Unveiling Hidden Visual Information: A Reconstruction Attack Against Adversarial Visual Information Hiding

Jonggyu Jang, *Member, IEEE*, Hyeonsu Lyu, *Student Member, IEEE*, Seongjin Hwang, and Hyun Jong Yang, *Member, IEEE*

***Abstract*—This article investigates the security vulnerabilities of adversarial example-based image encryption by executing data reconstruction (DR) attacks on encrypted images. A representative image encryption method is the adversarial visual information hiding (AVIH), which uses type-I adversarial example training to protect gallery datasets used in image recognition tasks. In the AVIH method, the type-I adversarial example approach creates images that appear completely different but are still recognized by machines as the original ones. Additionally, the AVIH method can restore encrypted images to their original forms using a predefined private key generative model. For the best security, assigning a unique key to each image is recommended; however, storage limitations may necessitate some images sharing the same key model. This raises a crucial security question for AVIH: *How many images can safely share the same key model without being compromised by a DR attack?* To address this question, we introduce a dual-strategy DR attack against the AVIH encryption method by incorporating 1) *generative-adversarial loss* and 2) *augmented identity loss*, which prevent DR from overfitting—an issue akin to that in machine learning. Our numerical results validate this approach through image recognition and re-identification benchmarks, demonstrating that our strategy can significantly enhance the quality of reconstructed images, thereby requiring fewer key-sharing encrypted images. The source code to reproduce the results will be available in https://github.com/jonggyujang0123/Hiding_person.**

***Index Terms*—Adversarial examples, data reconstruction (DR) attack, face recognition, image recognition, person re-identification.**

Received 8 August 2024; revised 30 January 2025; accepted 19 March 2025. Date of publication 28 April 2025; date of current version 4 September 2025. This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program supervised by then Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2025-2021-0-02048; in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through Korea Government (MSIT), 6G MIMO System Research under Grant 2021-0-00161; and in part by Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through Korea Government (MSIT), Development of 5G-A vRAN Research Platform under Grant RS-2024-00404972. *(Jonggyu Jang and Hyeonsu Lyu contributed equally to this work.) (Corresponding author: Hyun Jong Yang.)*

Jonggyu Jang is with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: jang255@purdue.edu).

Hyeonsu Lyu and Seongjin Hwang are with the Department of Electrical Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, Republic of Korea (e-mail: hslyu4@postech.ac.kr; sjh1753@postech.ac.kr).

Hyun Jong Yang is with the Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea (e-mail: hjyang@snu.ac.kr).

Digital Object Identifier 10.1109/TNNLS.2025.3555248

I. Introduction

MACHINE learning has evolved from a groundbreaking innovation to a widely adopted and promising technology across numerous fields. One key characteristic of machine learning is its dependency on data; machine-learning models are trained on data and often require additional user data for their application services. Recently, this dependency on data has raised significant privacy concerns [1], [2]. In the most straightforward case, storing and processing facial or body images in public cloud services for machine-learning applications can expose these images to unauthorized access and misuse [3].

A simple solution to mitigate this risk is on-device computing, where data are processed locally on mobile devices without transmitting it to the cloud [2]. However, mobile devices often have limited computing resources and battery life, making fully local processing impractical for many real-world scenarios. Consequently, cloud-based image recognition systems are widely used, wherein a *large gallery dataset of user images* is maintained on the cloud server to facilitate fast comparisons with incoming (query) images. This setup, while efficient, also heightens the risk of privacy breaches if the unencrypted gallery dataset is leaked or accessed by malicious attackers.

A. Backgrounds

To safeguard these gallery images, several defensive methods have been proposed to counteract the security and privacy risks, including

1) Hiding visual information in noisy images [1], [4].
2) Perceptual encryption (PE) [5], [6].
3) Homomorphic encryption (HE) [7], [8].[1]

While HE guarantees strong security, it also incurs excessive computation time, making it unsuitable for real-time cloud-based systems. PE necessitates retraining the service DNN, which significantly degrades inference quality after encryption. Conversely, the *adversarial visual information hiding* (AVIH) encryption method [1] encrypts gallery images into noisy images while preserving the output of the service model. As

[1]More comprehensive literature reviews are available in Section II.

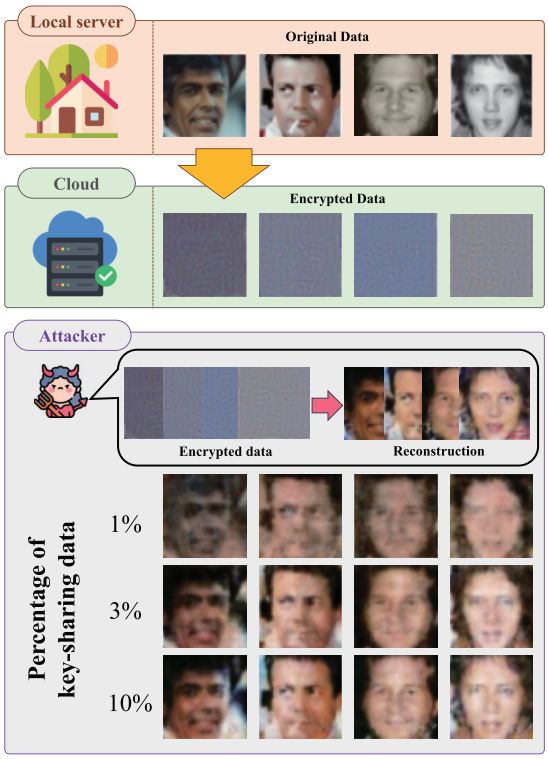

Fig. 1. Examples of the encryption method and the proposed attacker model. The local server stores the original image data. In a cloud service system, the local server offloads the computation of image recognition tasks (such as face recognition in our example) to the cloud server. Before sending raw data of the private images to the cloud server, the local server encrypts the original image data into a noisy image. The service model then processes both the original and encrypted images similarly. Our focus in this scenario is to highlight the privacy vulnerabilities of the encryption method through DR attacks.

shown in Fig. 1, the AVIH method offers practical advantages by maintaining inference speed and accuracy without requiring extensive retraining. However, despite its efficacy in preserving utility, *little work has investigated how robust AVIH truly is against adversaries attempting to reverse-engineer or reconstruct the original data.*

The concept of data reconstruction (DR) attack arises from the widespread belief that trained DNNs can retain information about their training data, with a simple example being inferring the membership of data [9]. Beyond merely inferring membership, researchers have developed DR attack methods for simple classifiers, though these reconstructed images often lack photorealism [10], [11], [12]. With the advent of generative models, new techniques have emerged to produce photorealistic images using methods such as advanced identity loss leveraging logits [13], supervised inversion [14], and adversarial examples [15], where those methods focus on finding appropriate latent vectors. In existing reconstruction attacks [9], [10], [11], [12], [13], [14], [15], [16], most of the attackers aim to recover the training data from trained neural networks; however, the existing attackers are only able to reconstruct the *images inspired by a similar style rather than recreate specific images.*

B. Motivations and Novelty

1) Research Question: In the AVIH method, assigning a unique key to each image is recommended for optimal security; however, storage limitations may necessitate some images sharing the same key model. Throughout this article, we address the following research question regarding the practical use of the AVIH method for cloud-based inference systems:

RQ:How securely can AVIH truly conceal visual information?

In the remainder of this article, we aim to solve the above research question by executing a DR attack against the AVIH method. Examples of the DR attack are depicted at the bottom of Fig. 1.

2) Novelty: Our study differentiates itself by not relying on pretrained generative model for finding latent vectors. Instead, we train an attacker key model that mimics the original key model, aiming to reconstruct images with high fidelity. The reconstruction quality improves as the number of key-sharing images increases, leveraging identity loss and adversarial training. While our approach aims for high-fidelity reconstruction, the quality of the reconstructed images is closely tied to the number of key-sharing images available during training. A larger number of key-sharing images provides stronger guidance through the augmented identity loss, resulting in finer details being preserved.

C. Our Findings

In this article, we aim to show that leveraging adversarial examples for visual information hiding [1] is *unsafe*, as depicted in the right part of Fig. 1. To this end, we propose a DR attack on the encrypted gallery set without access to the *original key model*, as depicted in Fig. 2. In the proposed method, we first randomly initialize an attacker key model. Then, we train the attacker key model to recover an image that consistently matches the service DNN output while ensuring photorealism. As a simple method, one can train the attacker key model with the identity loss between the service model outputs of encrypted data and the attacker key model's output. However, with only standard identity loss, the attacker key model's outputs are not similar to the original images. This is our main challenge in the generalization of the attacker key model, where overfitting (a concept similar to ordinary machine learning) occurs if few data share the same key model.

Our salient contributions are summarized as follows.

1) *Reconstruction Attack:* To the best of the authors' knowledge, our work is the first to demonstrate that a DR attack works in practical scenarios.[2]
2) *Resolving Overfitting 1 (Augmented Identity Loss):* To alleviate the overfitting issue in DR attacks, we propose augmented identity loss, which helps generalize the trained attacker key model.
3) *Resolving Overfitting 2 (Generative-Adversarial Loss):* In addition to augmented identity loss, we propose a GAN-based DR attack to improve the generalization and photorealism of the attacker outputs. Unlike previous studies [13], [14], [15] that use pretrained DNNs to find an appropriate latent vector, we train an auxiliary key

[2]As shown in previous studies [17], given a DNN's output features or logits, one can exactly reverse-engineer the DNN to reconstruct the original training dataset, though a strong assumption (homogeneous neural network) is required.

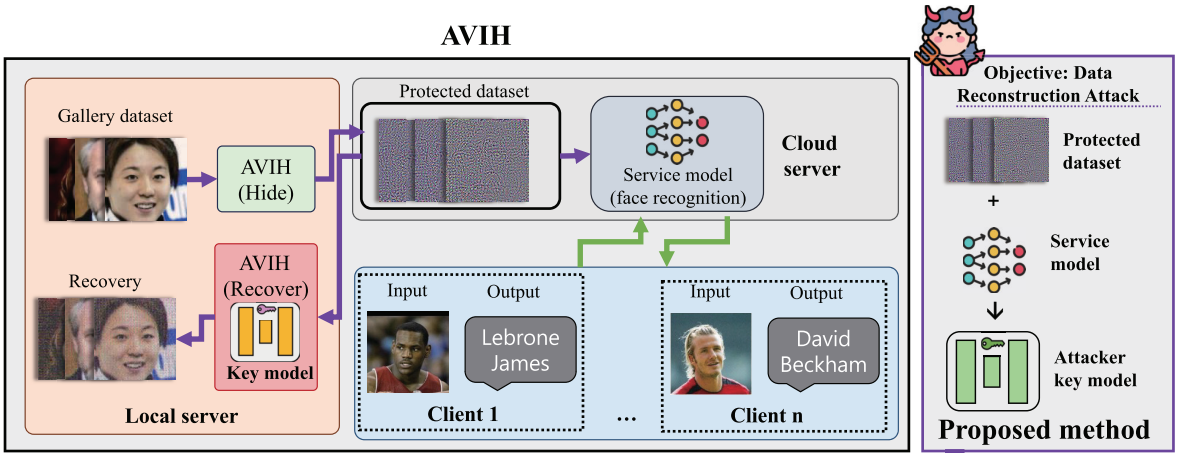

Fig. 2. Illustration of the AVIH method and the objective of our work is shown. Demonstrates that the gallery set is safeguarded and provided to the DNN in the cloud server (left). The protected image shows altered visual information that is completely different from the original, yet remains accurately identifiable. In the local server, the protected images can be recovered using its key DNN model. Illustration highlights our proposed method, which aims to design *a replica of the key model* without accessing the actual key model (right).

Fig. 3. Original images, encrypted images, reconstructed images, and our attack results. The numbers below the images refer the number of images sharing the same key model.

model that mimics the original key model while constraining local patch-level similarity with an auxiliary dataset.

4) *Vulnerabilities of the AVIH Method:* We validate that images encrypted by the AVIH method can be reconstructed by the proposed method for various tasks such as face recognition, human re-identification, and vehicle identification. For example, Fig. 3 shows the results of the proposed method by changing the number of key-sharing images for a face recognition model. Furthermore, our ablation studies show that the proposed method can enhance the quality of DR, highlighting the necessity of stronger privacy-preserving methods.

D. Organization

The remaining parts of this article are organized as follows. In Section II, we provide comprehensive reviews of existing visual information hiding methods and corresponding security and privacy attacks. Section III introduces the details of the AVIH encryption method proposed in [1]. Section IV details the proposed approach for DR against the AVIH method. Next, in Section V and Appendix A, we present our experimental setup and results for face recognition scenarios and re-identification scenarios, respectively. Finally, Section VI concludes the article with a discussion on the conclusion, limitations, extensibility, and future research directions.

II. Related Works on Hiding Visual Information

Several studies have focused on hiding visual information in machine-learning tasks, particularly during the inference stage.

