IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 March 2024, accepted 26 April 2024, date of publication 2 May 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396210

RESEARCH ARTICLE

DP Patch: ROI-Based Approach of Privacy-Preserving Image Processing With Robust Classification

ASSEM UTALIYEVA, (Student Member, IEEE), SEON-JIN HWANG, AND YOON-HO CHOI, (Member, IEEE)
School of Computer Science and Engineering, Pusan National University, Busan 46241, Republic of Korea

Corresponding author: Yoon-Ho Choi (yhchoi@pusan.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2022R1I1A3055233; in part by the NRF Grant funded by the Korean Government, Ministry of Science and ICT (MSIT), South Korea, under Grant RS-2023-00217689; and in part by the MSIT through the Information Technology Research Center (ITRC) Support Program, Supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP), under Grant IITP-2023-RS-2023-00259967.

ABSTRACT As machine and deep learning spread across diverse aspects of our society, the concerns about the privacy of the data are getting stronger, particularly in scenarios where sensitive information could be exposed as a result of various privacy attacks. This paper introduces a novel framework, DP Patch, aimed at addressing these privacy concerns in image data by considering sensitive objects that could be located within the image rather than considering the entire image as sensitive. DP Patch involves a multi-step pipeline, which consists of differential privacy image denoising and ROI-based sensitive object localization, followed by incorporating DP noise patches to obscure sensitive content. This process yields privacy-preserving images with enhanced utility compared to DP images. Furthermore, a custom model is presented that harnesses privacy-preserving and differentially private images to enrich feature representation and compensate for potential information loss, explicitly excluding the noisy patch from the training process. Experimental evaluations are conducted to assess the quality of the generated privacy-preserving images and to compare the performance of the custom model against state-of-the-art counterparts. Additionally, the proposed method undergoes evaluation under model inversion attacks, providing practical insights into its effectiveness.

INDEX TERMS Differential privacy, PPDL, privacy-preservation.

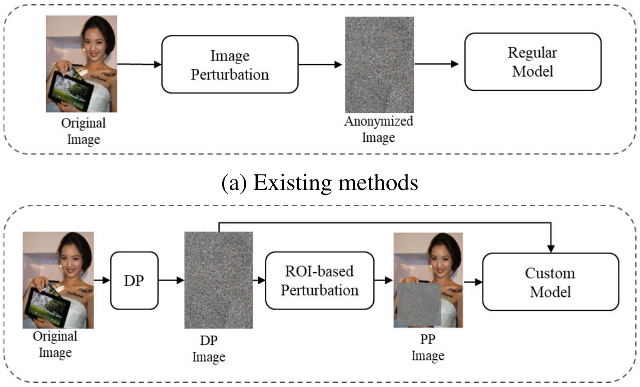
I. INTRODUCTION

The data has become the fuel for ubiquitous development as machine and deep learning methods are deployed in every aspect of our lives. However, such an increase in data needs leads to continuously increasing concerns about data privacy. For instance, Clearview AI faced fines and legal challenges in the U.K., Italy, France, and several other countries for its controversial practice of scraping images for its facial recognition database without user consent [1].

Adversaries can compromise data privacy in various ways, such as during transmission over the unsecured network or even reconstructed from the deep learning model trained on that data. Specifically, various privacy attacks such as model inversion [2] were introduced as an effective way to reconstruct the images from train data with high confidence due to the memorization characteristics of the deep learning models.

Various anonymization, obfuscation, and encoding methods have been actively studied to protect the privacy of image data. For instance, A. Huang et al. introduced InstaHide [3], which encodes train images by mixing multiple images and

The associate editor coordinating the review of this manuscript and approving it for publication was Peter Langendoerfer.



(a) Existing methods

(b) Proposed method

FIGURE 1. High-level overview of the difference between (a) existing and (b) proposed methods.

applying a pixel-wise mask. In comparison, T.Li and M.Choi proposed the DeepBlur [4] method, primarily focusing on facial obfuscation. Another way to protect the privacy of the images is to generate a synthetic dataset. There are various generative models such as variational autoencoders (VAE) [5], generative adversarial networks (GAN) [6], and diffusion models [7] adopted to the privacy-preservation tasks [8], [9].

Differential privacy (DP) is also recognized as a powerful concept to guarantee data privacy by adding random noise. There are various strategies for generating differentially private images. Adding random noise directly to the image is one of the most substantial ways to guarantee its privacy by making it utterly unrecognizable to the human eye. The incorporation of DP into the training process of generative models has been actively studied by researchers, resulting in variations of differentially private GANs [10], [11], [12]. Also, B.Liu et al. proposed a novel way of generating differentially private images by perturbating the feature space [13], while L. Fan proposed methods of differentially private image pixelization [14] and differentially private image singular value decomposition to obfuscate the images [15].

However, existing privacy-preserving methods have several limitations that hinder widespread usage of these methods. The first limitation is the privacy and utility tradeoff problem, which can be described as balancing the need to protect individuals' privacy with maintaining the usefulness of data for analysis and deep learning applications. One typical result of such a problem is that to protect the image's privacy from both human and machine adversaries, the image is made visually unrecognizable both to the human eyes and to the machine. Therefore, the accuracy of computer vision tasks performed on such anonymized and manipulated images is low. For instance, in the case of synthetic data, the quality and realism of generated image samples play a crucial role in its further usage. Meanwhile, in the case of differentially private images, the addition of random noise effects can distort critical features needed for computer vision tasks. Even though some of the methods, such as InstaHide [3], argue that generated images can both protect privacy and preserve high performance in image recognition tasks, N.Carlini et al. argue that encoding the train data only and feeding into the regular non-private learning algorithm is not enough to protect the data privacy [16], practically showing the successful image reconstruction attack on InstaHide.

Another limitation of existing methods is the assumption that the entire image is private. Consequently, privacy-preservation techniques are often applied to all pixels within the image, rendering it visually unrecognizable. However, it is essential to note that sensitive information may be confined to a small portion of the image or might not be present. Applying privacy-preservation techniques indiscriminately to pixels lacking sensitive content can unnecessarily compromise the image's overall usability. Our focus here is not on specific methods like facial de-identification [17], [18], which targets face recognition systems, but on general-purpose techniques.

To address these limitations, the paper proposes DP Patch, a novel framework that (1) generates privacy-preserving (PP) images from a specified region of interest (ROI) within differentially private images and (2) introduces a custom deep learning model designed explicitly for privacy-preserving images. In this method, differentially private images, which are unrecognizable visually and by machines, are used for secure transmission over networks. Subsequently, the privacy-preserving image is generated from the differentially private image, focusing on the ROI of sensitive objects within the image. Furthermore, a tailored model architecture is proposed to enhance the performance of computer vision tasks on these privacy-preserving images. Fig. 1 shows the high-level difference between the existing and proposed methods.

The main contribution of this paper can be summarized as follows:

- A privacy-preserving (PP) image generation technique is developed, involving creating PP images using noisy DP images and associated ROI of the sensitive objects. This method incorporates a pipeline for DP image denoising and sensitive object identification, followed by the addition of DP noise patches to obscure sensitive objects, resulting in PP images that retain higher utility compared to DP images. The generated PP images also include a supplementary binary mask indicating the application of DP noise patches.
- A Deep Learning (DL) model tailored for privacy-preserving image classification is presented in this paper. This model considers PP and DP images as inputs to enhance feature representation, compensating for potential loss during denoising. Incorporating channel-wise attention blocks allows for accentuating discriminative channels and suppressing less relevant ones during feature extraction. Masked feature maps are generated by utilizing a binary mask to identify the location of the noise patch. A weighted concatenation, assigning lower weight to DP images, serves as a feature fusion technique before feeding into the classification layers.

- Additionally, the DP Patch framework is presented, which combines privacy-preserving image generation modules and customized deep learning modules to guarantee higher utility in image classification and can be applied across various scenarios, including federated and distributed learning.

The rest of the paper is organized as follows. In section II, the preliminary information is introduced, and a brief overview of existing privacy-preserving image generation methods and their comparison is provided. In section III, the threat model needed to understand the motivation behind the proposed framework is presented. Next, an overall workflow and detailed description of the methodology of the proposed method are provided in section IV, and the privacy expectations are discussed in section V and experimentally evaluated in section VI. Finally, section VII concludes the paper.

II. PRELIMINARIES AND RELATED WORK

This section briefly describes the concept of DP, existing methods to generate privacy-preserving images, and an overview of the noise-robust deep learning models, including the image denoising techniques.

A. DIFFERENTIAL PRIVACY

DP is recognized as the gold standard for privacy preservation. It bounds the maximum information leakage by ensuring that the probability of any output does not increase by more than a factor of e^{ϵ} due to the presence or absence of a single individual's data, with an additional small allowance of δ for the probability of this guarantee being exceeded. This protection is achieved by injecting carefully calibrated random noise into the data.

