arXiv:2409.01062v5 [cs.LG] 15 Jun 2026

Random Erasing vs. Model Inversion: A Promising Defense or a False Hope?

Viet-Hung Tran* *h.tran@qub.ac.uk*
Temasek Laboratories, Singapore University of Technology and Design; The Queen's University Belfast

Ngoc-Bao Nguyen* *thibaongoc_nguyen@sutd.edu.sg*
Temasek Laboratories, Singapore University of Technology and Design

Son T. Mai *thaison.mai@qub.ac.uk*
The Queen's University Belfast

Hans Vandierendonck *h.vandierendonck@qub.ac.uk*
The Queen's University Belfast

Ira Assent *ira@cs.au.dk*
Aarhus University

Alex Kot *eackot@ntu.edu.sg*
Nanyang Technological University (NTU); VinUniversity, Hanoi, Vietnam

Ngai-Man Cheung† *ngaiman_cheung@sutd.edu.sg*
Temasek Laboratories, Singapore University of Technology and Design

Abstract

Model Inversion (MI) attacks pose a significant privacy threat by reconstructing private training data from machine learning models. While existing defenses primarily concentrate on model-centric approaches, the impact of data on MI robustness remains largely unexplored. In this work, we explore *Random Erasing (RE)*—a technique traditionally used for improving model generalization under occlusion—and uncover its surprising effectiveness as a defense against MI attacks.

Specifically, our novel feature space analysis shows that model trained with RE-images introduces a significant discrepancy between the features of MI-reconstructed images and those of the private data. At the same time, features of private images remain distinct from other classes and well-separated from different classification regions. These effects collectively degrade MI reconstruction quality and attack accuracy while maintaining reasonable natural accuracy. Furthermore, we explore two critical properties of RE including Partial Erasure and Random Location. First, *Partial Erasure* prevents the model from observing entire objects during training, and we find that this has significant impact on MI, which aims to reconstruct the entire objects. Second, the *Random Location* of erasure plays a crucial role in achieving a strong privacy-utility trade-off. Our findings highlight RE as a simple yet effective defense mechanism that can be easily integrated with existing privacy-preserving techniques. Extensive experiments of 37 setups demonstrate that our method achieves SOTA performance in privacy-utility tradeoff. The results consistently demonstrate the superiority of our defense over existing defenses across different MI attacks, network architectures, and attack configurations. For the first time, we achieve significant degrade in attack accuracy *without* decrease in utility for some configurations. Our code and additional results are available at: `https://ngoc-nguyen-0.github.io/MIDRE/`

* The first two authors contributed equally. † Corresponding author.