A. Homomorphic Encryption

HE is a cryptographic technique that allows computations to be performed on encrypted data, maintaining privacy while still producing an encrypted result that, when decrypted, matches the result of operations performed on the original data. In [18], HE for deep neural networks was proposed. Extending HE to deeper neural networks, a low-complexity encryption method for DNNs was introduced in [7] and [19]. Although HE effectively prevents privacy leakage, the state-of-the-art HE remains extremely slow for computing large neural networks.

B. Perceptual Encryption

In the inference stage, PE has emerged as a promising method for finding a suitable encrypted domain for visual images. In [20], a cycle-GAN-based PE method was proposed for medical images, requiring encryption/decryption keys. In [21], a more advanced method eliminated the need for these keys, using the cycle-GAN model itself as a unique encryption key. However, the purported security of PE is questionable. Numerous studies have demonstrated that PE is highly vulnerable to DR attacks, which can effectively restore original images even against pixel-based encryption [22] and learnable encryption [23]. This vulnerability underscores a critical flaw in PE methods, challenging their viability for robust security in practical applications.

C. Adversarial Examples

Adversarial examples are widely used in privacy-related deep learning technologies due to their versatility, such as for inserting watermarks in foundation models [24], [25]. For hiding visual information, unlike HE and PE, steganography [26] and AVIH [1] are practical approaches, as they can hide information or recover the original data with low computational complexity. More specifically, AVIH can guarantee the correctness of computational results. The variance-consistency

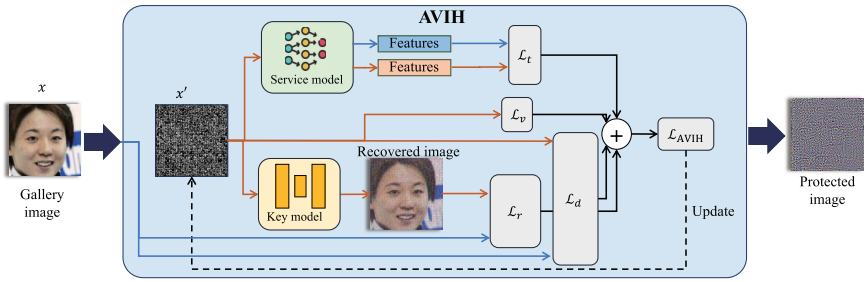

Fig. 4. AVIH method proposed in [1]. Given a key model, the AVIH method aims to generate type-I adversarial examples (noisy images) that produce very similar outputs for the target service model. To achieve this, the adversarial image x' is modified to minimize task loss (L_t), difference loss (L_d), recovery loss (L_r), and variance consistency loss (L_v). The details of these losses will be presented in Section III.

loss used in AVIH can efficiently encrypt visual information while preserving the correctness of DNN computations without requiring any retraining process, which could lead to additional privacy leakage. However, privacy-preserving aspects of AVIH have been discussed heuristically and do not meet stringent privacy constraints. In this work, we present the first approach that executes DR attacks against the AVIH method.

III. Preliminary: AVIH Encryption

In this section, we first briefly review the AVIH method. Consider a remote computing scenario where clients provide a gallery dataset and a target service model to a cloud server. If the original gallery dataset is sent to the server without encryption, the sensitive and private visual images that clients wish to keep confidential can be accessed by the server administrator. To address this, the AVIH method is proposed to secure the gallery dataset images while maintaining the service quality of the target service model. To ensure the reconstruction of the original data from the encrypted data, clients have their private and secure key model, accessible only to them. For image recognition tasks, such as face recognition [27], [28] or human re-identification [29], the encrypted gallery datasets can be used without a loss of accuracy. Additionally, while humans cannot recognize the original image from the encrypted version, the secure key model can accurately reconstruct the original image.

A. Overall Loss Function

In the AVIH method, the primary goal is to generate images that: 1) preserve the service model's output; 2) destroy the image's structural information; and 3) guarantee recovery of the original image with a key model. From the original image x_i, we generate the image x_i', where i denotes the index of the gallery images. The generated image x_i' is designed to minimize the following loss function:

$$L_{\text{AVIH}}\left(x_i, x_i'\right) = L_t\left(x_i, x_i'\right) - \lambda_1 L_d\left(x_i, x_i'\right) + \lambda_2 L_v\left(x_i'\right) + \lambda_3 L_r\left(x_i, x_i'\right). \tag{1}$$

The details of the AVIH loss function (1) are introduced in the remainder of this section. Fig. 4 illustrates the schematic of this method.

B. Task Loss (L_t)

The task loss is related to the preservation of the service model's output. With the service model $f_s(\cdot)$, the loss function is defined as follows:

$$L_t\left(x_i, x_i'\right) = \left\| f_s\left(x_i\right) - f_s\left(x_i'\right) \right\|_2 \tag{2}$$

where the distance metric is defined using ℓ_2 norm; however, other distance metrics such as ℓ_1 and ℓ_∞ norms can be also used.

C. Difference Loss (L_d)

The difference loss is related to destroying the image's structural information. Motivated by Type-I adversarial attacks, this loss function is defined by the ℓ_1 or ℓ_2 distance between x_i and x_i' as follows:

$$L_d\left(x_i, x_i'\right) = \left\| x_i - x_i' \right\|_2. \tag{3}$$

D. Recovery Loss (L_r)

Let us define the key model as $f_k(\cdot)$. To recover the original image, the recovery loss is defined by the ℓ_2 distance between the original image x_i and the recovered image $f_k(x_i')$ as follows:

$$L_r\left(x_i, x_i'\right) = \left\| x_i - f_k\left(x_i'\right) \right\|_2. \tag{4}$$

E. Variance Consistency Loss (L_v)

The variance consistency (VC) loss, proposed in [1], addresses the trade-off between protection quality and recovery quality. To formulate this loss function, we divide the encrypted image x_i' into N overlapping patches, each with height h and width w. Let us define the sum of the pixel values of the nth patch as S_n, where $S_n^{(r)}$, $S_n^{(g)}$, and $S_n^{(b)}$ denote the sum of pixel values for the red, green, and blue channels, respectively. The VC loss then measures the channel-wise variance of S_n as follows:

$$L_v\left(x_i'\right) = \texttt{var}\left(S_n^{(r)}\right) + \texttt{var}\left(S_n^{(g)}\right) + \texttt{var}\left(S_n^{(b)}\right) \tag{5}$$

where $\texttt{var}(S_n^{(r)})$ denotes the empirical variance of the sequence $S_1^{(r)}, S_2^{(r)}, \ldots, S_N^{(r)}$.

Remark 1: If we aim to obtain an image that is robust against potential attackers, we should set a higher weight on the VC loss. Consequently, the quality of the image recovered using the key model will be poor.

IV. Methodology

In this work, our primary focus is to demonstrate that sharing the same key models for hiding visual information poses potential security and privacy vulnerabilities.

A. Proposed DR Attack

Here, we first introduce the threat scenario of the AVIH method. Then, we present the proposed DR attack for the gallery dataset, where the visual information is hidden in noise-like images.

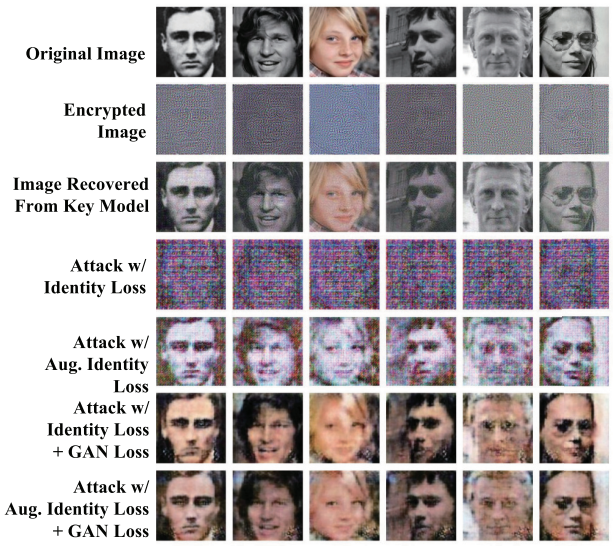

Fig. 5. Ablation study for the key features of our proposed work (augmented identity loss and GAN loss). Here, we assume that 3% of the encrypted data shares the same key model.

1) Threat Scenario: In this work, we follow the system model in the AVIH method, where the target service type is image recognition. In an image recognition system, the class of a target image can be identified by comparing it with images in the gallery dataset. As shown in Fig. 4, the AVIH method transforms the images in the gallery dataset into noise-like images while preserving the output of the service DNN.

Let us define the original gallery dataset as $\mathcal{G}$ and the encrypted gallery dataset as $\tilde{\mathcal{G}}$. We consider a threat scenario where a malicious attacker aims to reconstruct the hidden images from noise-like images with access to the: 1) *encrypted dataset* and 2) *service DNN model*. According to [1], the key model can reconstruct the original gallery dataset from the encrypted gallery dataset and is only available on the client-side.

More specifically, without making strong assumptions, since the service DNN weights are available at the remote server, we assume*a white box access scenario*, i.e., an attacker can access the weights of the service DNN model.

2) Motivation: In this work, we focus on the fact that a private key model can reconstruct the original gallery dataset. This implies that there is an *unknown but specific relationship* between the encrypted and original datasets. Motivated by this, we aim to mimic the functionality of the key model. However, we neither have access to the key model nor know how it was trained. To address this, we leverage the generative model [30], which is widely used for guaranteeing photorealism. However, generative model inversion attacks typically focus on mimicking the distribution of the training dataset and cannot reconstruct an image corresponding to a specific DNN output. To directly reconstruct the original gallery dataset, we consider the attacker key model as $a_k(\cdot;\theta)$, where θ denotes its weights. For simplicity, we use $a_k(\cdot;\theta)$ and $a_k(\cdot)$ interchangeably.

3) Identity Loss: With the given information, we can assume that the target key model also preserves the output of the service model. Additionally, since we have access to the weights of the service DNN, we can formulate the identity loss as follows:

$$L_{\mathrm{I}}(\tilde{\mathcal{G}}) = \mathbb{E}_{x'\sim\tilde{\mathcal{G}}} \left\| f_s\left(a_k\left(x_i';\theta\right)\right) - f_s\left(x_i'\right)\right\|_2 \tag{6}$$

where $\tilde{\mathcal{G}}$ denotes the encrypted images sharing the same key model. Instead of an ℓ_2-based loss, the identity loss function can be formulated to maximize the log-likelihood of the target class c or the cosine similarity of two feature vectors.

4) Overfitting in DR: In Fig. 5, we show examples of the original images and the encrypted images by the AVIH method. The figure also depicts the images recovered by the original key model and the attacker key model. As shown, the attacker model trained with identity loss in (6) recovers very noisy images compared to the original images or those recovered by the original key model. This issue is quite similar to *overfitting* in ordinary machine-learning problems. More specifically, there is a true relationship between the encrypted images and the original images, represented by the original key model. However, a few encrypted images are not sufficient to demonstrate this relationship using the service model. Therefore, in the remainder of this section, we propose: 1) augmented identity loss and 2) a GAN-based training scheme to secure the generalization of the trained attacker key model. The details of these methods are depicted in Fig. 6.

5) Augmented Identity Loss: In typical machine-learning model training, data augmentation is widely used to increase the validation/test accuracy of the trained model, i.e., for better generalization Similarly, to alleviate the overfitting issue in DR, we combine data augmentation with identity loss. Let us consider an image x' drawn from the encrypted dataset $\tilde{\mathcal{G}}$. Then, we may reconstruct the original image using $a_k(x';\theta)$. Unlike the canonical identity loss in (6), we apply random data augmentation before forwarding the image into the service model $f_x(\cdot)$. Denoting the random augmentation process as $T(\cdot)$, the identity loss in (6) can be redefined as follows:

$$L_{\mathrm{I}}(\tilde{\mathcal{G}}) = \mathbb{E}_{x'\sim\tilde{\mathcal{G}}} \left\| f_s\left(T\left(a_k\left(x_i';\theta\right)\right)\right) - f_s\left(x_i'\right)\right\|_2. \tag{7}$$

In our work, we consider the following data augmentation methods: 1) random horizontal flip; 2) random padding; and 3) random crop. In the right part of Fig. 6, we illustrate the concept of augmented identity loss. For further generalization, we also experimented with randomized smoothing on the reconstructed data $a_k(x';\theta)$; however, it did not produce notable differences.

6) Generative Model Inversion Attack: In this paragraph, we aim to resolve the overfitting issue using a GAN-based loss function. Intuitively, if we want to find an image that has the same output as the encrypted image x', there would be many possible images, most of which are unnatural. By reducing the number of cases by restricting the unnatural images, we can resolve the overfitting issue.