Previous studies primarily focused on adding differentially private random perturbations to statistical databases, as described in [28]. More recent approaches, however, extend the application of DP to the model's hyperparameters during the training process. This is exemplified in methods like Differential Privacy Stochastic Gradient Descent (DPSGD) [29], which aim to protect the deep learning model itself from privacy attacks, such as [2] and [30].

Nevertheless, even when a model is trained with differentially private hyperparameters, there remains a possibility that sensitive information can be recovered from the training set as a result of such privacy attacks. Therefore, one of the more effective strategies is to apply differentially private random noise directly to the data, thereby safeguarding the sensitive information from potential recovery. This approach, however, presents challenges due to the privacy-utility tradeoff problem, where highly private data may become less useful for subsequent analysis.

B. PRIVACY-PRESERVING IMAGE GENERATION METHODS

1) DP-BASED METHODS

L.Fan proposed generating differentially private images by combining the random perturbations with previously known methods such as image pixelization [14] and singular value decomposition [15]. DP-Image [13] and IdentityDP [25] are the frameworks to generate differentially private images mainly targeting the face obfuscation task by introducing random perturbations to the feature space.

In the case of generative models such as GANs [10], [11] and diffusion models [31], the DP is reached by injecting noise during the training process, usually by incorporating the DPSGD algorithm to ensure differentially private image generation. [19] utilized the PATE framework among multiple discriminators to ensure the generator model achieves DP. Unlike previous methods, J.Chen et al. introduced DPGEN [20] to generate images using the energy-guided network and Langevin Markov chain Monte Carlo sampling. DP-MERF [21] and PEARL [22] are the frameworks to generate images using differentially private mean embeddings.

2) OTHER METHODS

GANs are widely adapted in the face anonymization tasks such as DeepPrivacy [23] and CiaGAN [24] that utilize conditional GANs to replace the original face with another realistic face. Different from the approach that utilizes generative modeling, Ko et al. [26] proposed a structural image de-identification technique that aims to reduce the ability of humans to recognize the image while preserving the performance of computer vision task. Huang et al. introduced InstaHide [3] that encodes the images by mixing it with other randomly selected images. Yu et al. proposed IPrivacy [27] that can automatically identify the presence of sensitive information within the image using the multi-task learning algorithm and blurring it. However, this method does not utilize the potential benefits offered by more modern approaches such as automated object detection, that can be re-annotated according to the given task.

Table 1, provides a comprehensive summary of the existing methods in image privacy protection, evaluated against multiple criteria. These criteria include the specific target deemed as private and the consideration of a specialized deep learning model for the task. Based on our extensive review, DP Patch emerges as the first framework that simultaneously focuses on ROI-based protection of sensitive objects and incorporates a custom DL model specifically designed for processing such images.

C. NOISE-ROBUST DL MODELS

This section overviews the existing work in image denoising and noise-robust deep learning models in computer vision tasks.

1) IMAGE DENOISING

K.Zhang et al. introduced DnCNN [32]. This Gaussian denoiser leverages residual connections to perform high-quality image denoising under Gaussian noise up to $\sigma = 50$, and later presented FFDNET [33] that can denoise Gaussian noise up to $\sigma = 75$ with a single model.

TABLE 1. Characteristics of the image privacy protection methods.

Method Type	Method Name	Target			DL model
		Entire Image	Human Face	ROI	
Image Synthesis	DPGAN [10]	✓			Any
	GAN-Obfuscator [11]	✓			Any
	G-PATE [19]	✓			Any
	DPGEN [20]	✓			Any
	DP-MERF [21]	✓			Any
	PEARL [22]	✓			Any
	DeepPrivacy [23]	✓			Any
	CiaGAN [24]	✓			Any
Image Perturbation	DP-Pix [14]	✓			Any
	DP-SVD [15]	✓			Any
	DP-Image [13]		✓		Any
	Identity-DP [25]		✓		Any
	Structural De-Ident. [26]	✓			Any
	InstaHide [3]	✓			Any
	IPrivacy [27]			✓	Any
	DP Patch (proposed)			✓	**Custom**

However, standalone image denoising is not considered in this paper since the focus is not on the complete denoising of the image, which would make the application of DP ineffective.

2) NOISE-ROBUST DEEP LEARNING MODELS

Several works have studied the robustness of deep learning models to random noise. M. Momeny et al. proposed NR-CNN [34], which excludes noise pixels from the classification task, relying on the not corrupted pixels. While X.Meng et al. introduced the CNR-CNN [35], which denoises the noisy input prior to the face recognition task. [36] presents a different approach to extract the most discriminative features from the single noisy input in the final fully connected layer. However, existing methods primarily consider a small amount of random noise that slightly affects the computer vision tasks or may consider a large number of clean and uncorrupted pixels. The differentially private noise is way larger, and existing models are not able to perform robust computer vision tasks, nor are they considering the privacy preservation nature of the random noise.

D. OTHER IMAGE PROCESSING METHODS

Several other studies in the image retrieval and classification field incorporate multiple input images to augment the feature extraction, considering the data uncertainty. For instance, Regan and Khodayar [37] employs sparse representations and dictionary learning to efficiently classify the weather conditions from multiple input sources using the triplet graph network, while Saffari [38] incorporates multiple inputs to improve the feature discrimination and retrieval performance in the domain of traffic scene classification. These studies emphasize the importance of multi-input integration to advance the image classification task. Also, Qadir [39] proposed an active learning method that significantly improved the classification accuracy by selecting the most informative samples.

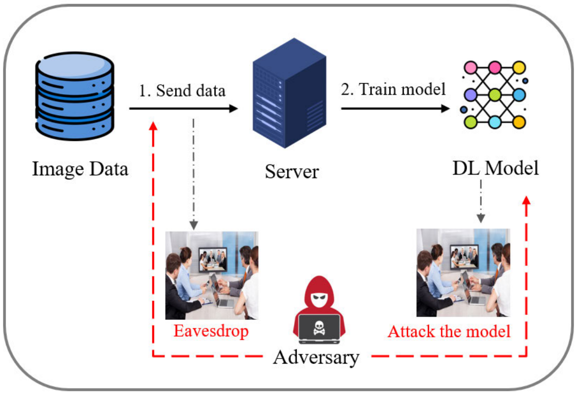


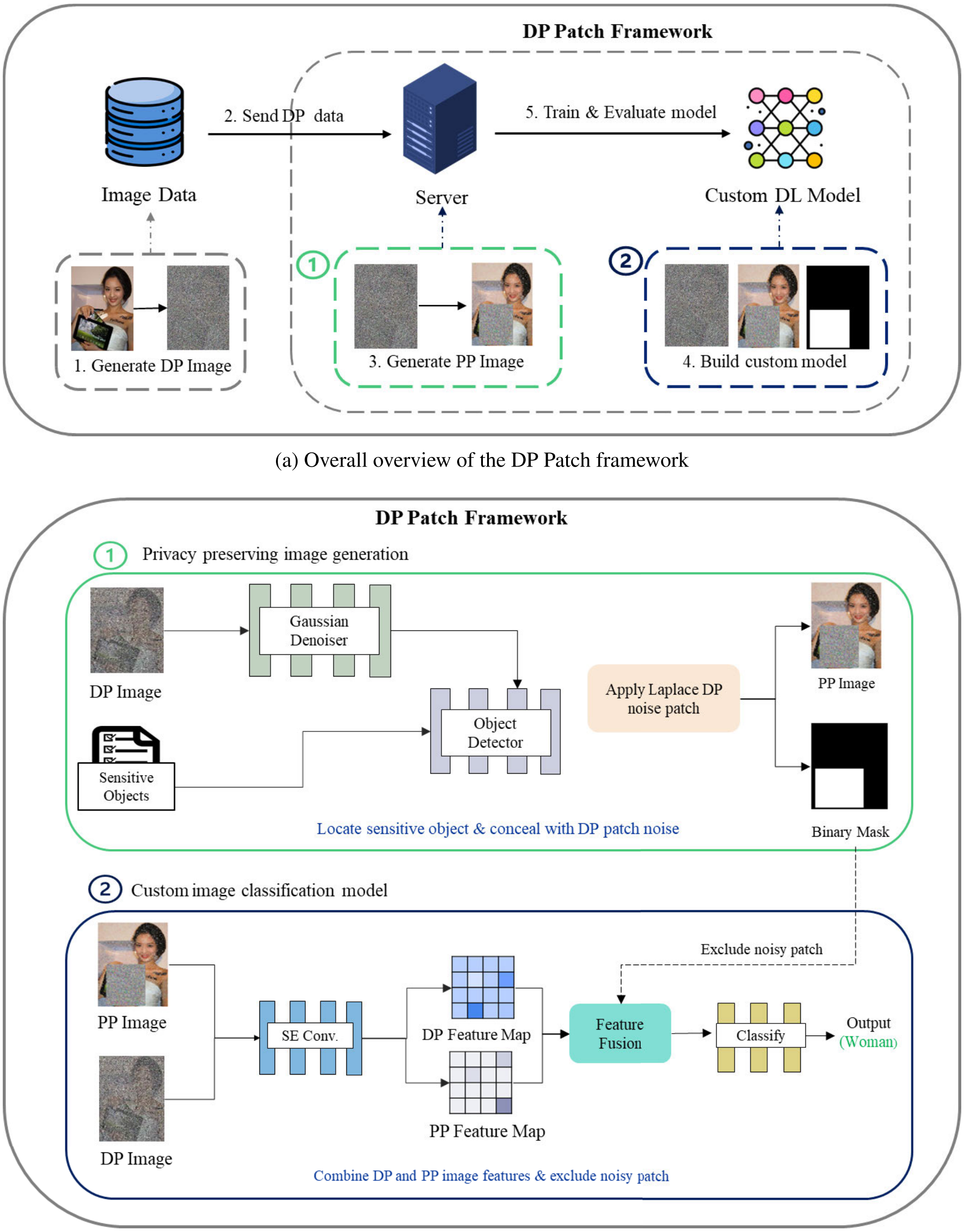
FIGURE 2. Example of privacy exposures.