1 Introduction

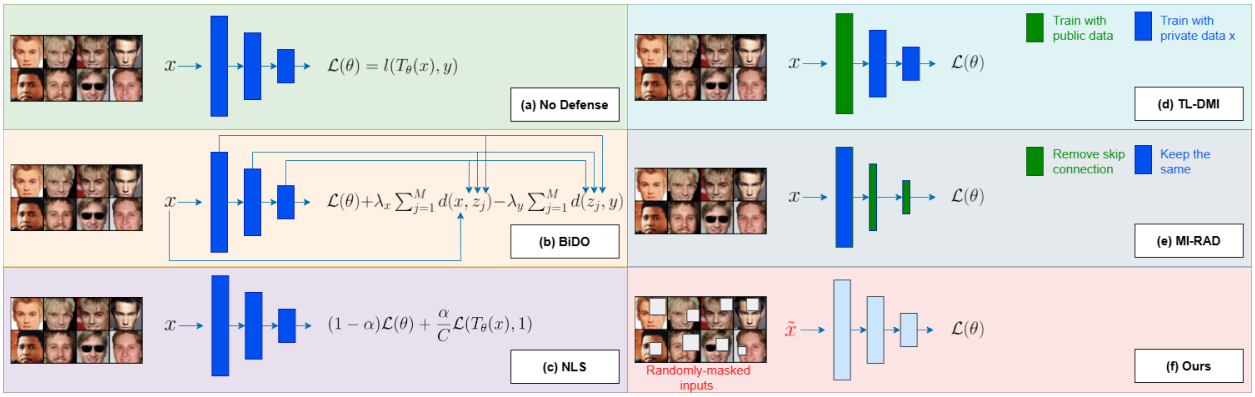


Figure 1: **Our Proposed Model Inversion (MI) Defense via Random Erasing (MIDRE).** (a) "No Defense": Training a model without MI defense. $\mathcal{L}(\theta)$ is the standard training loss, e.g., cross-entropy. Training a model with state-of-the-art MI defense (SOTA) (b) BiDO (Peng et al., 2022), (c) NLS (Struppek et al., 2024), and (d) TL-DMI (Ho et al., 2024), (e) MI-RAD (Koh et al., 2024) , (f) Our method. Studies in (Peng et al., 2022; Struppek et al., 2024) focus on **adding new loss** to the training objective in other to find the balance between model utility and privacy. Both TL-DMI (Ho et al., 2024) and MI-RAD (Koh et al., 2024) focus on **the model's parameters** to defend against MI. For our proposed method (f), the training procedure and objective are the same as that in (a) "No Defense". However, the training samples presented to the model are partially masked, thus, reducing private training sample's information encoded in the model and **preventing the model from observing the entire images**. Therefore, **MIDRE is different from other approaches and focuses on input data only to defend.** We find that this can significantly degrade MI attacks, which require substantial amount of private training data information encoded inside the model in order to reconstruct high-dimensional private images.

Machine learning and deep neural networks (DNNs) (LeCun et al., 2015) have demonstrated their utility across numerous domains, including computer vision (Voulodimos et al., 2018; O'Mahony et al., 2020), natural language processing (Otter et al., 2020), and speech recognition (Deng et al., 2013; Nassif et al., 2019). DNNs are now applied in critical areas such as medical diagnosis (Azad et al., 2021), medical imaging (Shen et al., 2017; Lundervold & Lundervold, 2019), facial recognition (Wang & Deng, 2021; Guo & Zhang, 2019; Masi et al., 2018), and surveillance (Zhou et al., 2021; Harikrishnan et al., 2019; Hashmi et al., 2021). However, the potential risks associated with the widespread deployment of DNNs raise significant concerns. In many practical applications, privacy violations involving DNNs can result in the leakage of sensitive and private data, eroding public trust in these applications. Defending against privacy violations of DNNs is of paramount importance.

One specific type of privacy violation is Model Inversion (MI) attacks on machine learning and DNN models. MI attacks aim to reconstruct private training data by exploiting access to machine learning models. Recent advancements in MI attacks including GMI (Zhang et al., 2020), KedMI (Chen et al., 2021), PPA (Struppek et al., 2022), MIRROR (An et al., 2022), IF-GMI (Qiu et al., 2024), PPDG-MI (Peng et al., 2024), PLG-MI (Yuan et al., 2023), and LOMMA (Nguyen et al., 2023) have achieved remarkable progress in attacking important face recognition models. This raises privacy concerns for models that are trained on sensitive data, such as face recognition, surveillance and medical diagnosis.

Related works. Existing MI defenses primarily focus on model-centric strategies like model gradients (Dwork, 2006; 2008), loss functions (Wang et al., 2021; Peng et al., 2022; Struppek et al., 2024), model features (Ho et al., 2024), and architecture designs (Koh et al., 2024) (see Fig. 1). Earlier works (Dwork, 2006; 2008) demonstrated the ineffectiveness of traditional Differential Privacy (DP) mechanisms against Model Inversion (MI) attacks. Recent research (Wang et al., 2021; Peng et al., 2022; Struppek et al., 2024) has explored the impact of loss functions on MI resilience. Wang et al. (2021) restricted the dependency between model inputs and outputs, while BiDO (Peng et al., 2022) focused on limiting the dependency

between model inputs and latent representations. To partially restore model utility, BiDO maximized the dependency between latent representations and outputs. Struppek et al. (2024) proposed using negative label smoothing factors as a defense. However, these loss function-based approaches often introduce conflicting objectives, leading to significant degradation in model utility. Recently, TL-DMI (Ho et al., 2024) restricted the number of layers to be encoded by the private training data, while MI-RAD (Koh et al., 2024) found that removing skip connections in final layers enhances robustness. However, both approaches experience difficulty in achieving competitive balance between utility and privacy.

Data augmentation, a technique that creates new, synthetic samples from existing data points, offers a promising avenue for enhancing model robustness. In this paper, we pioneer the investigation of Random Erasing (RE) (Zhong et al., 2020) for MI defense. RE, traditionally used to improve model generalization for detecting occluded objects by removing randomly a region in training samples, demonstrates its effectiveness as a powerful defense against MI attacks. In MI attacks, adversaries optimize reconstructed images to align with the target model's feature space representation of training samples. As will be shown in our novel analysis, thanks to RE, the target model's feature representations are inherently biased towards the RE-private images, the training data, rather than the private data. Consequently, **RE creates a discrepancy between the features of MI-reconstructed images and that of private images**, resulting in degraded MI attacks. Meanwhile, features of private images remain distinct from other classes, maintaining reasonable natural accuracy. Furthermore, we highlight **two crucial properties of RE that serve as an effective MI defense: *Partial Erasure* and *Random Location***. On the one hand, *Partial Erasure significantly reduces the amount of private information embedded in the training data, preventing the model from observing the entire image*, and consequently degrades the MI attacks. On the other hand, *Random Location improves the diversity of training data*, thereby, enhances the model utility. Our proposed method leads to substantial degradation in MI reconstruction quality and attack accuracy (See Sec. 3 for our comprehensive analysis and validation). Meanwhile, natural accuracy of the model is only moderately affected. Overall, we can achieve state-of-the-art performance in privacy-utility trade-offs as demonstrated in our extensive experiments of 37 setups – 9 SOTA MI attacks including both white-box, black-box, and label-only MI attacks, 11 model architectures (including vision transformer), 6 datasets and different resolution including 64×64, 160×160, and 224×224 resolution – and user study (in Supp.). Our contributions are:

- Our novel defense method, Model Inversion Defense via Random Erasing (MIDRE), is the first work to consider the well-known RE technique as a privacy protection mechanism, leveraging its powerful ability to reduce MI attack accuracy while maintaining model utility. All results support the SOTA effectiveness of a simple technique in addressing a critical security concern.
- We provide a deep understanding on feature space analysis of Random Erasing's defense effectiveness which leads to reduce of MI attacks in MIDRE model.
- Our analysis investigates two crucial properties of RE that serve as an effective MI defense: Partial Erasure and Random Location. With these two properties, our defense method degrades the attack accuracy while the impact on model utility is small.
- We conduct extensive experiments (Sec. 4, Supp.) and user study (Supp.) to demonstrate that our MIDRE can achieve SOTA privacy-utility trade-offs. Notably, in the high-resolution setting, our MIDRE is the first to achieve competitive MI robustness without sacrificing natural accuracy. Note that our method is very simple to implement and is complementary to existing MI defense methods.

2 Our Approach: Model Inversion Defense via Random Erasing (MIDRE)

2.1 Model inversion

A model inversion (MI) attack aims to reconstruct private training data from a trained machine learning model. The model under attack is called a *target model*, T_θ. The target model T_θ is trained on a private dataset $\mathcal{D}_{priv} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the private, sensitive data and y_i represents the corresponding ground truth label. For example, T_θ could be a face recognition model, and x_i is a face image of

an identity. The model is trained with standard loss function ℓ that penalizes the difference between model output $T_\theta(x)$ and y.

MI attacks. The underlying idea of MI attacks is to seek a reconstruction x that achieves maximum likelihood for a label y under T_θ:

$$\max_x \mathcal{P}(y; x, T_\theta) \tag{1}$$

In addition, some prior to improve reconstructed image quality can be included (Zhang et al., 2020; Chen et al., 2021). SOTA MI attacks (Zhang et al., 2020; Chen et al., 2021; Nguyen et al., 2023; Struppek et al., 2022) also apply GAN trained on a public dataset $\mathcal{D}_{pub}$ to limit the search space for x. $\mathcal{D}_{pub}$ has no identity intersection with $\mathcal{D}_{priv}$, assuming attackers can not access to any private samples.

2.2 Random Erasing (RE) as a defense

Random Erasing (RE) (Zhong et al., 2020) involves employing a random selection process to identify an region inside an image. Subsequently, this region is altered through the application of designated pixel values, such as zero or the mean value obtained from the dataset, resulting in *partial masking* of the image. Traditionally, RE is applied as a data augmentation technique to improve robustness of machine learning models in the presence of object occlusion (Zhong et al., 2020).

We propose a simple configuration of RE as a MI defense, requiring only one hyper-parameter. Given a training sample x with dimensions $W \times H$, we propose a square region erasing strategy to restrict private information leakage from x. We initiate by randomly selecting a starting point, denoted as (x_e, y_e), within the bounds of x. Next, we randomly select the erased area portion a_e within the specified range of $[0.1, a_h]$, guaranteeing at least 10% of x is erased during training, while a_h is the only hyper-parameter of our defense. The size of the erased region is $\sqrt{s_{RE}} \times \sqrt{s_{RE}}$ where $s_{RE} = W \times H \times a_e$ is the erased region. With the designated region, we determine the coordinates of the erased region $(x_e, y_e, x_e + \sqrt{s_{RE}}, y_e + \sqrt{s_{RE}})$. However, we need to ensure this selected region stays entirely within the boundaries of x, i.e. $x_e + \sqrt{s_{RE}} \leq W$, $y_e + \sqrt{s_{RE}} \leq H$. If the erased region extends beyond the image width or height, we simply repeat the selection process until we find a suitable square erased region that fits perfectly within x. We fill the erased regions with ImageNet mean pixel value (See Supp for a detailed discussion on the impact of the erased value) to obtain the RE-image. Note that RE is applied to all private training samples and the size and position vary each training iteration. We depict our method in Algorithm 1 (Supp.)