To make the reconstructed images natural, we consider an optimization problem that minimizes the augmented identity loss with a Jensen–Shannon (JS) divergence constraint between the auxiliary dataset and the reconstructed images as follows:

$$\min_{a_k(\cdot)} L_{\mathrm{I}}(\tilde{\mathcal{G}}), \quad \text{s.t. } D_{\mathrm{JS}}(a_k(\tilde{\mathcal{G}})\|\mathcal{X}) \le \epsilon \tag{8}$$

where $\mathcal{X}$ denotes the auxiliary dataset, and $D_{\mathrm{JS}}(a_k(\tilde{\mathcal{G}})\|\mathcal{X})$ denotes the JS divergence between the reconstructed images $a_k(x'), x' \sim \tilde{\mathcal{G}}$ and the auxiliary images $x'' \sim \mathcal{X}$. From this

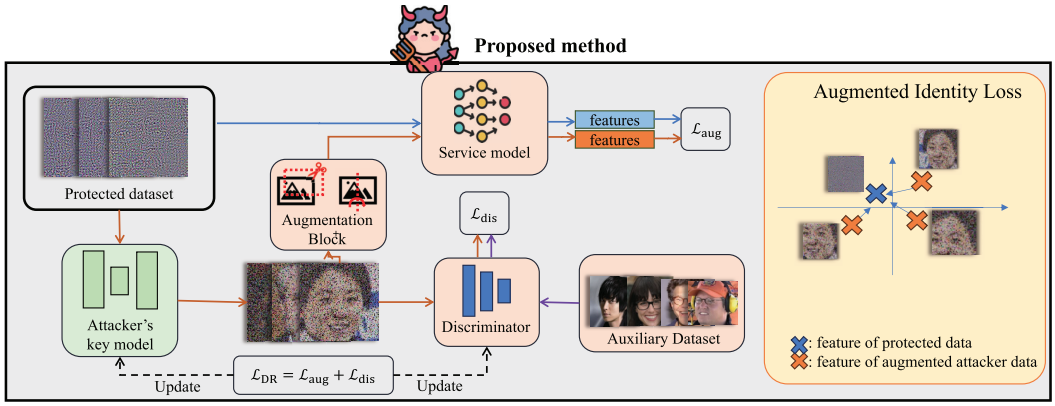

Fig. 6. Illustration of the proposed DR attack for the AVIH method. The attacker (possibly an honest-but-curious cloud server manager) is assumed to have access to the protected dataset and the auxiliary dataset. The attacker first initializes a key model. Then, the key model's weights are updated to be more photorealistic (discriminator loss, $\mathcal{L}_{\text{dis}}$) and to reconstruct the original image via feature matching (augmented identity loss, $\mathcal{L}_{\text{aug}}$).

optimization, we derive the Lagrangian of the problem (8) as follows:

$$\mathcal{L} = L_I\left(\tilde{\mathcal{G}}\right) + \lambda\left(D_{\text{JS}}\left(a_k\left(\tilde{\mathcal{G}}\right)\|\mathcal{X}\right) - \epsilon\right) \tag{9}$$

where $\mu \geq 0$. Let us consider the minimizer $a_k(\cdot;\theta)$ of the problem as $a_k^*(\cdot;\theta)$. Then, for any constant ϵ, there exists $\mu \geq 0$ that makes $a_k^*(\cdot;\theta)$ a minimizer of the original problem (8). The intuition is that for the minimizer $a_k^*(\cdot;\theta)$, a larger value of λ indicates a smaller ϵ in (8). Hence, we aim to minimize the Lagrangian in (9).

The next step is to convert the function in (9) into a GAN formulation. As shown in [30], the JS divergence minimization problem can be replaced by the GAN optimization problem. For brevity, we use simpler notation: the probability density function of the recovered images $x^* \sim a_k(\tilde{\mathcal{G}}) = p(x)$ and the auxiliary images $x'' \sim \mathcal{X} = q(x)$. The JS divergence in (9) can then be rewritten as follows:

$$\begin{aligned} D_{\text{JS}}\left(a_k\left(\tilde{\mathcal{G}}\right)\|\mathcal{X}\right)\Big) &= D_{\text{JS}}(p\|q) \\ &\propto \mathop{\mathbb{E}}_{x^*\sim p(x)}\left[\log\frac{2p(x^*)}{p(x^*)+q(x^*)}\right] \\ &\quad + \mathop{\mathbb{E}}_{x''\sim q(x)}\left[\log\frac{2q(x'')}{p(x'')+q(x'')}\right]. \end{aligned} \tag{10}$$

By defining a discriminator as $D(x)$, we can convert the JS divergence into a GAN formulation as follows:

$$\begin{aligned} D_{\text{JS}}(p\|q) &= \max_D \mathbb{E}_{x^*\sim p}\left[\log(D(x^*))\right] \\ &\quad + \mathbb{E}_{x''\sim q}\left[\log(1-D(x''))\right] \end{aligned} \tag{11}$$

where the optimal D is $(p(x)/(p(x)+q(x)))$. Then, the loss function for the attacker key model is defined by

$$L_{\text{key}}(\tilde{\mathcal{G}}) = \mathbb{E}_{x'\sim\tilde{\mathcal{G}}}\left[\log D(a_k(x';\theta))\right] + \lambda_1 \cdot L_1(\tilde{\mathcal{G}}) \tag{12}$$

where D is the discriminator model, and the last layer is activated by the hyperbolic tangent function. Similarly, the discriminator loss is defined as follows:

$$\begin{aligned} L_{\text{dis}}(\tilde{\mathcal{G}}) &= -\mathbb{E}_{x'\sim\tilde{\mathcal{G}}}\left[\log D(a_k(x';\theta))\right] \\ &\quad - \mathbb{E}_{x\sim\mathcal{D}_{\text{aux}}}\left[\log(1-D(x))\right]. \end{aligned} \tag{13}$$

Since the original images and the auxiliary images are not identical but belong to the same category (e.g., face images), we use a patch-GAN model for our optimization, where the discriminator D classifies true and false patches of the images. In Fig. 6, the attacker aims to generate photorealistic results by deceiving the discriminator.

Remark 2 (JS Divergence Versus KL Divergence): In our problem formulation (8), we use the JS divergence between the recovered images and auxiliary images as our constraint for photorealism. Another metric, KL divergence, is widely used to ensure similarity between two datasets. We have tried with the KL divergence formulation; however, it is closely related to variational inference, which requires a pretrained GAN. Since a pretrained GAN is not suitable for exact DR attacks, we use JS divergence as our constraint.

B. Ablation Study of Key Features

Before introducing our experimental results, we briefly present graphical examples of our attacker key model with and without our key features for resolving overfitting issues. In Section IV-A, we proposed the augmented identity loss and GAN-based training loss. In Fig. 5, we show our results on the AgeDB-30 dataset, using the CelebA dataset as the auxiliary dataset. As depicted in the fourth column, the results with the canonical identity loss recover images that are not very similar to the original ones. However, by leveraging augmented identity loss, the shape of the images can be recovered, though the colors are not realistic. On the other hand, using GAN-based training yields more photorealistic recovered images. Moreover, by combining both methods, the quality of the recovered images is further enhanced.

V. Experiments

In this section, we evaluate the proposed DR attack against AVIH. Since there has been no prior work on exact DR attack methods for deep neural network models, we measure the quality of the reconstruction using various metrics.[3] Instead of comparing with other methods, we conduct an ablation study on our key contributions: 1) augmented identity loss and 2) GAN-based training.

A. Experimental Details

We conduct two main experiments: one for the face recognition scenario and another for the object re-identification

[3]Only Haim et al. [17] have shown exact DR, but their method requires the target DNN to be a homogeneous neural network, which is not a practical assumption.

scenario. Both experiments are performed on a workstation equipped with an AMD Ryzen R9 5950x 16-core CPU and an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM.

1) Face Recognition Scenario: In the face recognition scenario, we perform DR attacks on three target face datasets: LFW [31], AgeDB-30 [32], and CFP-FP [33]. For the service model on the local server and cloud, we use ArcFace [27] and AdaFace [28] models with IR-18 and IR-50 backbones. For the DR attack, we need to find an auxiliary dataset for the face recognition model. To this end, we choose the CelebA [34] dataset as our auxiliary dataset, which is the most famous and commonly used dataset for face-related tasks. As an evaluation metric for face recognition, we use TAR@FAR =0.01 for the reconstructed images via our method. For this accuracy evaluation, we use the AdaFace model with an IR-101 backbone.

2) Implementation of AVIH: For the implementation of the AVIH method, we follow the hyperparameters used in the original paper [1]. For example, we update encrypted images (x') for 800 epochs, and the kernel size for the VC loss is set to 4. The weights for the difference loss, recovery loss, and VC loss are set to 0.03, 0.5, and 3.0, respectively. The key model is configured using the standard U-Net [35]. In our experiment, we slightly modify the gallery sets of the datasets to contain 2000 images. We then run the AVIH method on these 2000 gallery images, using 1000 images for training our attacker key model and the remaining 1000 images for evaluating the trained attacker key model.

3) Implementation of Our DR: In our DR attack, we train an attacker key model based on the U-Net structure.[4] We trained our key model for 1600 steps with a batch size of 32. The weight on the augmented identity loss is set to 30.0. For the augmented identity loss function, we use the following data augmentation methods: random horizontal flip, random padding of five pixels, and random cropping to the original size.

4) Evaluation Metrics for Reconstruction Quality Measurement: To measure the quality of the reconstructed images from our DR attack, we used the following evaluation metrics: 1) mean square error (MSE); 2) learned perceptual image patch similarity (LPIPS) [36]; 3) peak signal-to-noise ratio (PSNR); 4) contrastive language-image pretraining (CLIP) [37]; and 5) structural similarity index measure (SSIM) [38].

B. Accuracy and Similarity Metrics

In our experiments, before implementing our DR attack scheme, the AVIH [1] encrypts the gallery dataset of three face recognition datasets: AgeDB-30, LFW, and CFP-FP. We note that the encrypted gallery dataset successfully performs face recognition tasks for the target service model. For example, with the encrypted LFW gallery dataset, the cloud achieves a TPR accuracy of 98.40%, which is the same accuracy as with the original gallery dataset.

1) Accuracy Metrics: In Table I, we present the TPR performance metrics at FPR 0.01 across three datasets: AgeDB-30, LFW, and CFP-FP. The TPR accuracy values are provided for the proposed method (*Ours*) at different percentages of images sharing the same key model (1%, 3%, 10%, and 70%). We also compare the results of our method with the following: 1) original gallery dataset; 2) encrypted dataset; 3) dataset reconstructed by the key model; and 4) random face images.

[4]We tried other structures by modifying the U-Net structure, but the results were almost the same.

TABLE I

ACCURACY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS

	Same-key data (%)	TPR ↑		
		AgeDB-30	LFW	CFP-FP
Ours	1%	$0.628^{\pm 0.058}$	$0.612^{\pm 0.136}$	$0.554^{\pm 0.054}$
	3%	$0.746^{\pm 0.043}$	$0.825^{\pm 0.039}$	$0.630^{\pm 0.018}$
	10%	$0.761^{\pm 0.254}$	$0.737^{\pm 0.341}$	$0.663^{\pm 0.039}$
	70%	$\mathbf{0.924^{\pm 0.011}}$	$\mathbf{0.970^{\pm 0.006}}$	$\mathbf{0.817^{\pm 0.019}}$
Original	-	$0.980^{\pm 0.000}$	$0.998^{\pm 0.000}$	$0.971^{\pm 0.000}$
Protected	-	$0.211^{\pm 0.000}$	$0.149^{\pm 0.000}$	$0.579^{\pm 0.000}$
Key model	-	$0.971^{\pm 0.000}$	$0.998^{\pm 0.000}$	$0.964^{\pm 0.000}$
Random	-	$0.215^{\pm 0.000}$	$0.036^{\pm 0.000}$	$0.590^{\pm 0.000}$

TABLE II

QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY THE ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS. THE DATASET IS THE AGEDB-30 DATASET, AND THE USED BACKBONE NETWORK MODEL IS THE IR-18 NETWORK MODEL