III. THREAT MODEL

Let us consider the threat model illustrated in Fig. 2. In this context, two primary vectors that potentially compromise the privacy of image data are identified.

The first vulnerability occurs during the transmission of image data over network channels. Although encryption protocols such as SSL and TLS secure the communication channel, they are not entirely infallible [40], [41], [42]. For instance, Man-in-the-Middle (MITM) [43] attacks can still occur due to various factors, such as protocol misconfiguration, compromised certificate authorities, SSL stripping, and social engineering. As a result of a successful MITM attack, an attacker can intercept, alter, and capture the data between two communicating parties. Therefore, this stage is still prone to attacks that allow attackers to access original images and reveal sensitive information despite the implementation of communication-level security.

The second major threat involves the deep learning model trained on these images. If an adversary accesses this model, they could execute model inversion attacks as introduced by Fredrickson et al. [2], where an adversary infers sensitive



(a) Overall overview of the DP Patch framework

(b) Detailed view of the DP Patch framework

FIGURE 3. The DP patch framework.

information about instances by analyzing the model's outputs, like confidence scores. Zhang et al.introduced the Secret Revealer [44], a model inversion attack based on generative adversarial networks designed to accurately recover instances from the training set, posing a significant privacy risk, especially if the training set contains sensitive data.

IV. METHODOLOGY

This section provides a detailed description of the proposed method by explaining the end-to-end methodology.

A. OVERVIEW

The proposed DP Patch framework is designed to facilitate privacy-preserving image sharing and enhance computer vision models' ability to provide a robust classification of images. The overall workflow of the proposed method is shown in Fig. 3a, where the differentially private images are generated prior to the application of the DP patch framework to guarantee secure transmission over the network. Once the data is received, privacy-preserving images are generated based on the ROI of the sensitive object. Consequently,

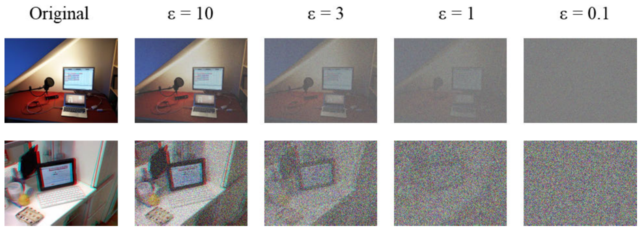


FIGURE 4. Examples of DP images generated with different ϵ values and fixed $\delta = 10^{-6}$.

generated privacy-preserving images and received differentially private images are fed into the custom deep learning model designed to handle both images.

The detailed overview of the proposed framework is shown in Fig. 3b, where the methodology is largely divided into two concepts *privacy preserving image generation* and *custom image classification model*.

The concept of *privacy preserving image generation* lies in hiding only objects that are recognized as sensitive rather than obscuring the entire image, which will distort the utility of that image. The key idea of privacy-preserving image generation is to run a denoising and object detection pipeline on the differentially private input, locate the ROI based on the predefined list of sensitive objects, and add patch noise of different DP distributions to hinder further restoration.

However, the generated privacy-preserving image can not guarantee the same performance in computer vision tasks as the original image due to the following limitations: (1) the denoising process results in a lower quality image, which may lose important information; (2) the noise patch that conceals the sensitive object may confuse the deep learning model trying to extract the features from that area; To resolve these limitations, we introduce a model that is customized for privacy-preserving images. The model takes multiple images as input, precisely the generated privacy-preserving image, noisy differentially private image, and binary mask. Here, the differentially private image is used as a source of a complimentary feature set to compensate for the information loss during the denoising process. In contrast, a binary mask is used to divert the model's attention from the noise patch.

B. DIFFERENTIALLY PRIVATE IMAGE GENERATION

1) NOTATIONS

Let us denote the original images as I_{orig}, differentially private images as I_{DP}, DP privacy parameter as ϵ and relaxation parameter as δ.

2) DESCRIPTION

Differentially private noise is applied to each pixel of the clean image using the Gaussian DP mechanism. This process can be formulated as $I_{DP} = I_{orig} + N(\epsilon, \delta)$, where $N(\epsilon, \delta)$ represents the noise calibrated according to the values of ϵ and δ. In this context, ϵ is manually defined in the range of 0.01 to 10. A value of 0.01 corresponds to the highest level of privacy, accompanied by the greatest degree of image distortion. Conversely, a value of 10 indicates the lowest level of privacy and minimal image distortion. The relaxation parameter δ assigned an extremely small value, such as 10^{-6}, to permit a minor deviation from the stringent privacy bounds. Fig. 4 illustrates the examples of DP images generated under different privacy parameter values with fixed δ values. Images generated with small ϵ values has the highest amount of noise added to the pixels.

The primary objective of employing this method for differentially private image generation is to ensure the secure transmission of images from the client side. The aim is to produce an unrecognizable image for human observers and automated systems.

It is important to note that current state-of-the-art image-denoising models are primarily designed to reduce noise and restore low-quality images. However, these models do not typically account for the unique noise distributions associated with differential privacy.

C. PRIVACY PRESERVING IMAGE GENERATION

1) NOTATIONS

Let us denote the list of sensitive objects as S. The denoising and object detection modules are denoted as D and O, respectively. The generated privacy-preserving image is also denoted as I_{PP}. Supplementary binary mask generated alongside the I_{PP} is denoted as BM.

2) DESCRIPTION

The procedure for creating a privacy-preserving image from a differentially private input I_{DP} is concisely outlined in algorithm 1. To generate the privacy-preserving image I_{PP} from the noisy I_{DP}, it is necessary to undertake image denoising and sensitive object detection tasks. It has been previously established that conventional, state-of-the-art image-denoising models cannot effectively handle the noise in a DP-enhanced image. Consequently, we have adapted the architecture of the DnCNN denoising model by incorporating additional residual connections. This modified model was then re-trained to accommodate the noise distribution characteristic of differential privacy, particularly Gaussian noise.

Considering the assumption that the DP noise adheres to a Gaussian distribution, the denoising module D can approximate the amount of random noise by estimating the value of σ and accordingly denoise the I_{DP} image as shown in line 2 of the algorithm 1.

The following equation can calculate the actual σ values of differentially private Gaussian noise addition:

$$\sigma^2 = \frac{2 \cdot \ln(1.25/\delta) \cdot \Delta f^2}{\epsilon^2} \tag{1}$$

Given the privacy parameter ε = [1,3,10] and relaxation parameter $\delta = [10^{-6}, 10^{-5}, 10^{-4}]$, and the sensitivity $\Delta f = 255$, the calculated σ values vary from 1100 ∼ 1350 for

Algorithm 1 Generate I_{PP}

```
1:  procedure Generate $I_{PP}(I_{DP}, \varepsilon, \delta)$
2:      $\sigma^2 \leftarrow \frac{2 \cdot \ln(1.25/\delta) \cdot \Delta f^2}{\epsilon^2}$        ▷ Estimate $\sigma$
3:      for each $I_{DP}$ do
4:          $I_{denoised} \leftarrow D(I_{DP}, \sigma)$
5:          $O_{sensitive} \leftarrow O(I_{denoised}) \cap S$
6:          Initialize $patch$ to zero
7:          for each $o \in O_{sensitive}$ do
8:              $patch(o) \leftarrow I_{denoised}(o) + DP(o)$
9:          end for
10:         $I_{PP} \leftarrow I_{denoised} + patch$
11:         Initialize $BM$ as zero array
12:         for each pixel $p$ in $I_{PP}$ do
13:             $BM(p) \leftarrow \begin{cases} 0, & \text{if } p \in patch \\ 1, & \text{otherwise} \end{cases}$
14:         end for
15:     end for
16: return $I_{PP}$
17: end procedure
```

$\epsilon = 1$, 370 ∼ 450 for $\varepsilon = 3$, and 110 ∼ 135 for $\varepsilon = 10$. The smaller the σ values, the easier it is to denoise the noisy image and the better the quality of the result. Thus, the quality of the denoised image, as well as the quality of the I_{PP} image to be further generated, directly depends on the ε value. Therefore, the trade-off between image quality and the denoising process is similar to the privacy-utility trade-off observed in regular DP mechanisms.