3 Analysis of privacy effect of MIDRE

In this section, we analyze the privacy impact of RE within our proposed MIDRE framework. We conduct a thorough analysis and demonstrate that RE can achieve a surprisingly state-of-the-art balance between utility and privacy. Specifically, when employed as a defense against MI attacks, RE is the first method to significantly reduce attack accuracy without compromising utility in certain configurations, whereas all prior MI defenses exhibit noticeable degradation in utility to achieve similar reductions in attack success. Experimental results in Sec. 4 further validate this finding.

We delve deeper into the mechanisms that underpin the effectiveness of RE. Importantly, we conduct a feature space analysis to explain RE's defense effectiveness, showing that model trained with MIDRE leads to a discrepancy between the features of MI-reconstructed images and that of private images, resulting in degrading of attack accuracy. At the same time, private images remain distinct from other classes and well-separated from different classification regions, maintaining reasonable natural accuracy. Furthermore, our analysis reveals that partial erasure, as implemented in RE, is a highly effective method for mitigating MI attacks. Particularly, to present the model with less private pixels during training, our approach of applying partial erasure while maintaining the original number of training epochs proves to be more effective than the alternative approach of reducing the number of epochs without using partial erasure. We attribute this to the fact that MI attacks rely on the target model to reconstruct the **entire image**, and RE's partial erasure prevents the target model from ever fully observing the entire image throughout the training process. Additionally, we show that applying partial erasure at random locations, as is done in RE, is more effective than erasure at fixed locations.

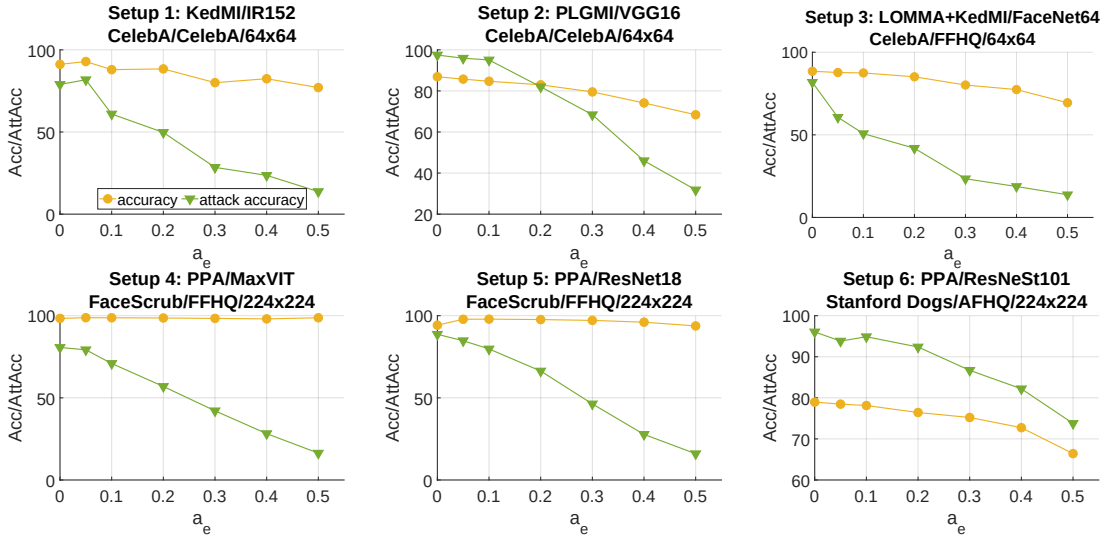
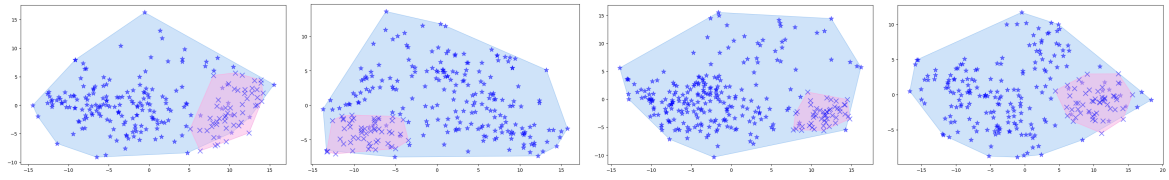