		MSE ↓	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
AgeDB-30 Dataset	Ours (1%)	$0.114^{\pm 0.007}$	$0.540^{\pm 0.008}$	$9.960^{\pm 0.268}$	$0.715^{\pm 0.006}$	$0.249^{\pm 0.003}$
	Ours (3%)	$0.071^{\pm 0.006}$	$0.450^{\pm 0.018}$	$11.847^{\pm 0.394}$	$0.746^{\pm 0.013}$	$0.270^{\pm 0.005}$
	Ours (10%)	$0.121^{\pm 0.149}$	$0.401^{\pm 0.069}$	$10.996^{\pm 2.766}$	$0.774^{\pm 0.017}$	$0.269^{\pm 0.046}$
	Ours (70%)	$\mathbf{0.063^{\pm 0.008}}$	$\mathbf{0.305^{\pm 0.017}}$	$\mathbf{12.464^{\pm 0.593}}$	$\mathbf{0.811^{\pm 0.009}}$	$\mathbf{0.298^{\pm 0.006}}$
	Original	$0.001^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
	Protected	$0.662^{\pm 0.000}$	$1.214^{\pm 0.000}$	$1.866^{\pm 0.000}$	$0.641^{\pm 0.000}$	$0.019^{\pm 0.000}$
	Key model	$0.022^{\pm 0.000}$	$0.367^{\pm 0.000}$	$17.525^{\pm 0.000}$	$0.837^{\pm 0.000}$	$0.302^{\pm 0.000}$
LFW Dataset	Ours (1%)	$0.142^{\pm 0.077}$	$0.527^{\pm 0.043}$	$9.120^{\pm 1.781}$	$0.747^{\pm 0.022}$	$0.246^{\pm 0.026}$
	Ours (3%)	$0.060^{\pm 0.007}$	$0.377^{\pm 0.009}$	$12.395^{\pm 0.492}$	$0.813^{\pm 0.017}$	$0.291^{\pm 0.004}$
	Ours (10%)	$0.140^{\pm 0.181}$	$0.352^{\pm 0.086}$	$11.232^{\pm 3.934}$	$0.832^{\pm 0.042}$	$0.276^{\pm 0.061}$
	Ours (70%)	$\mathbf{0.048^{\pm 0.008}}$	$\mathbf{0.268^{\pm 0.014}}$	$\mathbf{13.366^{\pm 0.673}}$	$\mathbf{0.865^{\pm 0.011}}$	$\mathbf{0.323^{\pm 0.006}}$
	Original	$0.001^{\pm 0.000}$	$0.008^{\pm 0.000}$	$32.067^{\pm 0.000}$	$0.981^{\pm 0.000}$	$0.477^{\pm 0.000}$
	Protected	$0.622^{\pm 0.000}$	$1.302^{\pm 0.000}$	$2.108^{\pm 0.000}$	$0.556^{\pm 0.000}$	$0.015^{\pm 0.000}$
	Key model	$0.006^{\pm 0.000}$	$0.243^{\pm 0.000}$	$23.071^{\pm 0.000}$	$0.919^{\pm 0.000}$	$0.385^{\pm 0.000}$
CFP-FP Dataset	Ours (1%)	$0.210^{\pm 0.020}$	$0.623^{\pm 0.019}$	$7.136^{\pm 0.419}$	$0.728^{\pm 0.011}$	$0.163^{\pm 0.003}$
	Ours (3%)	$0.214^{\pm 0.047}$	$0.566^{\pm 0.027}$	$7.272^{\pm 0.891}$	$0.755^{\pm 0.015}$	$0.177^{\pm 0.004}$
	Ours (10%)	$\mathbf{0.176^{\pm 0.030}}$	$0.456^{\pm 0.018}$	$\mathbf{8.007^{\pm 0.676}}$	$0.804^{\pm 0.010}$	$0.187^{\pm 0.004}$
	Ours (70%)	$0.210^{\pm 0.053}$	$\mathbf{0.408^{\pm 0.015}}$	$7.310^{\pm 0.982}$	$\mathbf{0.835^{\pm 0.005}}$	$\mathbf{0.204^{\pm 0.004}}$
	Original	$0.001^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
	Protected	$0.636^{\pm 0.000}$	$1.219^{\pm 0.000}$	$2.044^{\pm 0.000}$	$0.633^{\pm 0.000}$	$0.020^{\pm 0.000}$
	Key model	$0.028^{\pm 0.000}$	$0.435^{\pm 0.000}$	$16.377^{\pm 0.000}$	$0.825^{\pm 0.000}$	$0.285^{\pm 0.000}$

For all three datasets, TPR values of the proposed method gradually increase as more gallery images share the same key model. For instance, in the AgeDB-30 dataset, the proposed method achieves a TPR of 0.628 with 1% same-key data, which increases to 0.746 with 3%, 0.761 with 10%, and 0.924 with 70% same-key data.

On the other hand, with the original gallery datasets, the TPR accuracy values are sufficient for recognizing most of the query images. For example, the TPR value for the AgeDB-30 dataset is 98.0%. More importantly, the gallery dataset reconstructed by the original private key model performs almost the same as the original gallery dataset. The encrypted gallery set has significantly lower TPR values since the evaluation service model (IR-101 backbone) is different from the target service model (IR-18 backbone).

To summarize, the proposed method shows significant improvement in TPR by executing DR attacks against the AVIH encryption method.

Fig. 7. Graphical examples of the proposed method. The images in the first row with green borderlines show the original image and the images in the other rows show the reconstructed images by the proposed method.

2) Image Similarity Metrics: In Table II, we present the evaluation results for the three face recognition datasets. The numbers in the table are computed using each similarity metric between the original and reconstructed images. Similar to the accuracy metric benchmark in Table I, we evaluate the proposed method at different percentages of images sharing the same key model (1%, 3%, 10%, and 70%).

In the AgeDB-30 and LFW datasets, all similarity metrics improve as more images share a common key model. Interestingly, if only 3% of the images share the same privacy key model, the quality of reconstructed images is comparable to that of the true key model. For example, in the AgeDB-30 dataset, the PSNR for 3% shared key model images is 11.847 compared to 17.525 for the true key model images, and the mse for 3% shared key model images is 0.071 compared to 0.022 for the true key model images.

Furthermore, although pixel-based metrics such as PSNR and mse show some differences between different percentages of images sharing the same key model, other metrics like LPIPS and CLIP do not show significant differences. For instance, in the AgeDB-30 dataset, the LPIPS for 3% shared key model images is 0.450 compared to 0.367 for the true key model images, and the CLIP for 3% shared key model images is 0.746 compared to 0.837 for the true key model images.

For the CFP-FP dataset, the reconstruction quality is relatively lower compared to the other two datasets. This is because the reconstruction quality with the true key model serves as a performance cap for the replicated key model, where the true key model's reconstruction quality is relatively lower. However, similar to the other two datasets, perceptual quality metrics such as LPIPS and CLIP still perform well. For example, in the CFP-FP dataset, the LPIPS for the true key model is 0.435, which is comparable to 0.623 for 1% shared key model images. The CLIP metric also shows a consistent trend, with 0.825 for the true key model and 0.728 for 1% shared key model images.

3) Graphical Results: In Fig. 7, we present graphical examples of the proposed method. The first row with green borders shows the original images, while the subsequent rows display the reconstructed images with different percentages of leaked encrypted data (1%, 3%, 10%, and 70%).

With 1% leaked encrypted data, the reconstructed images are significantly distorted and blurred, making recognition difficult, which aligns with lower similarity scores. As the percentage increases to 3%, facial features become more distinguishable despite some blurring, showing noticeable improvement. At 10%, the images are clearer and more recognizable, and at 70%, the reconstructed images are very close to the original quality, supporting the highest similarity scores and demonstrating the method's effectiveness.

These results highlight that while pixel-based metrics like PSNR and mse show improvement, perceptual similarity metrics such as LPIPS and CLIP also indicate significant enhancements in image quality (in Table II). This improvement in reconstruction quality directly correlates with an increase in TPR (in Table I), further validating the robustness of the proposed method in maintaining high image similarity and effective DR as more images share the same key model.

TABLE III

ACCURACY AND QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY THE ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS. THE USED DATASET IS AGEDB-30, AND THE TARGET FACE RECOGNITION MODEL IS ADAFACE AND ARCFACE WITH THE IR-50 BACKBONE NETWORK MODEL

		TPR ↑	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
AdaFace & IR-50	Ours (1%)	$0.719^{\pm 0.045}$	$0.530^{\pm 0.012}$	$10.391^{\pm 0.377}$	$0.719^{\pm 0.007}$	$0.256^{\pm 0.004}$
	Ours (3%)	$0.861^{\pm 0.031}$	$0.433^{\pm 0.017}$	$12.331^{\pm 0.418}$	$0.747^{\pm 0.007}$	$0.284^{\pm 0.007}$
	Ours (10%)	$0.898^{\pm 0.020}$	$0.378^{\pm 0.016}$	$12.594^{\pm 0.630}$	$0.772^{\pm 0.011}$	$0.303^{\pm 0.005}$
	Ours (70%)	$\mathbf{0.932^{\pm 0.011}}$	$\mathbf{0.341^{\pm 0.027}}$	$\mathbf{12.599^{\pm 0.973}}$	$\mathbf{0.797^{\pm 0.013}}$	$\mathbf{0.310^{\pm 0.007}}$
	Original	$0.980^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
	Protected	$0.311^{\pm 0.000}$	$1.219^{\pm 0.000}$	$2.044^{\pm 0.000}$	$0.633^{\pm 0.000}$	$0.020^{\pm 0.000}$
	Key model	$0.952^{\pm 0.000}$	$0.435^{\pm 0.000}$	$16.377^{\pm 0.000}$	$0.825^{\pm 0.000}$	$0.285^{\pm 0.000}$
ArcFace & IR-50	Ours (1%)	$0.709^{\pm 0.044}$	$0.537^{\pm 0.010}$	$10.348^{\pm 0.233}$	$0.716^{\pm 0.007}$	$0.253^{\pm 0.004}$
	Ours (3%)	$0.823^{\pm 0.020}$	$0.440^{\pm 0.015}$	$11.787^{\pm 0.432}$	$0.753^{\pm 0.008}$	$0.278^{\pm 0.009}$
	Ours (10%)	$0.885^{\pm 0.012}$	$0.373^{\pm 0.016}$	$12.122^{\pm 0.599}$	$0.776^{\pm 0.008}$	$0.291^{\pm 0.004}$
	Ours (70%)	$\mathbf{0.917^{\pm 0.012}}$	$\mathbf{0.333^{\pm 0.022}}$	$\mathbf{12.338^{\pm 0.377}}$	$\mathbf{0.802^{\pm 0.007}}$	$\mathbf{0.304^{\pm 0.004}}$
	Original	$0.980^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
	Protected	$0.167^{\pm 0.000}$	$1.160^{\pm 0.000}$	$2.104^{\pm 0.000}$	$0.603^{\pm 0.000}$	$0.020^{\pm 0.000}$
	Key model	$0.982^{\pm 0.000}$	$0.222^{\pm 0.000}$	$20.664^{\pm 0.000}$	$0.868^{\pm 0.000}$	$0.351^{\pm 0.000}$

C. Results for Various Face Recognition Models

In Table III, we present the accuracy and similarity metrics for various backbones and face recognition schemes. Unlike the benchmarks in Tables I and II, the AdaFace and ArcFace face recognition models are used for evaluation, with their backbone configured as the IR-50 network. As shown in the table, similar to the previous results, all evaluation metrics of the proposed method improve as more encrypted images share the same key model. For example, if 70% of the gallery images share the same key model, the proposed method nearly achieves the reconstruction quality of the original key model. Interestingly, the proposed scheme can nearly achieve the perceptual similarity score of the original key model even when only 3% of the gallery images share the key model.

This experiment demonstrates that the proposed method can be generally applied to various face recognition models and backbone networks.

D. Ablation Study

In this section, we aim to study the effect of our key contributions: 1) augmented identity loss and 2) GAN-based key model training. To this end, we implement the proposed method for all cases, whether the key contributions exist

TABLE IV

ABLATION STUDY FOR OUR KEY CONTRIBUTIONS: 1) GAN-BASED TRAINING AND 2) AUGMENTED IDENTIFICATION (ID) LOSS. IN THIS TABLE, THE TARGET DATASET IS THE AGEDB-30 DATASET, WHERE THE TARGET SERVICE MODEL IS ADAFACE WITH THE IR-18 BACKBONE NETWORK. AS PRESENTED IN THIS TABLE, OUR KEY CONTRIBUTIONS EFFECTIVELY RESOLVE THE OVERFITTING ISSUES ON THE DR WHEN A SMALL PORTION OF THE GALLERY IMAGES SHARE THE SAME KEY MODEL