Upon acquiring the denoised image as shown in line 4 of the algorithm 1, the next step involves identifying the ROI, precisely the location of sensitive objects within the image, using the object detection module O. In this phase, a pre-trained object detection model, such as Faster R-CNN [45], is fine-tuned by incorporating additional instances of private object classes.

Subsequently, the object detection module O scans the image and proposes regions that might contain sensitive objects and then executes a comparative analysis between the list of detected objects and the predefined list of sensitive objects S as shown in line 5 of the algorithm 1. This comparison aims to identify objects that are common to both lists. It is essential to highlight that the pre-trained module O includes annotations for all potential sensitive objects as defined in the list S. The location of sensitive objects within the image can be mathematically represented using the following equation:

$$O_{sensitive} = O(D(I_{DP})) \cap S \tag{2}$$

Once the sensitive object within the image is accurately located, the proposed method proceeds to overlay patch DP noise onto the pixels encompassing the object's bounding box as shown in lines 7 to 10 of the algorithm 1. To augment the privacy protection of the sensitive object, noise sourced from the Laplace distribution is introduced, distinct from the normal Gaussian distribution used earlier. The essential characteristic of the Laplace distribution is its heavier tails, which implies a higher likelihood of generating noise with extreme values. Moreover, we generate the differentially private noised patch with a smaller privacy parameter ε in the range from 0.01 ∼ 1, calibrated according to the $Lap(\frac{\Delta f}{\varepsilon})$, to guarantee larger noise distribution added to the bounding box of the sensitive object. Consequently, this makes it more challenging to estimate and denoise the region affected by this noise accurately.

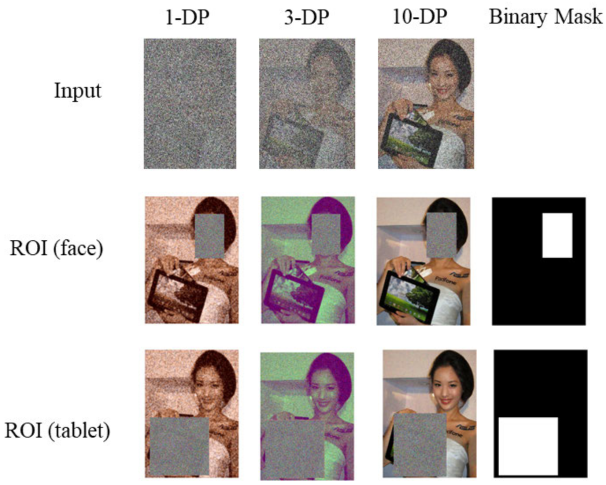


FIGURE 5. Privacy-preserving image generation examples from the $\epsilon = 1$, 3, and 10-differentially private images given different ROI of sensitive objects.

The process of generating privacy-preserving image I_{PP} can be summarized into the following equation:

$$I_{PP} = D(I_{DP}) + DP(O_{sensitive}) \tag{3}$$

In this process phase, an additional component is generated: the supplementary binary mask BM as shown in lines 8-10 of the algorithm 1. This mask explicitly indicates the areas where the DP noise patch has been applied to obscure the sensitive object. In particular, it sets '1' if the pixel is part of the area that corresponds to the non-sensitive part of the image, and '0' if the pixel corresponds to the area where differentially private noise was applied, thus protecting the sensitive part of the image.

Crucially, generating BM is deterministic, ensuring consistent and reproducible application across different instances. However, it is essential to note that BM is neither shared externally nor utilized for purposes other than its intended function within our system. The primary aim of creating BM is to assist in the seamless integration with a specialized deep learning module, which will be elaborated upon in subsequent sections of this paper.

It is important to note that in cases where the image does not contain any sensitive objects, the final output will solely result from the initial denoising process. Consequently, the

Algorithm 2 Multi-Input Feature Extraction and Classification

```
1:  function ConvSE_Block($x$)
2:      $x_{\text{conv}} \leftarrow \text{Conv}(x)$
3:      $x_{\text{SE}} \leftarrow \text{SE\_Block}(x_{\text{conv}})$
4:      return $x_{\text{SE}}$
5:  end function
6:  procedure Custom Model($I_{DP}, I_{PP}, BM$)    ▷ Extract attention-enhanced feature maps
7:      $FM_{PP} \leftarrow \text{ConvSE\_Block}(I_{PP})$
8:      $FM_{DP} \leftarrow \text{ConvSE\_Block}(I_{DP})$
9:      $FM_{PP_{\text{mask}}} \leftarrow FM_{PP} \odot BM$
10:     $FM_{DP_{\text{mask}}} \leftarrow FM_{DP} \odot BM$
11:     $FM_c \leftarrow \alpha \cdot FM_{PP_{\text{mask}}} \oplus \beta \cdot FM_{DP_{\text{mask}}}$
12:     $y \leftarrow \text{FC}(FM_c)$
13:     return $y$
14: end procedure
```

generated binary mask will not carry relevant information under these circumstances.

D. CUSTOM MODEL

1) NOTATIONS

Let us denote the feature extractor module as *ConvSE_Block*, and the extracted feature maps from I_{PP} and I_{DP} as FM_{PP} and FM_{DP}, respectively. The masked feature map is denoted as FM_{mask}, and the concatenated feature map is denoted as FM_c.

2) DESCRIPTION

In the proposed method, both I_{PP} and noisy I_{DP} are passed through the shared feature extractor base, where each convolutional block is followed by squeeze-and-excitation (SE) blocks [46] as shown in the lines 1-4 in the algorithm 2. In this context, the SE blocks function as channel-wise attention mechanisms, recalibrating the extracted feature representation after each block. They emphasize the most relevant features while suppressing the less relevant ones. Consequently, we obtain two attention-enhanced feature maps FM_{PP} and FM_{DP} containing the most discriminative features from both input sources as shown in lines 7 and 8 in the algorithm 2.

$$FM_{\text{mask}} = FM \odot BM \tag{4}$$

$$FM_c = \alpha \cdot FM_{PP_{\text{mask}}} \oplus \beta \cdot FM_{DP_{\text{mask}}} \tag{5}$$

The binary mask BM is generated alongside the I_{PP} image during the privacy-preserving image generation process. This mask indicates the area of the DP noise patch that conceals sensitive objects within the image, consisting of only 1s and 0s, where '1' in the mask represents pixels that are to be retained, and '0' represents the pixels that were obscured by differentially private patch noise.

By combining the binary mask BM with each of the attention-enhanced feature maps, we can obtain the masked $FM_{PP_{mask}}$ and $FM_{DP_{mask}}$, which effectively exclude the noisy DP patch from the process as shown in the equation 4 and lines 9 and 10 in the algorithm 2. Specifically, this procedure involves the element-wise multiplication of each feature in the feature map with its corresponding element in the binary mask. This operation preserves the features in FM that align with a 1 in BM and nullifies the features that align with a 0 in BM.

Subsequently, feature fusion is performed through a weighted concatenation of the masked attention-enhanced feature maps as shown in the equation 5, where the α and β represent the weights as shown in line 11 of the algorithm. The values of α and β are determined through empirical testing. It is important to mention, that less weight β is assigned to features extracted from the noisy input I_{DP} since the noisy input has too large pixel values distortion. However, if the generated I_{PP} lost most of the relevant information during the denoising process, an interesting synergy arises from the confluence of both feature maps. Even though the I_{DP} image was heavily perturbed by DP noise, the essence of the original information is still retained within this image, effectively compensating for information gaps that may have arisen during the denoising phase. Next, the concatenated feature map FM_c is fed to the model's further classification layers, which are defined according to the final task, number of classes, and other information.

Fig. 6 shows the operational example of the proposed custom model that takes the differentially private noisy image, generated privacy-preserving image as input. After extracting the most discriminative features from both images, the proposed model utilizes the binary mask to generate and concatenate the masked feature maps of both inputs. Such a method diverts the attention of the model from the sensitive object concealed by the DP patch and guarantees robust classification of the rest of the image.