Figure 2: **Our analysis shows that Random Erasing (RE) can lead to substantial degradation in MI reconstruction quality and attack accuracy, while natural accuracy of the model is only moderately affected.** In this analysis, we conduct 6 experimental setups with different *MI attacks/target models architecture/private/public datasets/image resolution*. We analyze the attack (green line) and natural accuracy (orange line) of the target models under different extents of random erasing applied in the training stage, using random erasing ratio $a_e = s_{RE}/(W \times H)$ as discussed in Sec. 2.2. To properly reconstruct private high-dimensional facial images of individuals, MI attacks require significant amount of private training data information encoded inside the model. We find that the model using RE by small percentages can significantly degrade MI attacks, with MI attack accuracy decreasing, for example, from 15.47% to 39.93%. However, the natural accuracy of the model only decreases slightly, less than 4%, as sufficient information remained in the RE-images for the model to learn to discriminate between individuals (Setup 1-3). We also observe a high degradation in MI attack accuracy while the model accuracy increases. For instance, model accuracy increases by 0.37%, while attack accuracy decreases by 69.39% (Setup 4). Overall, our defense demonstrates state-of-the-art privacy-utility trade-offs and can improve model utility in certain setups

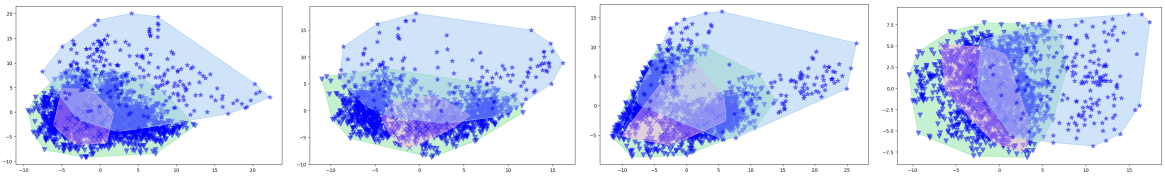
3.1 RE degrades MI significantly, achieving SOTA privacy-utility trade-off

In the analysis, we study attack accuracy and natural accuracy of a target model T_θ under different erased region portions a_e. Recall $a_e = s_{RE}/(W \times H)$, and $\sqrt{s_{RE}} \times \sqrt{s_{RE}}$ is the dimension of the erased region. For the target model, which is a face recognition model, in each setup, we employ the same architecture and hyper-parameters, while modifying the erased region portions a_e. Specifically, we fix the values of a_e for this analysis, to study the effect of a_e to model utility (accuracy) and model privacy (attack accuracy). We vary a_e from 0.0 (indicating no random erasing and the same as No Defense) to 0.5 (erasing 50% of each input image). After the training of T_θ, we proceed to evaluate its top 1 attack accuracy using SOTA MI attacks. This evaluation is conducted for all target models trained with different a_e. In order to ensure diversity in our study, we employ six distinct setups for model inversion attacks, target model architecture, private dataset, and public dataset, and both low- and high-resolution datasets.

RE has small impact on model utility while degrading MI attacks significantly. Fig. 2 summarizes the impact of erased portions on model performance and model inversion attacks. In all setups, we demonstrably improve robustness against MI attacks with small sacrifice to natural accuracy. For instance, introducing erased portions a_e at a ratio of 0.2 in Setup 1 caused a small 2.76% decrease in natural accuracy while the attack accuracy plummeted by 29.2%. This trend continues in Setup 2 – a 0.2 ratio of a_e led to a modest 3.92% decrease in natural accuracy, but a substantial 15.47% drop in attack accuracy. We note that in Setup 3, LOMMA+KedMI attack accuracy degrades by 39.93%. For high resolution images (Setup 4, 5), we observe an increase in model accuracy when using RE. In Setup 4, there is a significant 69.39% drop in attack accuracy while natural accuracy slightly increase (0.37%) when $a_e = 0.5$. Similar trend for Setup 5,



(a) NoDef, AttAcc = 88.67%, $f_{train}^{NoDef} = f_{priv}^{NoDef}$



(b) MIDRE, $a_e = 0.4$, AttAcc = 27.75%, $f_{train}^{MIDRE} = f_{RE}^{MIDRE}$

Figure 3: **Feature space analysis to show that, under MIDRE, f_{recon}^{MIDRE} and f_{priv}^{MIDRE} has a discrepancy, degrading MI attack.** We visualize penultimate layer activations of private images ($\star$ f_{priv}), RE-private images ($\blacktriangledown$ f_{RE}), and MI-reconstructed images ($\times$ f_{recon}) generated by both (a) NoDef and (b) our MIDRE model. We also visualize the convex hull for private images, RE-private images, and MI-reconstructed images. In (a), f_{recon}^{NoDef} closely resemble f_{priv}^{NoDef}, consistent with high attack accuracy. In (b), private images and RE-private images share some similarity but they are not identical, with partial overlap between f_{priv}^{MIDRE} and f_{RE}^{MIDRE}. Importantly, f_{recon}^{MIDRE} closely resembles f_{RE}^{MIDRE} as RE-private is the training data for MIDRE. This results in **a reduced overlap between f_{recon}^{MIDRE} and f_{priv}^{MIDRE}, explaining that MI does not accurately capture the private image features under MIDRE.** More visualization can be found in Supp.