# Leaked data	GAN Loss	Aug. ID. Loss	TPR ↑	MSE ↓	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
1%	✗	✗	$0.329^{\pm 0.110}$	$0.418^{\pm 0.050}$	$0.945^{\pm 0.073}$	$3.954^{\pm 0.512}$	$0.653^{\pm 0.072}$	$0.051^{\pm 0.005}$
	✗	✓	$0.524^{\pm 0.028}$	$0.271^{\pm 0.057}$	$0.923^{\pm 0.038}$	$6.036^{\pm 0.904}$	$\mathbf{0.772^{\pm 0.008}}$	$0.094^{\pm 0.009}$
	✓	✗	$0.517^{\pm 0.091}$	$0.116^{\pm 0.013}$	$0.571^{\pm 0.015}$	$9.840^{\pm 0.508}$	$0.709^{\pm 0.005}$	$0.236^{\pm 0.008}$
	✓	✓	$\mathbf{0.628^{\pm 0.058}}$	$\mathbf{0.114^{\pm 0.007}}$	$\mathbf{0.540^{\pm 0.008}}$	$\mathbf{9.960^{\pm 0.268}}$	$0.715^{\pm 0.006}$	$\mathbf{0.249^{\pm 0.003}}$
3%	✗	✗	$0.299^{\pm 0.134}$	$0.437^{\pm 0.060}$	$0.937^{\pm 0.109}$	$3.792^{\pm 0.600}$	$0.657^{\pm 0.057}$	$0.052^{\pm 0.004}$
	✗	✓	$0.549^{\pm 0.051}$	$0.323^{\pm 0.060}$	$0.897^{\pm 0.038}$	$5.318^{\pm 0.759}$	$\mathbf{0.781^{\pm 0.010}}$	$0.096^{\pm 0.010}$
	✓	✗	$0.543^{\pm 0.186}$	$0.130^{\pm 0.135}$	$0.519^{\pm 0.042}$	$10.346^{\pm 2.388}$	$0.719^{\pm 0.010}$	$0.244^{\pm 0.036}$
	✓	✓	$\mathbf{0.746^{\pm 0.043}}$	$\mathbf{0.071^{\pm 0.006}}$	$\mathbf{0.450^{\pm 0.018}}$	$\mathbf{11.847^{\pm 0.394}}$	$0.746^{\pm 0.013}$	$\mathbf{0.270^{\pm 0.005}}$
10%	✗	✗	$0.370^{\pm 0.084}$	$0.375^{\pm 0.034}$	$0.999^{\pm 0.054}$	$4.419^{\pm 0.379}$	$0.615^{\pm 0.041}$	$0.058^{\pm 0.004}$
	✗	✓	$0.662^{\pm 0.063}$	$0.264^{\pm 0.039}$	$0.700^{\pm 0.036}$	$6.244^{\pm 0.641}$	$\mathbf{0.807^{\pm 0.007}}$	$0.126^{\pm 0.009}$
	✓	✗	$\mathbf{0.787^{\pm 0.063}}$	$\mathbf{0.073^{\pm 0.010}}$	$0.413^{\pm 0.026}$	$\mathbf{11.832^{\pm 0.579}}$	$0.761^{\pm 0.014}$	$\mathbf{0.272^{\pm 0.013}}$
	✓	✓	$0.761^{\pm 0.254}$	$0.121^{\pm 0.149}$	$\mathbf{0.401^{\pm 0.069}}$	$10.996^{\pm 2.766}$	$0.774^{\pm 0.017}$	$0.269^{\pm 0.046}$
70%	✗	✗	$0.860^{\pm 0.022}$	$0.193^{\pm 0.049}$	$0.716^{\pm 0.046}$	$7.431^{\pm 1.020}$	$0.802^{\pm 0.011}$	$0.119^{\pm 0.013}$
	✗	✓	$0.891^{\pm 0.021}$	$0.187^{\pm 0.042}$	$0.490^{\pm 0.035}$	$7.581^{\pm 1.040}$	$\mathbf{0.832^{\pm 0.005}}$	$0.197^{\pm 0.013}$
	✓	✗	$\mathbf{0.929^{\pm 0.009}}$	$0.068^{\pm 0.010}$	$0.341^{\pm 0.027}$	$12.153^{\pm 0.659}$	$0.796^{\pm 0.008}$	$\mathbf{0.304^{\pm 0.007}}$
	✓	✓	$0.924^{\pm 0.011}$	$\mathbf{0.063^{\pm 0.008}}$	$\mathbf{0.305^{\pm 0.017}}$	$\mathbf{12.464^{\pm 0.593}}$	$0.811^{\pm 0.009}$	$0.298^{\pm 0.006}$

or not. In Table IV, we present the quantitative results of our ablation study. At lower percentages of leaked encrypted data, particularly 1% and 3%, we observe significant performance improvements due to the ablation study configurations. For example, incorporating both GAN loss and augmented identity loss at 1% leaked data increases the TPR from 0.329 to 0.628 and reduces the mse from 0.418 to 0.114. Similarly, for 3% leaked data, the TPR increases from 0.299 to 0.746 and the mse decreases from 0.437 to 0.071. These enhancements demonstrate the effectiveness of the ablation configurations in improving reconstruction quality when the amount of leaked data is minimal, aligning with the common belief regarding overfitting: less data leads to higher overfitting.

When the percentage of leaked encrypted data is higher, such as 10% and 70%, the performance improvements from ablation studies are relatively smaller but still notable. GAN-based training continues to enhance performance metrics. For instance, at 70% leaked data, the TPR increases from 0.860 to 0.929, and the mse decreases from 0.193 to 0.063. This improvement is attributed to the increased amount of available data, which helps the reconstruction quality approach that of the original key model, thereby mitigating overfitting issues. Consequently, as more data become available, the model benefits from better generalization, leading to enhanced reconstruction fidelity.

While the GAN-based training significantly improves metrics such as TPR, mse, LPIPS, PSNR, and SSIM, a noticeable decline in the CLIP score is observed in Table IV, especially for the results with few key-sharing images. This decrease can be explained by the inherent objective of the GAN loss. The GAN loss encourages the reconstructed images to align with the statistical distribution of a general third-party dataset, which promotes the generation of realistic human-like images. However, this generalization can reduce the specific alignment with the ground truth images (consider color/gray-scale images), as measured by the CLIP score, which evaluates semantic similarity. For example, as shown in Fig. 5, our method with GAN loss always produces color images, even if the original ones are gray-scale images.

TABLE V

ACCURACY AND QUALITY MEASUREMENTS OF THE IMAGES RECONSTRUCTED BY THE ORIGINAL KEY MODEL, PROPOSED APPROACH, AND STYLEGAN-BASED METHOD

	Methods		
	Original Key	Ours	StyleGAN
TPR ↑	$0.971^{\pm 0.000}$	$\mathbf{0.924^{\pm 0.011}}$	$0.798^{\pm 0.014}$
MSE ↓	$0.022^{\pm 0.000}$	$\mathbf{0.063^{\pm 0.008}}$	$0.391^{\pm 0.003}$
LPIPS ↓	$0.367^{\pm 0.000}$	$\mathbf{0.305^{\pm 0.017}}$	$0.420^{\pm 0.001}$
PSNR ↑	$17.525^{\pm 0.000}$	$\mathbf{12.464^{\pm 0.593}}$	$4.454^{\pm 0.027}$
CLIP ↑	$0.837^{\pm 0.000}$	$0.811^{\pm 0.009}$	$\mathbf{0.814^{\pm 0.001}}$
SSIM ↑	$0.302^{\pm 0.000}$	$\mathbf{0.298^{\pm 0.006}}$	$0.154^{\pm 0.001}$

E. Comparison With StyleGAN-Based Attacker

This section analyzes the performance comparison with the StyleGAN-based reconstruction attack. As a baseline method, we revise the reconstruction attack method proposed in [39] suitable for our scenario, where this method aims to find latent vectors in the latent space of the StyleGAN [40] matching with the given facial feature vectors. In this experiment, we assume that the proposed method can access 70% of the encrypted images.

Table V summarizes the accuracy and quality measurements of the images reconstructed by the original key model, proposed method, and StyleGAN-based method. Most importantly, the proposed method outperforms the StyleGAN-based method across most metrics. This is because the proposed method aims to reconstruct the original key model rather than finding images with similar feature vectors. Interestingly, the proposed method and the StyleGAN-based method have similar CLIP scores. In our opinion, the CLIP score concentrates on semantic similarity rather than structural similarity. Thus, it gives a high score if two images are semantically similar such as hairstyle, age, gender, and race.

Fig. 8 visually illustrates the qualitative comparisons between the original images, our reconstructed images, and StyleGAN-based reconstructions. Visually, our method preserves texture consistency better than the StyleGAN-based

Fig. 8. Graphical examples of the original images, proposed method, and StyleGAN-based method. The images in the first row with green borderlines show the original image. The images in the second and third rows show the reconstructed images using the proposed and StyleGAN-based methods, respectively.

method. On the other hand, the StyleGAN-based method, although exhibiting smoother textures with similar facial features, often deviates significantly from the original structure, resulting in worse pixel-wise or structural metrics in Table V.

VI. Discussion

A. Conclusion

This study investigates the potential vulnerabilities of the AVIH method [1] by proposing a DR attack, highlighting the need for additional privacy protection methods in online image recognition systems. Our findings emphasize that if 1% of the gallery dataset shares the same key model, the key model's functionality can be reconstructed, leading to a successful DR attack.

B. Limitations and Extensibility

Although we implement our method for the AVIH method [1], it can be extended to other cloud-based machine-learning systems where neural network-based key models are used for reconstructing original data. Since the work in [1] was recently published, few follow-up papers have appeared. However, we believe our method can be extended to all future works related to ML-based cloud-based systems.

C. Future Research Direction

Here, we discuss the defense method against our work. One might consider assigning a unique neural network key model to each gallery image; however, this is extremely memory inefficient. Instead, we could assign an additional key image to each gallery image, where the original image can be reconstructed only when the key image and key model match exactly. If these images do not match, the reconstructed image would be another natural image. None of the previous studies have proposed a defense method like this; however, since this is beyond the scope of our work, we leave this for future research.

Appendix A
Additional Experimental Results on Vehicle and Person Re-Identification

A. Implementation Details

1) Person and Vehicle Re-Identification Scenario: For image recognition tasks other than face images, we target to reconstruct the original images for the gallery set of the following datasets: vehicle re-identification dataset (VeRi [41]) and pedestrian re-identification dataset (Market-1501 [42]). For both datasets, the service model is chosen as TransReID model [29] with ViT backbone [43]. For the auxiliary datasets, we use the Stanford Car [44] dataset and LPW [45] dataset for vehicle/human images, respectively. The accuracy on the re-identification is measured by mAP, rank-1, and rank-5 accuracy, where these metrics are evaluated by the TransReID model with the DeiT backbone [46].

2) Implementation of AVIH and Our DR Attack: The implementation details of the AVIH and the proposed method for re-identification tasks are similar to those in the face recognition experiment. The difference is that the VC loss kernel size is configured as 8 for the vehicle dataset, and our key model is trained for 800 steps.

TABLE VI
Accuracy Measurements of the Original Gallery Images, Protected Images, Reconstructed Images by Original Key Model, and Our DR Attack Results, Where the Datasets are VeRi (Vehicle) and Market-1501 (Person) Datasets

		Rank-1 ↑	Rank-5 ↑	mAP ↑
Market-1501 (Person)	Ours (1%)	$0.015^{\pm 0.015}$	$0.045^{\pm 0.040}$	$0.025^{\pm 0.019}$
	Ours (3%)	$0.147^{\pm 0.033}$	$0.296^{\pm 0.061}$	$0.151^{\pm 0.033}$
	Ours (10%)	$0.429^{\pm 0.147}$	$0.637^{\pm 0.212}$	$0.389^{\pm 0.130}$
	Ours (70%)	$\mathbf{0.772}^{\pm \mathbf{0.013}}$	$\mathbf{0.905}^{\pm \mathbf{0.012}}$	$\mathbf{0.685}^{\pm \mathbf{0.015}}$
	Original	$0.912^{\pm 0.000}$	$0.969^{\pm 0.000}$	$0.830^{\pm 0.000}$
	Protected	$0.002^{\pm 0.000}$	$0.039^{\pm 0.000}$	$0.020^{\pm 0.000}$
	Key model	$0.881^{\pm 0.000}$	$0.953^{\pm 0.000}$	$0.797^{\pm 0.000}$
VeRi (Vehicle)	Ours (1%)	$0.020^{\pm 0.009}$	$0.045^{\pm 0.015}$	$0.031^{\pm 0.008}$
	Ours (3%)	$0.052^{\pm 0.030}$	$0.109^{\pm 0.059}$	$0.060^{\pm 0.028}$
	Ours (10%)	$0.195^{\pm 0.137}$	$0.327^{\pm 0.217}$	$0.163^{\pm 0.108}$
	Ours (70%)	$\mathbf{0.520}^{\pm \mathbf{0.064}}$	$\mathbf{0.740}^{\pm \mathbf{0.055}}$	$\mathbf{0.406}^{\pm \mathbf{0.044}}$
	Original	$0.900^{\pm 0.000}$	$0.976^{\pm 0.000}$	$0.719^{\pm 0.000}$
	Protected	$0.018^{\pm 0.000}$	$0.074^{\pm 0.000}$	$0.032^{\pm 0.000}$
	Key model	$0.707^{\pm 0.000}$	$0.874^{\pm 0.000}$	$0.529^{\pm 0.000}$

B. Accuracy Metrics

Table VI presents the accuracy measurements for the original gallery images, protected images, and reconstructed images by the original key model, and our DR attack results using the TransReID method with ViT backbone. The evaluations are conducted on two datasets: Market-1501 (Person) and VeRi (Vehicle).

For both datasets, all the accuracy matrices are enhanced as more gallery images share a common key model. For instance, in market-1501 results, if only 1% of images share the same key model, the performance is quite poor, with a Rank-1 accuracy of 0.015, Rank-5 accuracy of 0.045, and mAP of 0.025. As the percentage of shared key models increases to 3%, the performance improves significantly, with Rank-1 increasing to 0.147, Rank-5 to 0.296, and mAP to 0.151. When 70% of the images share the same key model, the reconstructed images achieve near-original performance with Rank-1 at 0.772, Rank-5 at 0.905, and mAP at 0.685.