V. PRIVACY DISCUSSION

The DP Patch framework proposed in this study is meticulously designed to safeguard the privacy of sensitive objects within shared and analyzed images. The framework's protection operates at two primary levels:

- Initial transmission security: The initial layer of protection is provided by utilizing differentially private images for network transmission. This approach ensures that even if an adversary intercepts an image instance, it remains uninterpretable to both human and machine adversaries, as demonstrated in Fig. 7 when the adversary attempts to eavesdrop on the transmission. The proposed method serves as a complementary safeguard alongside the existing communication-level security protocols, which are enabled by default to provide protection in cases of protocol malfunction, as detailed in section III.
- Protection against deep learning model attacks: The second layer of protection guards against privacy attacks targeting deep learning models, which aim to reconstruct images containing sensitive information from the

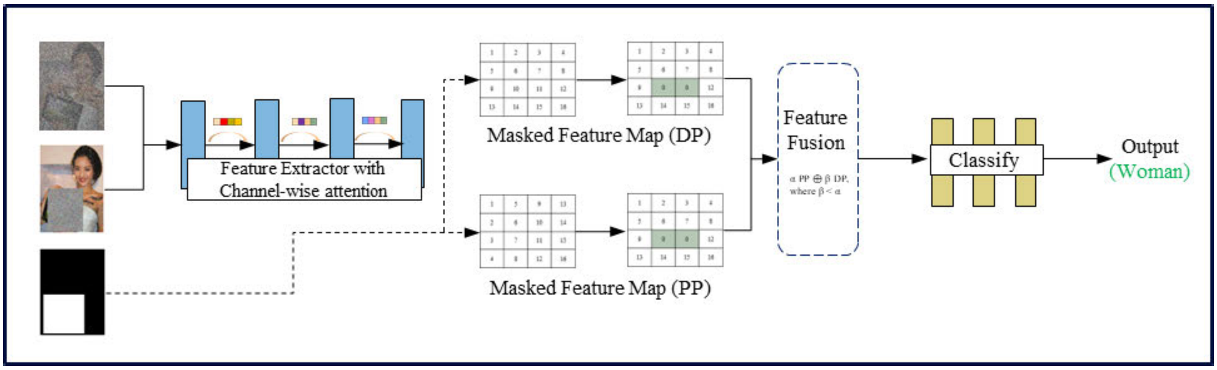


FIGURE 6. Operational example of the proposed custom model that takes privacy-preserving, differentially private, and binary mask as an input.

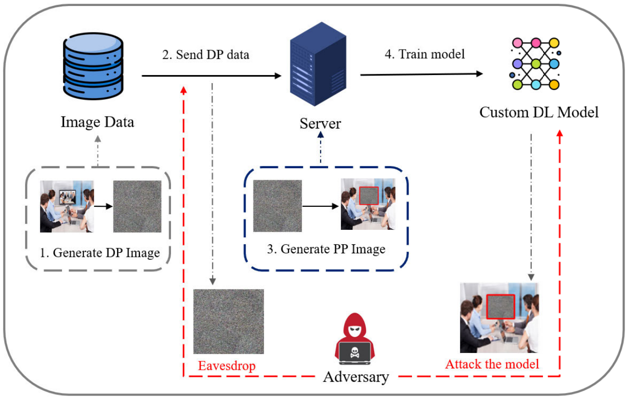


FIGURE 7. Example of privacy exposure prevention with the proposed DP Patch framework.

training set. Our methodology focuses on concealing only the sensitive objects within an image, if present, while leaving the non-sensitive parts denoised. A custom-designed deep learning model facilitates this ROI-based concealment. The model is trained to disregard the DP noise patch that masks sensitive objects, thus enhancing the utility of the images. This architecture ensures that the model does not memorize the sensitive objects, as its attention is deliberately diverted away from them. Consequently, in the event of a successful model inversion attack, any reconstructed image would lack the sensitive object, given its absence during the training phase, as shown in Fig. 7, when the adversary attempts to attack the trained model.

An adversary might intercept the noised differentially private images during transmission and attempt to denoise them using existing denoising models or by developing their own denoiser. However, within the proposed DP patch framework, the architecture of the Gaussian denoiser has been altered, rendering it a black box to the adversary. Furthermore, even if the attacker manages to reverse-engineer the model architecture, achieving successful denoising would be challenging. In the proposed method, the denoiser model was trained on a vast dataset of noised and denoised pairs to familiarize it with various noise distributions corresponding to different privacy parameters ε values. Consequently, the likelihood of an adversary accurately deducing the exact ε value and mastering the noise distribution is considerably low.

VI. EVALUATION

The proposed DP Patch framework is evaluated by answering the following research questions:

- RQ1: How effective are generated ROI-based privacy-preserving images in hiding sensitive objects?
- RQ2: How does the proposed custom model improve the classification of the generated privacy-preserving images?
- R3: How does the ablation of certain components affect the accuracy of the proposed framework?
- RQ4: How good is the performance of the proposed framework compared to the state-of-the-art method?
- RQ5: How effective is the proposed framework in defending against image reconstruction attacks?

1) EXPERIMENTAL SETUP

All implementation and experiments were performed on the environment with Windows 10, AMD Ryzen 5 3600 6-Core Processor, 16 Gb RAM, NVIDIA(R) GeForce RTX 2080 Ti GPU, and Python 3.8.

2) IMPLEMENTATION

This study uses the Open Images Dataset V4 [47], a huge dataset of nearly 9 million images for various computer vision tasks such as classification, object detection, and instance segmentation. This dataset is sourced from the image hosting service and contains over 15 million objects from 600 class instances. Even though the images with PII are removed, there is still a high possibility of finding non-trivial sensitive objects.

For our experiments, we meticulously selected a subset of the dataset comprising 500,000 images across 100 categories deemed potentially sensitive (such as mobile phones,

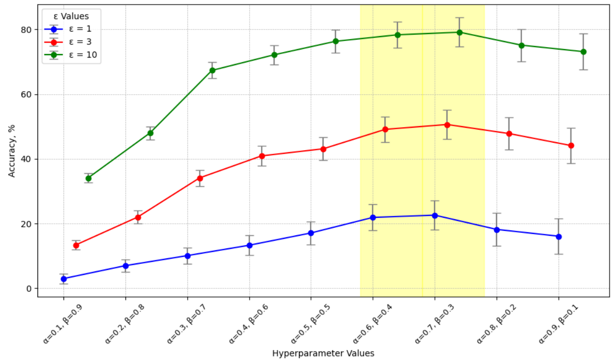


FIGURE 8. Selection of the optimal weight values.

computer screens, and envelopes). This sampling was conducted randomly to ensure representativeness.

To generate differentially private images, we perturb each pixel in the image with Gaussian noise calibrated with the following parameters: privacy parameter $\epsilon = [1,3,10]$, and relaxation parameter $\delta = [10^{-6}, 10^{-5}, 10^{-4}]$ to evaluate the different levels of image distortion.

The implementation of the proposed DP Patch framework is divided into (1) a privacy-preserving image generation module and (2) a customized deep learning module.

In the case of the first module, the denoising and object detection pipeline is implemented with the DnCNN [32] as a denoiser network and Faster R-CNN [45] with inception as an object detection network. Here, both networks are trained to suit the dataset to be used, the Gaussian noise distribution of different levels, and a rich set of annotations that might be listed as sensitive objects. To hide the sensitive object, we perturb the pixels within the bounding box with DP noise from Laplace distribution, calibrated by privacy parameter values $\epsilon = [0.1 \sim 1]$.

In the case of the second module, the multi-class image classification model is implemented. The feature extractor module consists of multiple convolutional blocks followed by a channel-wise attention mechanism implemented with SE blocks [46]. Next, the obtained feature maps are combined with a binary mask, so noisy patches and features are set to 0. The feature fusion module is implemented as a layer of weighted concatenation, where the optimal weights α and β were experimentally determined, as shown in figure 8. Intuitively, α and β should not be selected to extremes, considering that I_{DP} image may not be highly representative on their own. Hence, the weight values are chosen to best complement the feature representation of the I_{PP} image, which experimentally happened to be in the margin of 0.6 - 0.4 and 0.7 - 0.3, respectively.

The time complexity of the proposed framework is denoted as $O(H \times W)$, depending on the height and width of the image or the total number of pixels. Assume that each part of the framework is implemented and optimized and that all operations scale linearly with the number of pixels in the image.

TABLE 2. Average ROI SSIM score between generated I_{PP} and corresponding original image.