attack accuracy drops from 88.67% to 27.75% when $a_e = 0.4$ while natural accuracy increases 1.83%. In conclusion, *using RE-images during training significantly degrades MI attack while impact on natural accuracy is small.*

3.2 Feature space analysis of Random Erasing's defense effectiveness

In this section, we present a novel observation that explains RE's defense effectiveness. We observe **Property P1: Model trained with RE-private images following our MIDRE leads to a discrepancy between the features of MI-reconstructed images and that of private images**, resulting in degrading of attack accuracy.

The following analysis explains why MIDRE has **Property P1**. We use the following notation: f_{train}, f_{priv}, f_{RE}, and f_{recon} represent the features of training images, private images, RE-private images, and MI-reconstructed images, respectively. To extract these features, we first train the target model without any defense (NoDef) and another target model with our MIDRE. Then, we pass images into these models to obtain the penultimate layer activations. Specifically, we input private images into the models to obtain f_{priv}. Next, we apply RE to private images, pass these RE-private images into the models to obtain f_{RE}. We also perform MI attacks to obtain reconstructed images from NoDef model (resp. MIDRE model), and then feed them into the NoDef model (resp. MIDRE model) to obtain f_{recon}. We use the same experimental setting as in Sec. 3.3. Then, we visualize penultimate layer activations $f_{priv}, f_{RE}, f_{recon}$ by both NoDef and our MIDRE model. We use $a_e = 0.4$ to train MIDRE and to generate RE-private images. Additionally, we visualize the convex hull of these features.

Features of MI-reconstructed images tend to match features of training data. SOTA MI attacks aim to reconstruct images that maximize the likelihood under the target model (Eq. 1) in order to extract training data (which possess a high likelihood under the target model). Under attacks of high accuracy,

f_{recon} tends to match the features of training data f_{train} (Nguyen et al., 2023).

Evidence. In Fig. 3 (a), as the training data of NoDef is private images $f_{train}^{NoDef} = f_{priv}^{NoDef}$, we observe that in NoDef model, f_{recon}^{NoDef} overlaps f_{priv}^{NoDef}, i.e. there is significant overlap between the pink and blue polygons. In Fig. 3 (b), the MIDRE model is trained with RE-private images $f_{train}^{MIDRE} = f_{RE}^{MIDRE}$, and as a result, pink polygon f_{recon}^{MIDRE} and green polygon f_{RE}^{MIDRE} overlap. This confirms **features of reconstructed images tend to match the features of training data**, i.e. f_{priv}^{NoDef} in NoDef and f_{RE}^{MIDRE} in MIDRE.

Mismatch in feature space of MIDRE. MIDRE is trained using RE-private images and is generalizable to images without RE as shown in (Zhong et al., 2020). Under MIDRE target model, f_{RE}^{MIDRE} and f_{priv}^{MIDRE} have partial overlaps, but they are not identical. Meanwhile, f_{recon}^{MIDRE} tend to match with f_{RE}^{MIDRE} (RE-private images are training data for MIDRE, and follows the discussion above). Therefore, f_{recon}^{MIDRE} do not replicate f_{priv}^{MIDRE}, significantly degrading the MI attack.

Evidence. In Fig. 3 (b), f_{RE}^{MIDRE} and f_{priv}^{MIDRE} are partial overlap. Importantly, f_{recon}^{MIDRE}, which overlaps with f_{RE}^{MIDRE} as explained above, only partially overlaps with f_{priv}^{MIDRE}, suggesting MI attacks fail to guide the reconstructed features to replicate private features. Consequently, **MIDRE introduces a discrepancy between MI-reconstructed and private images in feature space of the target model, degrading the attack accuracy.**

Remark. Note that the mismatch between f_{RE}^{MIDRE} and f_{priv}^{MIDRE} does not cause the reduction of model utility (see Fig. 4). This is because the private images remain distinct from other classes and distant from other classification regions, even when their representations are partially overlapped with RE-private images (the training data).

