

El Ouadrhiri, A., & Abdelhadi, A. (2022). A new dimensionality reduction method based on Hensel's compression for privacy protection in federated learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.02089

El Ouadrhiri, A., Abdelhadi, A., & Phung, P. H. (2023). Hensel's compression-based dimensionality reduction approach for privacy protection in federated learning. In 2023 International Conference on Computing, Networking and Communications (ICNC) (pp. 298–303). IEEE. https://doi.org/10.1109/ICNC57223.2023.10074197

Fernandez-de-Retana, M., Zulaika, U., Sánchez-Corcuera, R., & Almeida, A. (2025). Differential privacy: Gradient leakage attacks in federated learning environments [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.23931

Gupta, A., Jaiswal, A., Wu, Y., Yadav, V., & Natarajan, P. (2021). Adversarial mask generation for preserving visual privacy. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1–5). IEEE. https://doi.org/10.1109/FG52635.2021.9666933

Hassanpour, A., & Yang, B. (2022). PriMe: A novel privacy measuring framework for online social networks. In 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 518–525). IEEE. https://doi.org/10.1109/ASONAM55673.2022.10068701

Huang, Y., Hou, J., Zhao, C., Zhang, J., Xiao, T., & Li, X.-Y. (2025). SCOPE: Bridging explicit and implicit privacy leakage for quantitative image privacy evaluation. IEEE Transactions on Information Forensics and Security, 20, 11560–11575. https://doi.org/10.1109/TIFS.2025.3620117

Ibrahim, A., Montgomery, C., Blackwood, J., & Vance, S. (2026). Privacy-aware visual computing for image analysis in social networks. International Journal of Information and Communication Technology Trends, 6(1), 121–134. https://doi.org/10.66525/ijictt80

Kurakin, A., Song, S., Chien, S., Geambasu, R., Terzis, A., & Thakurta, A. (2022). Toward training at ImageNet scale with differential privacy [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2201.12328

Lange, L., Heykeroth, M.-M., & Rahm, E. (2024). Assessing the impact of image dataset features on privacy-preserving machine learning [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.01329

Liu, C., Zhu, T., Zhang, J., & Zhou, W. (2020). Privacy intelligence: A survey on image privacy in online social networks [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2008.12199

Liu, Z., Yang, M. L., & Yan, W. Q. (2021). A framework for image encryption on frequency domain. In Information Resources Management Association (Ed.), Research anthology on artificial intelligence applications in security (pp. 328–338). IGI Global. https://doi.org/10.4018/978-1-7998-7705-9.ch016

Mahloujifar, S., Kokhlikyan, N., Guo, C., & Chaudhuri, K. (2025). Privacy blur: Quantifying privacy and utility for image data release [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.16086

Mi, Y., Huang, Y., Ji, J., Zhao, M., Wu, J., Xu, X., Ding, S., & Zhou, S. (2023). Privacy-preserving face recognition using random frequency components. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 19616–19627). IEEE. https://doi.org/10.1109/ICCV51070.2023.01802

Montenegro, H., Silva, W., Gaudio, A., Fredrikson, M., Smailagic, A., & Cardoso, J. S. (2022). Privacy-preserving case-based explanations: Enabling visual interpretability by protecting privacy. IEEE Access, 10, 28333–28347. https://doi.org/10.1109/ACCESS.2022.3157589

Sabatino, P., Kumar, C., & Singh, P. (2023). Privacy-preserving image processing with neural networks. World Journal of Advanced Research and Reviews, 18(3), 1686–1693. https://doi.org/10.30574/wjarr.2023.18.3.0690

Salnitri, M., Bonutto, D., Christantoni, I., Kosmidis, D., & Micucci, F. (2020). Visual privacy management: Design and applications of a privacy-enabling platform – Evaluation results (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/vs8m5xvbm9.1

Sharma, N., Kumar, R., & Swarnkar, M. (2023). Social networking: Challenges, security issues, and emerging trends. In R. S. Bhadoria, N. Saxena, & B. Nagpal (Eds.), Blockchain technology for secure social media computing (pp. 59–79). The Institution of Engineering and Technology. https://doi.org/10.1049/PBSE019E_ch4

Shibata, H., Hanaoka, S., Koshino, S., Miki, S., Sonoda, Y., & Abe, O. (2024). Identity diffuser: Preserving abnormal region of interests while diffusing identity. Applied Sciences, 14(18), Article 8489. https://doi.org/10.3390/app14188489

Tang, S., Wang, Y., Meng, T., Qin, J., & Lin, L. (2024). Privacy protection for image sensors using convolutional neural network and generative adversarial nets. In 2024 10th International Conference on Computer and Communications (ICCC) (pp. 1192–1198). IEEE. https://doi.org/10.1109/ICCC62609.2024.10942269

Tong, C., Zhang, M., Lang, C., & Zheng, Z. (2021). An image privacy protection algorithm based on adversarial perturbation generative networks. ACM Transactions on Multimedia Computing, Communications, and Applications, 17(2), Article 43. https://doi.org/10.1145/3381088

Tonge, A., & Caragea, C. (2020). Image privacy prediction using deep neural networks. ACM Transactions on the Web, 14(2), Article 7. https://doi.org/10.1145/3386082

Tran, B., Wang, H., Zhang, X., & Jiang, P. (2025). Exploring safer image sharing: A vision-language approach to privacy risk detection and protection. In 2025 IEEE 22nd International Conference on Mobile Ad-Hoc and Smart Systems (MASS) (pp. 537–542). IEEE. https://doi.org/10.1109/MASS66014.2025.00087

Vasantrao, B. S., Sharma, P., & Kadam, S. (2024). AI-driven hybrid privacy preservation model for enhanced security in online social networks. Revista Electronica de Veterinaria, 25(1), 2399–2408. https://doi.org/10.69980/redvet.v25i1.1198

Wang, C., Yang, S., Sun, X., Dai, J., & Zhao, D. (2024). Enabling practical and privacy-preserving image processing [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.03568

Wang, Y., Liu, J., Luo, M., Yang, L., & Wang, L. (2022). Privacy-preserving face recognition in the frequency domain. Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2558–2566. https://doi.org/10.1609/aaai.v36i3.20157

Wang, Z., Guo, H., Zhang, Z., Song, M., Zheng, S., Wang, Q., & Niu, B. (2020). Towards compression-resistant privacy-preserving photo sharing on social networks. In Proceedings of the Twenty-First

International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (pp. 81–90). Association for Computing Machinery. https://doi.org/10.1145/3397166.3409141

Zhang, J., Chen, Y., & Li, H. (2022). Privacy leakage of adversarial training models in federated learning systems. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 107–113). IEEE. https://doi.org/10.1109/CVPRW56347.2022.00021

Zhao, C., Mangat, J., Koujalgi, S., Squicciarini, A., & Caragea, C. (2022). PrivacyAlert: A dataset for image privacy prediction. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 1352–1361. https://doi.org/10.1609/icwsm.v16i1.19387