Profile	Image type	ROI SSIM
Case A	I_{PP} ($\epsilon = 10$)	86%
	I_{PP} ($\epsilon = 3$)	65%
	I_{PP} ($\epsilon = 1$)	53%
Case B	I_{PP} ($\epsilon = 1$)	57%
	I_{PP} ($\epsilon = 3$)	63%
	I_{PP} ($\epsilon = 10$)	89%

TABLE 3. Average accuracy & confidence score of the sensitive object detection task.

Profile	Image type	Accuracy	Conf. Score
Case A	Original	100%	79%
	I_{PP} ($\epsilon = 10$)	93%	67%
	I_{PP} ($\epsilon = 3$)	74%	60%
	I_{PP} ($\epsilon = 1$)	64%	57%
Case B	Original	100%	74%
	I_{PP} ($\epsilon = 10$)	89%	69%
	I_{PP} ($\epsilon = 3$)	70%	56%
	I_{PP} ($\epsilon = 1$)	59%	50%

A. RQ1. HOW EFFECTIVE ARE GENERATED PRIVACY-PRESERVING IMAGES IN HIDING SENSITIVE OBJECTS DEFINED IN EACH CORRESPONDING SSI PROFILE?

This section examines the efficacy of privacy-preserving images generated from differentially private inputs with ϵ values of 1, 3, and 10 in concealing sensitive objects defined as ROI targets. We assess the structural similarity between the generated I_{PP} images and the original images I_{orig}, then evaluate the sensitive object detection accuracy.

Our experiments were executed under two cases of regions of interest (sensitive objects):

- Case A: ROI is defined as information about individuals, such as human faces, persons, names, and vehicle registration plates.
- Case B: ROI is defined as objects with digital information, e.g., tablet screens, computer screens, computer monitors, and mobile phones.

The structural similarity index metric (SSIM) is conventionally employed to evaluate the perceptual quality of the image in relation to its original counterpart. However, straightforward evaluation of the I_{PP} and I_{orig} images is not possible due to the DP noise patch present in the arbitrary I_{PP}.

To effectively evaluate the structural similarity between the two images, we measure the SSIM within the region of interest, which is the image without the noise patch. Since the supplementary binary mask is generated during the privacy-preserving image generation step, it can be also used to perform an effective ROI SSIM calculation.

Table 2 presents the average ROI SSIM score for the privacy-preserving images I_{PP} generated from the $\epsilon = 1$, 3, and 10 - differentially private images and comparing them with the original images. The average ROI SSIM score for the ROI case A was 53%, 65%, and 86%, while for ROI case B was 57%, 63%, and 89% for the

TABLE 4. Performance comparison of multiple computer vision models (Open Images [47] dataset).

Profile	Image Type	Type	VGG16	ResNet50	InceptionV3	Proposed
N/A	Original	Val	98.5%	97.8%	98.9%	N/A
		Test	97.4%	93.1%	95.0%	
Case A	I_{PP} ($\epsilon = 10$)	Val	54.9%	41.3%	49.6%	**75.4%**
		Test	35.4%	37.1%	42.7%	**69.3%**
	I_{PP} ($\epsilon = 3$)	Val	15.6%	16.9%	21.1%	**46.3%**
		Test	11.2%	13.4%	17.3%	**49.1%**
	I_{PP} ($\epsilon = 1$)	Val	8.9%	7.5%	12.0%	**19.4%**
		Test	10.5%	10.1%	11.1%	**22.5%**
Case B	I_{PP} ($\epsilon = 10$)	Val	65.1%	48.1%	55.8%	**79.1%**
		Test	49.2%	50.6%	44.1%	**74.9%**
	I_{PP} ($\epsilon = 3$)	Val	20.1%	18.3%	14.5%	**50.6%**
		Test	14.7%	12.6%	7.10%	**47.1%**
	I_{PP} ($\epsilon = 1$)	Val	8.98%	5.45%	10.5%	**22.6%**
		Test	10.1%	5.05%	6.00%	**25.1%**

ϵ values of 1, 3, and 10, respectively. These scores indicate that the generated privacy-preserving images maintain high structural similarity to the original images, making them well-suited for image recognition tasks while effectively obfuscating sensitive objects. To evaluate how well the proposed method can localize sensitive objects within the image, the performance of the object detection module on the generated privacy-preserving image is evaluated and compared it to the original images. Table 3 shows the results for the sensitive object detection task for the original images and privacy-preserving images I_{PP} generated from the $\epsilon =$ 1, 3, and 10 - differentially private images and compared to the detection rate of original images. For ROI case A, the accuracy of the sensitive object detection task was 64%, 74%, 93%, and 100% for ϵ values of 1, 3, 10, and the original images, respectively. The corresponding average confidence scores were 57%, 60%, 67%, and 79%. For ROI case B, the accuracy was 59%, 70%, 89%, and 100%, while the average confidence scores were 50%, 56%, 69%, and 74% for ϵ values of 1, 3, 10, and the original images, respectively.

These results show that both the accuracy and average confidence score of sensitive object detection diminishes as the ϵ value decreases. This decline can be attributed to the increased distortion of the input image. The accuracy in ROI case A is slightly higher than in case B. This difference can be attributed to the nature of the objects in case B, which predominantly include smaller items like mobile phones, making them more susceptible to the effects of differential privacy noise addition and subsequent denoising processes.

> ***Answer to RQ1:*** The generated privacy-preserving images I_{PP} show sufficiently high ROI structural similarity and acceptable rate of sensitive object detection compared to their corresponding original image counterparts.

B. RQ2. HOW DOES THE PROPOSED CUSTOM MODEL IMPROVE THE CLASSIFICATION OF THE GENERATED PRIVACY-PRESERVING IMAGES?

In this section, the efficacy of the proposed custom model for privacy-preserving image processing tasks is explored by contrasting its performance against that of a standard model. Specifically, comparisons are made between the proposed model and state-of-the-art image classification models based on VGG16, ResNet, and InceptionV3 architectures, using various input images such as I_{PP} generated from differentially private inputs with ϵ values of 1, 3, and 10, as well as original images.

Table 4 shows the comparative results of model performance under consistent parameters, including a batch size of 32, 100 epochs, and a dataset split into training, validation, and test sets in a 70:15:15 ratio. It is worth noting that during this experiment, privacy-preserving images with a minor DP noise patch were assessed, enabling the model to evaluate the remaining portions of the image. Additionally, experiments within this section were conducted across two distinct cases of sensitive objects considered ROI, as introduced in section VI-A.

As observed from the table, the performance of the three state-of-the-art models is strong when provided with original input images without any visual distortion. For instance, the VGG16 model achieved a validation accuracy of up to 98.5%, and a test accuracy of up to 97.4%. Similarly, the ResNet50 model achieved validation and test accuracy metrics of up to 97.8% and 93.1% respectively. Meanwhile, the InceptionV3 model achieved a validation accuracy of up to 98.9% and a test accuracy of up to 95.0%.

Considering the ROI case A, designed to conceal human-related information, the performance of the VGG16 model on the privacy-preserving images generated from the $\epsilon =$ 10 - differentially private input decreased, resulting in the validation accuracy of 54.9%, and the test accuracy of 35.4%. Similarly, the validation accuracy of the ResNet50 and the InceptionV3 model was 41.3% and 49.6%, while the test accuracy was by as much as 37.1% and 42.7%, respectively. For the I_{PP} images generated from the ϵ values equal to 3 and 1, the validation accuracy of the VGG16 model was by as much as 15.6% and 8.9%, and the test accuracy was by as much as 11.2% and 10.5%, respectively. For the ResNet50 model, the validation accuracy was by as much as 16.9% and 7.5%, while the test accuracy was by as much as 13.4% and 10.1% for privacy parameter ϵ values $=$ 3 and 1. InceptionV3 model also showed poor performance

TABLE 5. Performance comparison of multiple computer vision models (ImageNet [48] dataset).

Profile	Image Type	VGG16	ResNet50	InceptionV3	Proposed
N/A	Original	92.7%	86.0%	79.5%	N/A
Case A	$I_{PP}(\epsilon = 10)$	54.3%	44.0%	39.8%	**73.1%**
	$I_{PP}(\epsilon = 3)$	33.5%	27.2%	20.4%	**47.6%**
	$I_{PP}(\epsilon = 1)$	11.4%	9.7%	10.4%	**25.0%**

for the privacy-preserving images generated from $\epsilon = 1$ and 3 images, such as 21.1% and 12.0% for the validation accuracy, and 17.3%, 11.1% for the test accuracy.

The ROI case B also exhibited a similar trend of the model validation/test accuracy degradation as the ϵ value decreases, with the occasional instances of improved performance, such as the validation accuracy of the VGG16 model with the I_{PP} (ϵ=10) of 65.1%, or worsened performance, such as the test accuracy of the InceptionV3 model with the I_{PP} ($\epsilon = 3$) of 7.4%.