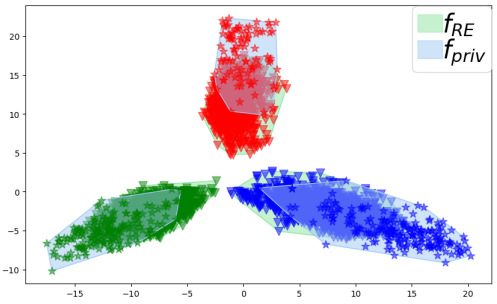


Figure 4: **MIDRE target model achieves high accuracy despite partial overlap of f_{RE}^{MIDRE} and f_{priv}^{MIDRE}.** We visualize the penultimate layer activations of RE-private images and private images for three identities. While f_{RE}^{MIDRE} and f_{priv}^{MIDRE} do not completely overlap, the model can still classify private images with high accuracy. This is because the private images remain distinct from other classes and distant from other classification regions, even when their representations are partially shared with RE-private images (the training data). We remark that RE randomly erases different regions from the images in different iterations, preventing the model to learn shortcut features and forcing the model to learn intrinsic features and become more generalizable beyond training data. More visualization can be found in Supp.

3.3 Importance of partial erasure and random location for privacy-utility trade-off

In addition to two properties discussed in Sec. 3.2 which contribute to outstanding effectiveness of applying RE to degrade MI, we analyse in this section two properties of Random Erasing that are: **Property P2: Partial Erasure**, and **Property P3: Random Location**. To investigate the effect of each property, we conduct the experiment using the following setup: We use T = ResNet-18 (Simonyan & Zisserman, 2014), D_{priv} = Facecrub (Ng & Winkler, 2014), D_{pub} = FFHQ (Karras et al., 2019), attack method = PPA (Struppek et al., 2022). To evaluate the effectiveness of **Partial Erasure** and **Random Location**, we conduct experiments on three schemes: **Entire Erasing (EE), Fixed Erasing (FE)**, and **Random Erasing (RE)**. These schemes are compared against a No Defense baseline, which is trained for 100 epochs without any defense. In the Entire Erasing (EE) scheme, we randomly reduce the number of erased samples per ID in each training epoch to simulate varying pixel concealment levels. Specifically, we train the model for 100 epochs, and for each epoch, we erase randomly 50%, 40%, 30%, 20%, 10%, and 0% of images per ID. For Fixed Erasing (FE), a fixed location within each image is erased throughout the entire training process. However, the erased location varied across different images. For Random Erasing (RE), the location of erased areas is randomly selected for each image and training iteration. We train the RE model for 100 epochs with different values of the erasure ratio, a_e = 0.5, 0.4, 0.3, 0.2, 0.1 corresponding to 50%, 40%, 30 %, 20%, and 10% pixel concealment, respectively.

Table 1: We compare three different techniques for pixel concealment, to reduce the amount of private information presented to the model during training. The results show that simply reducing epochs as in "Entire Erasure" is insufficient for degrading attack performance. Meanwhile, RE improves model utility while degrading attack accuracy effectively.

Concealment		**Partial Erasure**				**Entire Erasing**	
		Random Erasing		**Fixed Erasing**			
	a_e	**Acc (↑)**	**AttAcc (↓)**	**Acc (↑)**	**AttAcc (↓)**	**Acc (↑)**	**AttAcc (↓)**
0%	0	97.69	87.12	97.69	87.12	97.69	87.12
10%	0.1	**97.91**	**79.76**	96.13	85.26	97.33	87.83
20%	0.2	**97.64**	**66.32**	96.79	69.81	97.52	87.03
30%	0.3	97.14	**46.30**	96.13	50.71	**97.53**	89.15
40%	0.4	96.05	**27.75**	93.10	28.49	**97.30**	89.03
50%	0.5	93.77	15.98	86.69	**14.86**	**97.21**	87.59

Property P2 brings the privacy effect to defend against MI attacks. By erasing portions of training images, it reduces the amount of private information exposed to the model during training. By erasing more information, we can effectively degrade the accuracy of privacy attacks. Importantly, partial erasures prevent the model from seeing **entire images**. Consequently, RE-images provide less information about the entirety of the face, such as inter-pupillary distance, relative distances between the eyes, nose, and mouth, the position of the cheekbones, etc. Note that, features need to be presented to the model many times during training in order for the model to learn the features, as suggested by the common practice of using multiple epochs to train a robust model. **Property P2** reduces the frequency of presenting the features to the model during training. Such reduced frequency using partial erasure makes it more difficult for the model to memorize the identity features.

Evidence. In Tab. 1, in terms of degrading the attack, partial erase (fixed or random) is more effective than entire erase (reduce numbers of sample per ID) although the percentages of pixel concealment are the same. Specifically, EE (reduce 50% images per ID) is significantly more vulnerable to attacks than RE and FE (50% image areas are erased, trained in 100 epochs), suffering approximately 71% higher in attack accuracy.