Compared to the results in face recognition, the accuracy results are not good enough; however, as will be discussed later in Appendix A–C, the perceptual similarity of our proposed method closely achieves the images reconstructed by the private key model.

TABLE VII

RECONSTRUCTION QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS

		MSE ↓	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
Market-1501	Ours (1%)	$0.221^{\pm 0.166}$	$0.709^{\pm 0.055}$	$8.031^{\pm 3.371}$	$0.828^{\pm 0.012}$	$0.216^{\pm 0.055}$
	Ours (3%)	$0.078^{\pm 0.009}$	$0.557^{\pm 0.023}$	$11.170^{\pm 0.507}$	$0.853^{\pm 0.009}$	$0.296^{\pm 0.013}$
	Ours (10%)	$0.082^{\pm 0.123}$	$0.471^{\pm 0.088}$	$12.984^{\pm 3.138}$	$0.876^{\pm 0.012}$	$0.315^{\pm 0.056}$
	Ours (70%)	$\mathbf{0.033}^{\pm \mathbf{0.005}}$	$\mathbf{0.399}^{\pm \mathbf{0.031}}$	$\mathbf{14.986}^{\pm \mathbf{0.649}}$	$\mathbf{0.891}^{\pm \mathbf{0.007}}$	$\mathbf{0.351}^{\pm \mathbf{0.019}}$
	Original	$0.000^{\pm 0.000}$	$0.009^{\pm 0.000}$	$36.831^{\pm 0.000}$	$0.980^{\pm 0.000}$	$0.492^{\pm 0.000}$
	Protected	$0.409^{\pm 0.000}$	$1.438^{\pm 0.000}$	$3.945^{\pm 0.000}$	$0.571^{\pm 0.000}$	$0.015^{\pm 0.000}$
	Key model	$0.003^{\pm 0.000}$	$0.408^{\pm 0.000}$	$25.447^{\pm 0.000}$	$0.927^{\pm 0.000}$	$0.422^{\pm 0.000}$
VeRi	Ours (1%)	$0.288^{\pm 0.184}$	$0.787^{\pm 0.026}$	$6.434^{\pm 2.743}$	$0.715^{\pm 0.003}$	$0.167^{\pm 0.028}$
	Ours (3%)	$0.228^{\pm 0.210}$	$0.720^{\pm 0.048}$	$7.864^{\pm 3.032}$	$0.731^{\pm 0.007}$	$0.201^{\pm 0.034}$
	Ours (10%)	$0.297^{\pm 0.253}$	$0.668^{\pm 0.080}$	$7.209^{\pm 3.850}$	$0.751^{\pm 0.027}$	$0.215^{\pm 0.059}$
	Ours (70%)	$\mathbf{0.083}^{\pm \mathbf{0.022}}$	$\mathbf{0.598}^{\pm \mathbf{0.025}}$	$\mathbf{11.059}^{\pm \mathbf{1.079}}$	$\mathbf{0.793}^{\pm \mathbf{0.009}}$	$\mathbf{0.267}^{\pm \mathbf{0.014}}$
	Original	$0.000^{\pm 0.000}$	$0.010^{\pm 0.000}$	$34.907^{\pm 0.000}$	$0.975^{\pm 0.000}$	$0.487^{\pm 0.000}$
	Protected	$0.410^{\pm 0.000}$	$1.385^{\pm 0.000}$	$3.963^{\pm 0.000}$	$0.593^{\pm 0.000}$	$0.017^{\pm 0.000}$
	Key model	$0.026^{\pm 0.000}$	$0.759^{\pm 0.000}$	$16.172^{\pm 0.000}$	$0.867^{\pm 0.000}$	$0.276^{\pm 0.000}$

C. Similarity Metrics

In Table VII, the evaluation of the reconstructed images on both datasets is presented using perceptual similarity and pixel-based similarity metrics. For the Market-1501 dataset, the mse decreases as more images share a common key model, with the lowest mse of 0.033 for 70% shared key model images. Although pixel-based metrics like PSNR improve significantly from 8.031 (1%) to 14.986 (70%), perceptual similarity metrics such as LPIPS and CLIP also show significant enhancement, with LPIPS decreasing from 0.709 to 0.399 and CLIP increasing from 0.828 to 0.891.

Similarly, in the VeRi dataset, both perceptual and pixel-based similarity metrics show improvement. For instance, the PSNR increases from 6.434 (1%) to 11.059 (70%), while LPIPS decreases from 0.787 to 0.598 and CLIP increases from 0.715 to 0.793. These improvements in perceptual metrics indicate that the reconstructed images, even with a higher percentage of shared key models, maintain a high level of visual similarity to the original images.

These results suggest that while pixel-based metrics like PSNR and mse improve, perceptual similarity metrics such as LPIPS and CLIP also indicate significant enhancements in image quality even nearly achieving the original key model. This shows high image similarity and effective privacy protection as more images share the same key model.

D. Graphical Results

In Fig. 9, we present graphical examples of the proposed method. The first row with green borderlines shows the original images, while the subsequent rows display the reconstructed images with different percentages of leaked encrypted data (1%, 3%, 10%, and 70%).

With 1% leaked encrypted data, the reconstructed images are significantly distorted and blurred, making recognition difficult, which aligns with lower similarity scores. As the percentage increases to 3%, facial features become more distinguishable despite some blurring, showing noticeable improvement. At 10%, the images are clearer and more recognizable, and at 70%, the reconstructed images are very close to the original quality, supporting the highest similarity scores and demonstrating the method's effectiveness.

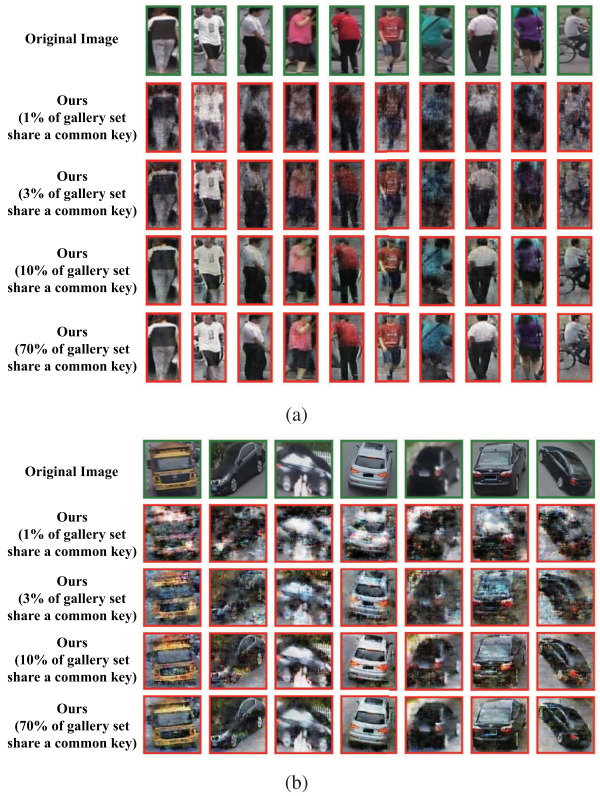

Fig. 9. Examples of the reconstructed images by the proposed method. The index of each image is randomly chosen. The images with the green-rectangular borderline are the original images. In the re-identification task, we evaluate two datasets. (a) Market-1501 dataset and (b) VeRi Dataset.

TABLE VIII

RECONSTRUCTION QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS

	TPR ↑	MSE ↓	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
Ours (1%)	$0.540^{\pm 0.063}$	$0.118^{\pm 0.006}$	$0.678^{\pm 0.015}$	$9.521^{\pm 0.236}$	$0.726^{\pm 0.004}$	$0.267^{\pm 0.005}$
Ours (3%)	$0.525^{\pm 0.057}$	$\mathbf{0.115}^{\pm \mathbf{0.009}}$	$0.636^{\pm 0.009}$	$\mathbf{9.621}^{\pm \mathbf{0.303}}$	$0.732^{\pm 0.005}$	$0.271^{\pm 0.005}$
Ours (10%)	$0.663^{\pm 0.068}$	$0.124^{\pm 0.008}$	$0.568^{\pm 0.014}$	$9.355^{\pm 0.295}$	$0.754^{\pm 0.010}$	$\mathbf{0.281}^{\pm \mathbf{0.006}}$
Ours (70%)	$\mathbf{0.796}^{\pm \mathbf{0.157}}$	$0.131^{\pm 0.079}$	$\mathbf{0.555}^{\pm \mathbf{0.047}}$	$9.441^{\pm 1.625}$	$\mathbf{0.799}^{\pm \mathbf{0.014}}$	$0.280^{\pm 0.026}$
Original	$0.977^{\pm 0.000}$	$0.001^{\pm 0.000}$	$0.032^{\pm 0.000}$	$33.375^{\pm 0.000}$	$0.979^{\pm 0.000}$	$0.483^{\pm 0.000}$
Protected	$0.171^{\pm 0.000}$	$0.495^{\pm 0.000}$	$1.362^{\pm 0.000}$	$3.121^{\pm 0.000}$	$0.665^{\pm 0.000}$	$0.014^{\pm 0.000}$
Key model	$0.979^{\pm 0.000}$	$0.004^{\pm 0.000}$	$0.478^{\pm 0.000}$	$23.984^{\pm 0.000}$	$0.909^{\pm 0.000}$	$0.366^{\pm 0.000}$

APPENDIX B
ADDITIONAL EXPERIMENTAL RESULTS ON LARGER FACE IMAGES

In this section, we utilize the AgeDB-30 dataset to evaluate the effectiveness of the proposed method on larger size data. To this end, we resize the AgeDB-30 dataset to 224×224 pixels, which is named AgeDB-30-L. In this experiment, we use the AdaFace face recognition model with the IR-18 backbone network.

A. Accuracy and Similarity Metrics

In Table VIII, we show the accuracy and similarity metrics for the AgeDB-30-L dataset. Compared to results of Table IV, the quality of reconstruction attacks is slightly degraded because there are more uncertainties in finding the key model with the same number of given encrypted samples.

Fig. 10. Examples of the reconstructed images by the proposed method for the AgeDB-30-L dataset experiments. The index of each image is randomly chosen. The images with the green rectangular borderline are the original images.

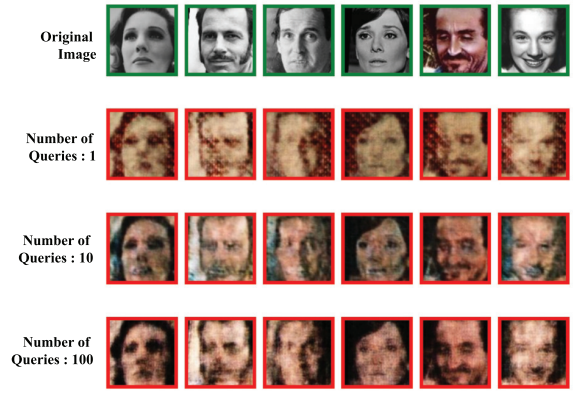

Fig. 11. (Black box) Examples of the reconstructed images by the proposed method for the AgeDB-30 dataset experiments. The index of each image is randomly chosen. The images with the green rectangular borderline are the original images.

B. Graphical Results

In addition to the numerical results provided in Table VIII, we also show the graphical results for the AgeDB-30-L dataset. In Fig. 10, we depict the images reconstructed by the proposed method for various numbers of key-sharing data. As shown in the figure, the quality of the reconstructed images is enhanced as more data share the same key model. By doing this experiment, we can show that the proposed method works well for larger face image datasets. For further verification, please refer to the results of the other datasets such as the VeRi vehicle dataset and the market-1501 dataset.

APPENDIX C
BLACKBOX ATTACKS

This section presents the performance of our approach on the *black box* setting, where only the outputs of the service DNN are available to the attacker. In this scenario, we estimate the gradient of the service model via finite-difference method [47], which is generally used in zeroth-order optimization methods. Let us define the confidence level of the image $\mathbf{x}$ computed by the service DNN as $s(x)$ (in our method, we use cosine similarity). Then, the estimated gradient can be obtained by

$$\tilde{g}(x) = \frac{1}{N_{\mathrm{FD}}} \sum_{i=1}^{N_{\mathrm{FD}}} \frac{s(x + \alpha_{\mathrm{FD}} n_i) - s(x)}{\alpha_{\mathrm{FD}}} \tag{14}$$

where N_{FD} and α_{FD} denote the number of queries and the perturbation coefficient in the finite-difference-based gradient estimation, respectively. With the estimated gradient in (14), the identity loss in (6) can be replaced by

$$L_{\mathrm{I}}(\tilde{\mathcal{G}}) = \mathbb{E}_{x' \sim \tilde{\mathcal{G}}} \| \mathrm{sg}(x') + \tilde{g}(x') - x' \|_2 \tag{15}$$

where $\mathrm{sg}(x')$ denotes the image copied from x' without gradient propagation (in PyTorch, "`x.detach().copy()`").