Different from the aforementioned computer vision models that consider the privacy-preserving image as a whole, incorporating the DP noise patch, and suffering from the large input image quality degradation, the proposed custom model is designed explicitly for privacy-preserving image classification. Therefore, the are no experiment results corresponding to the original images. As can be observed from the table 4, the validation and test accuracy of the proposed model in the ROI case A was by as much as 75.4% and 69.3% for $\epsilon = 10$, 46.3% and 49.1% for $\epsilon = 3$, and 19.4% and 22.5 % for $\epsilon = 1$. In the ROI case B, the validation and test accuracy were 79.1% and 79.4% for $\epsilon = 10$, 50.6% and 47.1% for $\epsilon = 3$, and 22.6% and 25.1 % for $\epsilon = 1$.

Table 5 shows the test accuracy of the existing image classification models and the proposed method under the ImageNet [48] dataset, generated in a similar manner under ROI case A. The performance of the DP Patch was by as much as 73.1%, 47.6%, and 25.0% for $\varepsilon = 10, 3$, and 1, respectively.

The proposed method shows an average performance improvement of approximately 24.8%, 34.3%, and 16.17% for $\varepsilon = 10$, 3, and 1, respectively, compared to other models for Open Images dataset, and 26.0%, 20.5%, and 14.5% for $\varepsilon = 10, 3$, and 1, respectively for ImageNet dataset. Currently, the performance degradation problem linked to differential privacy remains an unavoidable trade-off for privacy guarantees. Consequently, the proposed custom model's performance follows the same trend as regular state-of-the-art models while yielding notably improved outcomes.

> ***Answer to RQ2:*** The proposed custom model shows satisfactory performance in multi-class classification of the generated privacy-preserving images given various privacy parameter ϵ values.

C. RQ3: HOW DOES THE ABLATION OF CERTAIN COMPONENTS AFFECT THE ACCURACY OF THE PROPOSED FRAMEWORK?

The ablation study was conducted to demonstrate the effects of the binary mask BM and the differentially private image I_{DP} on the performance of the proposed custom model. Table 6 summarizes the accuracy of the custom model with one of the components ablated, comparing it to the baseline accuracies of 25.1%, 47.1%, and 74.9% for $\varepsilon = 1$, 3, and 10, respectively.

TABLE 6. Accuracy of the model with ablated components.

Component	Accuracy		
	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 10$
BM	11.0%	13.6%	24.8%
I_{DP}	24.0%	43.4%	67.8%
N/A	25.1%	47.1%	74.9%

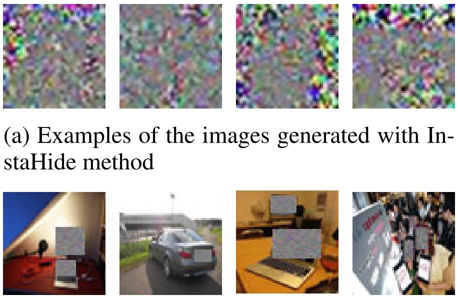
The first row corresponds to the ablation of the binary mask BM and its combination with attention-enhanced feature maps. The accuracy of the model was 11.0%, 13.6%, and 24.8% for $\varepsilon = 1, 3$, and 10. Since the primary goal is to exclude features corresponding to sensitive objects concealed by the noisy DP patch, removing this component of the framework forces the model to classify based on the weighted combination of I_{PP} and I_{DP} alone, including the noisy patches if present within the I_{PP} image.

The second row of Table 6 corresponds to the ablation of the I_{DP} input image from the custom model. The accuracy of the model was 24.0%, 43.4%, and 67.8% for $\varepsilon = 1, 3$, and 10. The primary goal of the noisy I_{DP} is to complement the I_{PP} image, especially if the denoising process results in significant information loss. However, this process is also highly dependent on the amount of random noise applied and the outcome of the denoising process. When this component of the framework is removed, the model attempts to classify the I_{PP} with the noisy patch excluded.

> ***Answer to RQ3:*** The binary mask component has a greater impact on the performance of the proposed custom model compared to the complementary power of I_{DP}.

D. RQ4: HOW GOOD IS THE PERFORMANCE OF THE PROPOSED FRAMEWORK COMPARED TO THE STATE-OF-THE-ART METHOD?

This section compares the proposed DP Patch framework with the state-of-the-art method called InstaHide [3]. The primary concept behind InstaHide is to randomly mix an image's pixels with those of other images. For this experiment, the ‘‘inside-dataset'' variant of InstaHide was employed, which intermixes the training data images with pixels from other images within the same dataset.



(a) Examples of the images generated with InstaHide method

(b) Examples of the images generated with the proposed DP Patch method

FIGURE 9. Comparision of the (a) InstaHide and (b) proposed DP Patch methods.

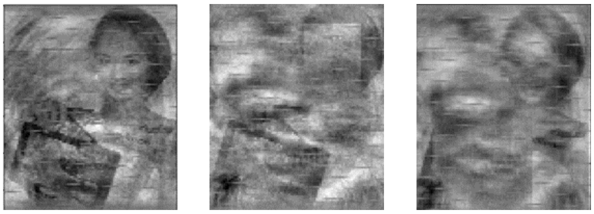
Fig. 9 shows the generated images from the InstaHide and proposed DP Patch framework. As observable from fig. 9a, encrypted images are not recognizable to the human eyes as a result of the pixel mixing. However, the regular VGG16 model trained on that data shows robust classification results of 87.1%. Fig 9b shows the privacy-preserving images generated by the proposed framework from the $\epsilon = 10$ differentially private input, where only arbitrary sensitive objects were concealed with the patch of heavy $\epsilon = 0.1$ noise drawn from the Laplace distribution. At the same time, the accuracy of the custom model was about $75 \sim 80\%$.

However, the primary challenge in evaluating both methods stems from their distinct approaches. While InstaHide treats the entirety of an image as private and randomly mixes pixels of that image with those of other images, the DP Patch approach deems only specific objects as private.

> ***Answer to RQ4:*** While the proposed framework's distinct approach makes direct comparison with other methods challenging, DP Patch still showcases commendable performance relative to state-of-the-art solutions.

E. RQ5: HOW EFFECTIVE IS THE PROPOSED FRAMEWORK IN DEFENDING AGAINST IMAGE RECONSTRUCTION ATTACKS?

As a practical evaluation method, a model inversion attack is performed that tries to recover the instance sample from the training set by reverse engineering using the model's output. In this experiment, a simplified version of the model inversion attack introduced by [2] is implemented, which considers the confidence score produced alongside the model prediction. The higher the confidence score of the model when predicting specific samples, the more likely it is that it has seen that sample before during the training phase. Specifically, the reconstructed image is created by iteratively modifying an input image (random noise) to minimize a loss function associated with the model's prediction until the image resembles the targeted class.



(a) Original image (b) I_{PP} ($\epsilon = 10$) (c) I_{PP} ($\epsilon = 10$)

FIGURE 10. Example of the model inversion attack on the (a) original image; (b) generated I_{PP} in ROI case A that conceals human face; (c) generated I_{PP} in ROI case B that conceals tablet screen.

Fig. 10 shows the example of a model inversion attack on the model to reconstruct the woman with a tablet image sample. In the case of the original image without any visual distortions, the reconstructed image is shown in Fig. 10a, where the woman's face and the tablet screen were reconstructed successfully. Fig. 10b and 10c show the reconstructed images for the I_{PP} images generated from $\epsilon = 10$-differentially private input under ROI cases A and B. In our experiments, ROI case A tries to conceal all human-related information, such as face, and ROI case B tries to conceal all digital-related information, such as tablet screen. As observable from the result, the sensitive information was not reconstructed in either case, even though the rest of the image could be successfully reconstructed. Since the noisy patch is excluded from the training process of the custom model, the model inversion attack is likely to reconstruct only the non-sensitive part of the image but omit the sensitive object hidden under the DP noise patch.

> ***Answer to RQ5:*** The proposed method shows satisfactory results in defending against image reconstruction attacks, preventing the concealed sensitive objects from recovery.

VII. CONCLUSION

This paper introduces a novel framework, DP Patch, designed to facilitate the secure transmission of image data over networks and protect sensitive objects within images from being recovered through model attacks while maintaining adequate performance in computer vision tasks. The proposed DP Patch framework encompasses two key modules: (1) the generation of privacy-preserving images from differentially private input, which effectively conceals sensitive objects within the image based on the ROI of the sensitive object, and (2) a custom model tailored for privacy-preserving images, which processes both noisy differentially private images and privacy-preserving images to ensure robust classification. Furthermore, an experimental evaluation was conducted to assess the quality of the privacy-preserving images by comparing the performance of the custom model with state-of-the-art models. Specifically, the proposed method demonstrates an average performance improvement of approximately 24.8%, 34.3%, and 16.17% for $\epsilon = 10$, 3, and 1, respectively, over other models. Additionally, the proposed framework effectively counters image reconstruction attacks. As we look towards future research directions, several avenues emerge from our work, including adapting the proposed framework to specific domains, such as medical imaging, incorporating alternative DP image generation techniques, strengthening the proposed method for more complex computer vision tasks, and considering it for real-world deployment.