Discussion of the Results: In Table IX and Fig. 11, we demonstrate the effectiveness of our method in a black box setting, where only query-based access to the service DNN is available. In Table IX, the results show that the reconstruction performance is enhanced as the number of queries N_{FD} increases. This is because the estimated gradient in (14) is getting more accurate as N_{FD} increases. For example, at $N_{\mathrm{FD}} = 1$, the TPR is 0.098, and the mse is 0.306, while at $N_{\mathrm{FD}} = 100$, the TPR increases to 0.506, and the mse decreases to 0.202. In graphical results in Fig. 11, increasing N_{FD} results in reconstructed images that are progressively close to the original images.

TABLE IX
(BLACK BOX) THE ACCURACY AND RECONSTRUCTION QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS BY VARYING THE NUMBER OF QUERIES N_{FD}

	TPR ↑	MSE ↓	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
Ours (1)	$0.098^{\pm 0.126}$	$0.306^{\pm 0.066}$	$0.604^{\pm 0.030}$	$5.619^{\pm 1.085}$	$0.699^{\pm 0.022}$	$0.168^{\pm 0.019}$
Ours (10)	$0.387^{\pm 0.256}$	$0.226^{\pm 0.134}$	$\mathbf{0.565^{\pm 0.058}}$	$7.464^{\pm 2.213}$	$0.709^{\pm 0.025}$	$0.204^{\pm 0.038}$
Ours (100)	$\mathbf{0.506^{\pm 0.210}}$	$\mathbf{0.202^{\pm 0.144}}$	$0.569^{\pm 0.057}$	$\mathbf{8.040^{\pm 2.365}}$	$\mathbf{0.726^{\pm 0.020}}$	$\mathbf{0.213^{\pm 0.037}}$
White-box	$0.761^{\pm 0.254}$	$0.121^{\pm 0.149}$	$0.401^{\pm 0.069}$	$10.996^{\pm 2.766}$	$0.774^{\pm 0.017}$	$0.269^{\pm 0.046}$
Original	$0.977^{\pm 0.000}$	$0.001^{\pm 0.000}$	$0.032^{\pm 0.000}$	$33.375^{\pm 0.000}$	$0.979^{\pm 0.000}$	$0.483^{\pm 0.000}$
Protected	$0.171^{\pm 0.000}$	$0.495^{\pm 0.000}$	$1.362^{\pm 0.000}$	$3.121^{\pm 0.000}$	$0.665^{\pm 0.000}$	$0.014^{\pm 0.000}$
Key model	$0.979^{\pm 0.000}$	$0.004^{\pm 0.000}$	$0.478^{\pm 0.000}$	$23.984^{\pm 0.000}$	$0.909^{\pm 0.000}$	$0.366^{\pm 0.000}$

To summarize, the results demonstrate the feasibility and effectiveness of our proposed method even in a black box setting, where only query-based access to the target model is available. The results highlight the versatility of our approach, as it does not rely on white box assumptions, making it applicable to a wider range of real-world scenarios.

APPENDIX D
VC LOSS TRADEOFF

In this section, we present additional experiments to validate the effect of the VC loss weights used in the AVIH encryption on the reconstruction quality of the proposed method. The weight λ_2 of the VC loss in (1) plays as a key parameter in the AVIH method, balancing variance consistency and leading a tradeoff between the quality of encryption and the recovery process.

A. Accuracy and Similarity Metrics

Table X indicates the impact of varying the VC loss weight λ_2 on both encryption quality and reconstruction quality. As observed in the table, increasing λ_2 leads to a noticeable decline in the quality of images reconstructed by *the original key model*. For example, with $\lambda_2 = 0.3$, the key model achieves an LPIPS of 0.683, while with $\lambda_2 = 3.0$, the LPIPS is significantly degraded to 1.057. Since our attack paradigm relies on

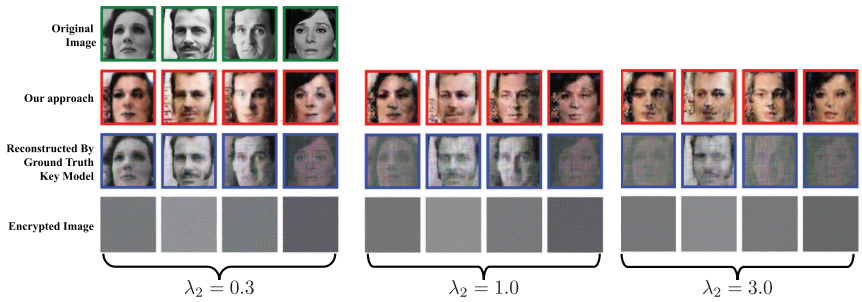

Fig. 12. (VC Loss Tradeoff) Examples of the reconstructed images by the proposed method for the AgeDB-30 dataset. The index of each image is randomly chosen. The images with the green rectangular borderline are the original images. The red and blue rectangular borderline images denote the reconstructed images via our method and the original key model, respectively.

TABLE X

(VC LOSS TRADEOFF) THE ACCURACY AND QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY THE ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS. THE USED DATASET IS AGEDB-30, AND THE TARGET FACE RECOGNITION MODEL IS ADAFACE WITH THE IR-18 BACKBONE NETWORK MODEL. TO VERIFY THE TRADEOFF BETWEEN THE VC LOSS WEIGHT λ_2 AND THE RECONSTRUCTION QUALITY, WE IMPLEMENT THE EXPERIMENTS FOR $\lambda_2 \in \{0.3, 1.0, 3.0\}$

		TPR ↑	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
	Original	$0.980^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
$\lambda_2 = 0.3$	Ours (1%)	$0.428^{\pm 0.139}$	$0.589^{\pm 0.028}$	$9.157^{\pm 1.423}$	$0.717^{\pm 0.009}$	$0.222^{\pm 0.020}$
	Ours (3%)	$0.538^{\pm 0.102}$	$0.509^{\pm 0.037}$	$10.107^{\pm 2.042}$	$0.739^{\pm 0.012}$	$0.235^{\pm 0.027}$
	Ours (10%)	$0.593^{\pm 0.035}$	$0.400^{\pm 0.019}$	$10.902^{\pm 0.341}$	$0.776^{\pm 0.012}$	$0.252^{\pm 0.006}$
	Ours (70%)	$0.649^{\pm 0.042}$	$0.369^{\pm 0.013}$	$11.106^{\pm 0.342}$	$0.792^{\pm 0.015}$	$0.254^{\pm 0.004}$
	Key model	$0.877^{\pm 0.000}$	$0.638^{\pm 0.000}$	$14.533^{\pm 0.000}$	$0.810^{\pm 0.000}$	$0.229^{\pm 0.000}$
$\lambda_2 = 1.0$	Ours (1%)	$0.298^{\pm 0.141}$	$0.609^{\pm 0.027}$	$8.464^{\pm 1.895}$	$0.711^{\pm 0.010}$	$0.214^{\pm 0.026}$
	Ours (3%)	$0.411^{\pm 0.056}$	$0.528^{\pm 0.009}$	$10.517^{\pm 0.421}$	$0.734^{\pm 0.008}$	$0.236^{\pm 0.004}$
	Ours (10%)	$0.373^{\pm 0.140}$	$0.442^{\pm 0.043}$	$9.033^{\pm 2.376}$	$0.772^{\pm 0.014}$	$0.221^{\pm 0.028}$
	Ours (70%)	$0.461^{\pm 0.097}$	$0.396^{\pm 0.023}$	$9.939^{\pm 0.383}$	$0.787^{\pm 0.008}$	$0.230^{\pm 0.007}$
	Key model	$0.777^{\pm 0.000}$	$0.830^{\pm 0.000}$	$12.018^{\pm 0.000}$	$0.789^{\pm 0.000}$	$0.183^{\pm 0.000}$
$\lambda_2 = 3.0$	Ours (1%)	$0.190^{\pm 0.054}$	$0.628^{\pm 0.011}$	$9.042^{\pm 0.393}$	$0.707^{\pm 0.008}$	$0.210^{\pm 0.006}$
	Ours (3%)	$0.279^{\pm 0.083}$	$0.565^{\pm 0.018}$	$9.802^{\pm 0.396}$	$0.728^{\pm 0.009}$	$0.220^{\pm 0.006}$
	Ours (10%)	$0.170^{\pm 0.093}$	$0.447^{\pm 0.024}$	$9.116^{\pm 1.873}$	$0.769^{\pm 0.008}$	$0.215^{\pm 0.021}$
	Ours (70%)	$0.196^{\pm 0.092}$	$0.414^{\pm 0.022}$	$8.787^{\pm 1.428}$	$0.785^{\pm 0.020}$	$0.210^{\pm 0.016}$
	Key model	$0.494^{\pm 0.000}$	$1.057^{\pm 0.000}$	$8.525^{\pm 0.000}$	$0.760^{\pm 0.000}$	$0.135^{\pm 0.000}$

mimicking the key model, this reduction in the original key model also impacts the performance of our approach. This is evident in most of the metrics; for instance, the TPR decreases from 0.649 at $\lambda_2 = 0.3$ to 0.196 $\lambda_2 = 0.196$.

B. Graphical Results

Fig. 12 visually demonstrates the gradual decline in the reconstruction quality as λ_2 increases. This graphical representation underscores how higher VC loss weights result in more degraded reconstructions, both for our approach and the original key model. In the figure, while our method provides human-like images injected by the GAN loss, the details of the textures are degraded as λ_2 increases. Even for the original key model, it cannot provide reliable results as λ_2 increases. Therefore, using excessively high VC loss weights is impractical in scenarios where reconstruction fidelity is critical. Our proposed method is inherently tied to the reconstruction performance of the key model, making this tradeoff an important consideration in the overall design.

Fig. 13. (Augmented identity loss probability) Examples of the reconstructed images by the proposed method for the AgeDB-30 dataset experiments. To verify the impact of the augmented identity loss, we implement our method for $P_{\text{aug}} \in \{0.01, 0.03, 0.1, 0.3, 1.0\}$.

TABLE XI

(AUGMENTED IDENTITY LOSS PROBABILITY) THE ACCURACY AND QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY THE ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS. THE USED DATASET IS AGEDB-30, AND THE TARGET DNN MODEL IS ADAFACE WITH THE IR-18 BACKBONE. WE IMPLEMENT OUR METHOD FOR $P_{\text{AUG}} \in \{0.01, 0.03, 0.1, 0.3, 1.0\}$

		TPR ↑	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
	Original	$0.980^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
	Encrypted	$0.211^{\pm 0.000}$	$1.214^{\pm 0.000}$	$1.866^{\pm 0.000}$	$0.641^{\pm 0.000}$	$0.019^{\pm 0.000}$
	Key Model	$0.971^{\pm 0.000}$	$0.367^{\pm 0.000}$	$17.525^{\pm 0.000}$	$0.837^{\pm 0.000}$	$0.302^{\pm 0.000}$
Ours (1%)	$P_{\text{aug}} = 0.01$	$0.523^{\pm 0.052}$	$0.566^{\pm 0.015}$	$9.691^{\pm 0.499}$	$0.706^{\pm 0.010}$	$0.238^{\pm 0.007}$
	$P_{\text{aug}} = 0.03$	$0.591^{\pm 0.061}$	$0.553^{\pm 0.014}$	$10.070^{\pm 0.371}$	$0.711^{\pm 0.005}$	$0.247^{\pm 0.006}$
	$P_{\text{aug}} = 0.1$	$0.534^{\pm 0.189}$	$0.560^{\pm 0.041}$	$9.013^{\pm 1.652}$	$0.718^{\pm 0.013}$	$0.233^{\pm 0.025}$
	$P_{\text{aug}} = 0.3$	$0.598^{\pm 0.149}$	$0.543^{\pm 0.040}$	$9.501^{\pm 1.507}$	$0.718^{\pm 0.012}$	$0.241^{\pm 0.023}$
	$P_{\text{aug}} = 1.0$	$0.628^{\pm 0.058}$	$0.540^{\pm 0.008}$	$9.960^{\pm 0.268}$	$0.715^{\pm 0.006}$	$0.249^{\pm 0.003}$
Ours (3%)	$P_{\text{aug}} = 0.01$	$0.618^{\pm 0.080}$	$0.497^{\pm 0.013}$	$10.896^{\pm 0.320}$	$0.723^{\pm 0.010}$	$0.250^{\pm 0.008}$
	$P_{\text{aug}} = 0.03$	$0.686^{\pm 0.055}$	$0.484^{\pm 0.019}$	$11.397^{\pm 0.513}$	$0.733^{\pm 0.012}$	$0.260^{\pm 0.010}$
	$P_{\text{aug}} = 0.1$	$0.553^{\pm 0.269}$	$0.507^{\pm 0.071}$	$9.715^{\pm 3.377}$	$0.728^{\pm 0.014}$	$0.237^{\pm 0.051}$
	$P_{\text{aug}} = 0.3$	$0.712^{\pm 0.236}$	$0.461^{\pm 0.065}$	$10.987^{\pm 2.786}$	$0.738^{\pm 0.011}$	$0.262^{\pm 0.043}$
	$P_{\text{aug}} = 1.0$	$0.746^{\pm 0.043}$	$0.450^{\pm 0.018}$	$11.847^{\pm 0.394}$	$0.746^{\pm 0.013}$	$0.270^{\pm 0.005}$

APPENDIX E
VARIOUS AUGMENTED IDENTITY LOSS PROBABILITY

In this experiment, we show the effect of the augmented identity loss in (7). While the effectiveness of the augmented identity loss has been examined in Section V through an ablation study (Table IV), we conducted further experiments for an in-depth analysis. Specifically, we introduce the augmented identity loss probability P_{aug}, which determines the probability of applying augmented identity loss during training. In other words, we use the augmented identity loss in (7) with the probability P_{aug}, and use the standard identity loss in (6) with the probability $1 - P_{\text{aug}}$. The results shown in Table XI

TABLE XII

(WEIGHT ON THE AUGMENTED IDENTITY LOSS) THE ACCURACY AND QUALITY MEASUREMENTS OF THE ORIGINAL GALLERY IMAGES, PROTECTED IMAGES, RECONSTRUCTED IMAGES BY THE ORIGINAL KEY MODEL, AND OUR DR ATTACK RESULTS. THE USED DATASET IS AGEDB-30, AND THE TARGET DNN MODEL IS ADAFACE WITH THE IR-18 BACKBONE. WE IMPLEMENT OUR METHOD FOR $\lambda_1 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$

		TPR ↑	LPIPS ↓	PSNR ↑	CLIP ↑	SSIM ↑
	Original	$0.980^{\pm 0.000}$	$0.009^{\pm 0.000}$	$29.563^{\pm 0.000}$	$0.966^{\pm 0.000}$	$0.461^{\pm 0.000}$
	Encrypted	$0.211^{\pm 0.000}$	$1.214^{\pm 0.000}$	$1.866^{\pm 0.000}$	$0.641^{\pm 0.000}$	$0.019^{\pm 0.000}$
	Key Model	$0.971^{\pm 0.000}$	$0.367^{\pm 0.000}$	$17.525^{\pm 0.000}$	$0.837^{\pm 0.000}$	$0.302^{\pm 0.000}$
Ours (10%)	$\lambda_1 = 10^{-4}$	$0.389^{\pm 0.176}$	$0.501^{\pm 0.042}$	$8.536^{\pm 2.241}$	$0.749^{\pm 0.025}$	$0.211^{\pm 0.030}$
	$\lambda_1 = 10^{-3}$	$0.479^{\pm 0.152}$	$0.470^{\pm 0.059}$	$9.520^{\pm 1.547}$	$0.753^{\pm 0.017}$	$0.226^{\pm 0.025}$
	$\lambda_1 = 10^{-2}$	$0.745^{\pm 0.065}$	$0.412^{\pm 0.042}$	$10.740^{\pm 1.138}$	$0.769^{\pm 0.023}$	$0.257^{\pm 0.018}$
	$\lambda_1 = 10^{-1}$	$0.787^{\pm 0.188}$	$0.379^{\pm 0.042}$	$11.499^{\pm 1.619}$	$0.778^{\pm 0.020}$	$0.278^{\pm 0.024}$
	$\lambda_1 = 10^{0}$	$0.789^{\pm 0.196}$	$\mathbf{0.370^{\pm 0.054}}$	$11.237^{\pm 2.087}$	$0.779^{\pm 0.014}$	$0.274^{\pm 0.032}$
	$\lambda_1 = 10^{1}$	$\mathbf{0.812^{\pm 0.021}}$	$0.407^{\pm 0.007}$	$\mathbf{12.045^{\pm 0.344}}$	$\mathbf{0.792^{\pm 0.014}}$	$\mathbf{0.283^{\pm 0.008}}$

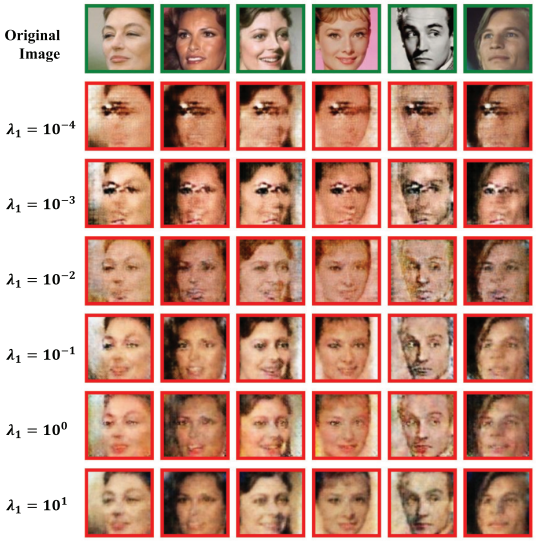

Fig. 14. (Weight on the Augmented Identity Loss) Examples of the reconstructed images by the proposed method for the AgeDB-30 dataset experiments. To verify the impact of the augmented identity loss weight, we implement our method for $\lambda_1 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$.

demonstrate that the proposed method effectively leverages the augmented identity loss. By systematically adjusting P_{aug}, we observe a clear improvement in reconstruction quality, as reflected in both quantitative metrics and visual examples in Fig. 13.

APPENDIX F
WEIGHTS ON THE IDENTITY LOSS

In this section, we further analyze the effect of the augmented identity loss weights on the reconstruction results. To this end, we implement our method for various weights λ_1 in (12), where $\lambda_1 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. In Table XII, we show the results of the proposed method for various values of λ_1. As shown in the table, the reconstruction performance of the proposed method is highly affected by the value of λ_1. With the very small values of λ_1 $(10^{-4}, 10^{-3})$, the loss function of our approach is rarely affected by the identity loss; hence, the reconstruction quality is relatively bad compared to the higher values of λ_1. As the value of λ_1 increases, the reconstruction quality is gradually enhanced, which is reflected in the visual examples in Fig. 14.

REFERENCES

[1] Z. Su, D. Zhou, N. Wang, D. Liu, Z. Wang, and X. Gao, "Hiding visual information via obfuscating adversarial perturbations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4333–4343.

[2] G. Chollet et al., "Privacy preserving personal assistant with on-device diarization and spoken dialogue system for home and beyond," 2024, *arXiv:2401.01146*.

[3] A. Singh and K. Chatterjee, "Cloud security issues and challenges: A survey," *J. Netw. Comput. Appl.*, vol. 79, pp. 88–115, Feb. 2017.

[4] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 657–672.

[5] T. Xiang, Y. Yang, H. Liu, and S. Guo, "Visual security evaluation of perceptually encrypted images based on image importance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4129–4142, Nov. 2020.

[6] S. Guo, T. Xiang, X. Li, and Y. Yang, "PEID: A perceptually encrypted image database for visual security evaluation," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1151–1163, 2020.

[7] E. Lee et al., "Low-complexity deep convolutional neural networks on fully homomorphic encryption using multiplexed parallel convolutions," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12403–12422.

[8] D. Kim and C. Guyot, "Optimized privacy-preserving CNN inference with fully homomorphic encryption," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2175–2187, 2023.

[9] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Comput. Surv.*, vol. 54, no. 11s, pp. 1–37, Jan. 2022.

[10] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. USENIX Conf. Secur. Symp.*, Aug. 2014, pp. 17–32.

[11] 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.

[12] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, "Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes," in *Proc. 15th Annu. Conf. Privacy, Secur. Trust (PST)*, Aug. 2017, pp. 115–11509.

[13] N.-B. Nguyen, K. Chandrasegaran, M. Abdollahzadeh, and N.-M. Cheung, "Re-thinking model inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16384–16393.

[14] Z. Tian, L. Cui, C. Zhang, S. Tan, S. Yu, and Y. Tian, "The role of class information in model inversion attacks against image deep learning classifiers," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 4, pp. 2407–2420, Jul. 2024.

[15] S. Zhou, T. Zhu, D. Ye, X. Yu, and W. Zhou, "Boosting model inversion attacks with adversarial examples," *IEEE Trans. Dependable Secur. Comput.*, vol. 21, no. 3, pp. 1451–1468, May 2023.

[16] J. Jang, H. Lyu, and H. J. Yang, "Patch-MI: Enhancing model inversion attacks via patch-based reconstruction," 2023, *arXiv:2312.07040*.

[17] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Jan. 2022, pp. 22911–22924.

[18] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 201–210.

[19] J.-W. Lee et al., "Privacy-preserving machine learning with fully homomorphic encryption for deep neural network," *IEEE Access*, vol. 10, pp. 30039–30054, 2022.

[20] Y. Ding et al., "DeepEDN: A deep-learning-based image encryption and decryption network for Internet of Medical Things," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1504–1518, Feb. 2021.

[21] 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.

[22] T. Chuman and H. Kiya, "A jigsaw puzzle solver-based attack on block-wise image encryption for privacy-preserving DNNs," *Proc. SPIE*, vol. 12592, pp. 335–340, Mar. 2023.

[23] A. MaungMaung and H. Kiya, "Generative model-based attack on learnable image encryption for privacy-preserving deep learning," 2023, *arXiv:2303.05036*.

[24] Y. Tang et al., "Watermarking vision-language pre-trained models for multi-modal embedding as a service," 2023, *arXiv:2311.05863*.
[25] T. Qiao et al., "A novel model watermarking for protecting generative adversarial network," *Comput. Secur.*, vol. 127, Apr. 2023, Art. no. 103102.
[26] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Dec. 2017, pp. 2066–2076.
[27] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[28] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality adaptive margin for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18750–18759.
[29] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15013–15022.
[30] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 1–9.
[31] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Workshop Faces 'Real-Life' Images: Detection, Alignment, Recognit., Univ. Massachusetts, Boston, MA, USA, Tech. Rep. inria-00321923, 2007.
[32] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1997–2005.
[33] S. Sengupta, J. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.
[34] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[37] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 1–15. [Online]. Available: https://aclanthology.org/2021.emnlp-main.595v2.pdf
[38] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.
[39] H. Otroshi Shahreza and S. Marcel, "Face reconstruction from facial templates by learning latent space of a generator network," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2024, pp. 1–18.
[40] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.
[41] X. Liu, W. Liu, T. Mei, and H. Ma, "PROVID: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 645–658, Mar. 2018.
[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[43] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2020, pp. 1–21.
[44] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.
[45] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proc. Assoc. Adv. Artif. Intell. (AAAI)*, Jan. 2017, pp. 7347–7354.
[46] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 1–11.
[47] Y. Nesterov and V. Spokoiny, "Random gradient-free minimization of convex functions," *Found. Comput. Math.*, vol. 17, pp. 527–566, Apr. 2017.

Jonggyu Jang (Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, in 2017 and 2021, respectively.

From March 2021 to February 2023, he was a Post-Doctoral Researcher at the Future IT Innovation Laboratory at Pohang University of Science and Technology (POSTECH) (mandatory mislitary service), Pohang, Republic of Korea. From March 2023 to August 2024, he was a Post-Doctoral Researcher at the Department of Electrical Engineering, POSTECH. From October 2024 to February 2025, he was a Post-Doctoral Researcher at the Institute of New Media and Communications, Seoul National University, Seoul, Republic of Korea. Since December 2024, he has been a Post-Doctoral Researcher at Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His fields of interests are theory and machine learning applications of wireless communications and communication networks.

Hyeonsu Lyu (Student Member, IEEE) received the B.S. degree in mathematical science and the M.S. degree in electrical engineering from Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea.

His research interests are developing next-generation wireless communication systems and designing AI- and robot-integrated systems in wireless networks.

Seongjin Hwang received the B.S. degree in electrical engineering from Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea, in 2022, where he is currently pursuing the Ph.D. degree with the School of Electrical Engineering.

His research interests include optimization theory for machine learning, and algorithms for AI-mediated communication.

Hyun Jong Yang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, in 2004, 2006, and 2010, respectively.

From August 2010 to August 2011, he was a Research Fellow at Korea Institute Ocean Science Technology (KIOST), Daejeon. From October 2011 to October 2012, he worked as a Post-Doctoral Researcher at the Electrical Engineering Department, Stanford University, Stanford, CA, USA. From October 2012 to August 2013, he was a Staff II Systems Design Engineer, Broadcom Corporation, Sunnyvale, CA, USA, where he developed physical-layer algorithms for LTE-AN MIMO receivers. In addition, he was a delegate of Broadcom in 3GPP standard meetings for RAN1 Rel-12 technologies. From September 2013 to July 2020, he was an Assistant/Associate Professor at the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea. From July 2020 to August 2024, he was an Associate Professor at the Department of Electrical Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea. Since September 2024, he has been an Associate Professor at the Department of Electrical And Computer Engineering, Seoul National University, Seoul, Republic of Korea. His fields of interests are privacy-preserving robot systems, deep-learning theory and algorithms, and signal processing.