REFERENCES

[1] M. Heikkilä, "The walls are closing in on clearview AI," MIT Technol. Rev., Cambridge, MA, USA, 2022. Accessed: May 24, 2022. [Online]. Available: https://www.technologyreview.com/2022/05/24/1052653/clearview-ai-data-privacy-uk/

[2] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.

[3] Y. Huang, Z. Song, K. Li, and S. Arora, "InstaHide: Instance-hiding schemes for private distributed learning," in *Proc. Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 4507–4518.

[4] T. Li and M. S. Choi, "DeepBlur: A simple and effective method for natural image obfuscation," 2021, *arXiv:2104.02655*.

[5] Z. Wan, Y. Zhang, and H. He, "Variational autoencoder based synthetic data generation for imbalanced learning," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–7.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–12.

[7] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233*.

[8] Y. Wu, F. Yang, and H. Ling, "Privacy-protective-GAN for face de-identification," 2018, *arXiv:1806.08906*.

[9] W. Sirichotedumrong and H. Kiya, "A GAN-based image transformation scheme for privacy-preserving deep neural networks," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 745–749.

[10] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, *arXiv:1802.06739*.

[11] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren, "GANobfuscator: Mitigating information leakage under GAN via differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2358–2371, Sep. 2019.

[12] K. Liu, B. Tan, and S. Garg, "Subverting privacy-preserving GANs: Hiding secrets in sanitized images," in *Proc. AAAI*, vol. 35, May 2021, pp. 14849–14856.

[13] H. Xue, B. Liu, M. Ding, T. Zhu, D. Ye, L. Song, and W. Zhou, "DP-image: Differential privacy for image data in feature space," 2021, *arXiv:2103.07073*.

[14] L. Fan, "Image pixelization with differential privacy," in *Data and Applications Security and Privacy XXXII*, F. Kerschbaum and S. Paraboschi, Eds. Cham, Switzerland: Springer, 2018, pp. 148–162.

[15] L. Fan, "Practical image obfuscation with provable privacy," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 784–789.

[16] N. Carlini, S. Deng, S. Garg, S. Jha, S. Mahloujifar, M. Mahmoody, S. Song, A. Thakurta, and F. Tramer, "Is private learning possible with instance encoding?" 2021, *arXiv:2011.05315*.

[17] H. Chi and Y. H. Hu, "Face de-identification using facial identity preserving features," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2015, pp. 586–590.

[18] K. Yang, J. Yau, L. Fei-Fei, J. Deng, and O. Russakovsky, "A study of face obfuscation in ImageNet," 2022, *arXiv:2103.06191*.

[19] Y. Long, B. Wang, Z. Yang, B. Kailkhura, A. Zhang, C. Gunter, and B. Li, "G-pate: Scalable differentially private data generator via private aggregation of teacher discriminators," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2965–2977.

[20] J.-W. Chen, C.-M. Yu, C.-C. Kao, T.-W. Pang, and C.-S. Lu, "DPGEN: Differentially private generative energy-guided network for natural image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8377–8386.

[21] F. Harder, K. Adamczewski, and M. Park, "DP-MERF: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 1819–1827.

[22] S. P. Liew, T. Takahashi, and M. Ueno, "PEARL: Data synthesis via private embeddings and adversarial reconstruction learning," 2021, *arXiv:2106.04590*.

[23] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2019, pp. 565–578.

[24] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: Conditional identity anonymization generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5446–5455.

[25] Y. Wen, B. Liu, M. Ding, R. Xie, and L. Song, "IdentityDP: Differential private identification protection for face images," *Neurocomputing*, vol. 501, pp. 197–211, Aug. 2022.

[26] D.-H. Ko, S.-H. Choi, J.-M. Shin, P. Liu, and Y.-H. Choi, "Structural image de-identification for privacy-preserving deep learning," *IEEE Access*, vol. 8, pp. 119848–119862, 2020.

[27] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "IPrivacy: Image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 5, pp. 1005–1016, May 2017.

[28] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang., Program.* Venice, Italy: Springer, 2006, pp. 1–12.

[29] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 308–318.

[30] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.

[31] T. Dockhorn, T. Cao, A. Vahdat, and K. Kreis, "Differentially private diffusion models," 2022, *arXiv:2210.09929*.

[32] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[33] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[34] M. Momeny, A. M. Latif, M. A. Sarram, R. Sheikhpour, and Y. D. Zhang, "A noise robust convolutional neural network for image classification," *Results Eng.*, vol. 10, Jun. 2021, Art. no. 100225.

[35] X. Meng, Y. Yan, S. Chen, and H. Wang, "A cascaded noise-robust deep CNN for face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3487–3491.

[36] Y. Ding, Y. Cheng, X. Cheng, B. Li, X. You, and X. Yuan, "Noise-resistant network: A deep-learning method for face recognition under noise," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 1–14, Dec. 2017.

[37] J. Regan and M. Khodayar, "A triplet graph convolutional network with attention and similarity-driven dictionary learning for remote sensing image retrieval," *Exp. Syst. Appl.*, vol. 232, Dec. 2023, Art. no. 120579.

[38] M. Saffari, M. Khodayar, and S. M. J. Jalali, "Sparse adversarial unsupervised domain adaptation with deep dictionary learning for traffic scene classification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 4, pp. 1139–1150, Jul. 2023.

[39] M. S. Qadir and G. Bilgin, "Active learning with Bayesian CNN using the BALD method for hyperspectral image classification," *Mesopotamian J. Big Data*, vol. 2023, pp. 55–62, Jun. 2023.

[40] F. Bozkurt, M. Kara, M. A. Aydın, and H. H. Balik, "Exploring the vulnerabilities and countermeasures of SSL/TLS protocols in secure data transmission over computer networks," in *Proc. IEEE 12th Int. Conf. Intell. Data Acquisition Adv. Comput. Systems: Technol. Appl. (IDAACS)*, Sep. 2023, pp. 400–407.

[41] Y. R. Siwakoti and D. B. Rawat, "Investigating security vulnerability related to exposure and TLS ecosystem in IoT devices," in *Proc. IEEE 24th Int. Conf. Inf. Reuse Integr. Data Sci. (IRI)*, Aug. 2023, pp. 7–12.

[42] A. F. Abate, A. Castiglione, L. Cimmino, D. De Angelis, S. Flauto, and A. Volpe, "On the (in)Security and weaknesses of commonly used applications on large-scale distributed systems," in *Proc. 24th Int. Conf. Control Syst. Comput. Sci. (CSCS)*, France, May 2023, pp. 572–579.

[43] A. Mallik, "Man-in-the-middle-attack: Understanding in simple words," *Cyberspace: Jurnal Pendidikan Teknologi Informasi*, vol. 2, no. 2, p. 109, Jan. 2019.

[44] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 250–258.

[45] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–12.

[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[47] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy, "OpenImages: A public dataset for large-scale multi-label and multi-class image classification," Dataset, 2017. [Online]. Available: https://github.com/openimages

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

ASSEM UTALIYEVA (Student Member, IEEE) received the B.S. and M.S. degrees from Pusan National University, Busan, South Korea, in 2020 and 2022, respectively, where she is currently pursuing the Ph.D. degree in computer science and engineering. Her research interests include differential privacy and security for artificial intelligence.

SEON-JIN HWANG received the B.E. and M.S. degrees from Pusan National University, Busan, South Korea, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in computer science and engineering. His research interests include security for artificial intelligence, software security, and malware detection.

YOON-HO CHOI (Member, IEEE) received the M.S. and Ph.D. degrees from the School of Electrical and Computer Engineering, Seoul National University, Seoul, South Korea, in 2004 and 2008, respectively. He was a Postdoctoral Scholar with Seoul National University, in 2008, and Pennsylvania State University, University Park, PA, USA, in 2009. From 2010 to 2012, he was a Senior Engineer with Samsung Electronics. From 2012 to 2014, he was an Assistant Professor with the Department of Convergence Security, Kyonggi University, Suwon, South Korea. He is currently a Professor with the School of Computer Science and Engineering, Pusan National University, Busan, South Korea. His research interests include privacy-preserving deep learning, adversarial examples, anomaly detection, deep packet inspection for high-speed intrusion prevention, and the IoT security for realizing secure computer systems and networks. He has served as a TPC member and an editor for various international conferences and journals.