Property P3 recovers the model utility. While information reduction can improve privacy, it may also negatively impact model utility if too much information is erased. Fixing the erasing location for an image means some identity feature of this image will never be presented to the model, model may not have adequate information to learn effectively. RE avoids this issue. As the location of erased area is changed in each training iteration, RE improves the diversity of the training data and ensures that the model still observes a significant portion of the image.

Evidence. In Tab. 1, RE improves the model accuracy while maintains the same attack accuracy as FE in different erased portion ratio a_e. For instance, RE has higher model accuracy than FE by 7.08% with a_e = 0.5. With a_e = 0.3 and 0.4, RE has higher accuracy and lower attack accuracy than EE model, showing that privacy effect of RE.

4 Experiments

4.1 Experimental Setting

To demonstrate the generalisation of our proposed MI defense, we carry out multiple experiments using different SOTA MI attacks on various architectures. In addition, we also use different setups for public and private data. The summary of all experiment setups is shown in Tab. 2. In total, we conducted 37 experiment setups to demonstrate the effectiveness of our proposed defense MIDRE.

Dataset: We follow the same setups as SOTA attacks (Zhang et al., 2020; Nguyen et al., 2023; Struppek et al., 2022) and defense (Peng et al., 2022; Struppek et al., 2024; Ho et al., 2024) to conduct the experiments on four datasets including: CelebA (Liu et al., 2015), FaceScrub (Ng & Winkler, 2014), VGGFace2 (Cao

Table 2: In total, we conduct 37 experiment setups to demonstrate the effectiveness of MIDRE.

Attack	T	$\mathcal{D}_{priv}$	$\mathcal{D}_{pub}$	Resolution
GMI (Zhang et al., 2020) KedMI (Chen et al., 2021) LOMMA (Nguyen et al., 2023) PLGMI (Yuan et al., 2023) RLBMI (Han et al., 2023) BREPMI (Kahla et al., 2022)	VGG16 (Simonyan & Zisserman, 2014) IR152 (He et al., 2016) FaceNet64 (Cheng et al., 2017)	CelebA	CelebA/FFHQ	64×64
PPA (Struppek et al., 2022)	ResNet18 (He et al., 2016) ResNet101 (He et al., 2016) ResNet152 (He et al., 2016) DenseNet121 (Huang et al., 2017) DenseNet169 (Huang et al., 2017) MaxVIT (Tu et al., 2022)	Facescrub	FFHQ	224×224
	ResneSt101	Stanford Dogs	AFHQ Dogs	
MIRROR (An et al., 2022)	Inception-V1 (Inc) ResNet50 (He et al., 2016)	VGGFace2	FFHQ	160×160 224×224
IF-GMI (Qiu et al., 2024)	ResNet18 ResNet152	Facescrub	FFHQ	224×224

et al., 2018), and Stanford Dogs (Dataset, 2011). We use FFHQ (Karras et al., 2019) and AFHQ Dogs (Choi et al., 2020) for the public dataset. We strictly follow (Zhang et al., 2020; Nguyen et al., 2023; Struppek et al., 2022; An et al., 2022; Peng et al., 2022; Struppek et al., 2024; Ho et al., 2024; Koh et al., 2024) to divide the datasets into public and private set. See Supp for the details of datasets.

Model Inversion Attacks. To evaluate the effectiveness of our proposed defense MIDRE, we employ a comprehensive suite of state-of-the-art MI attacks. This includes various attack categories: white-box and label-only, one type of black-box attack. To assess robustness at high resolutions, we employ PPA (Struppek et al., 2022) and IF-GMI (Qiu et al., 2024) against attacks targeting 224×224 pixels and MIRROR (An et al., 2022) against attacks targeting 224×224 and 160×160 pixels. For low resolution 64×64 pixels, we leverage four SOTA white-box attacks: GMI (Zhang et al., 2020), KedMI (Chen et al., 2021), PLG-MI (Yuan et al., 2023), and LOMMA (Nguyen et al., 2023) (including LOMMA+GMI and LOMMA+KedMI). Additionally, we incorporate RLBMI (Han et al., 2023) for black-box attack and BREPMI (Kahla et al., 2022) for label-only attack. We strictly replicate the experimental setups in (Zhang et al., 2020; Chen et al., 2021; Yuan et al., 2023; Nguyen et al., 2023; Struppek et al., 2022; Peng et al., 2022; Han et al., 2023; An et al., 2022) to ensure a fair comparison between NoDef (the baseline model without defense), existing state-of-the-art defenses, and our proposed method, MIDRE.

Target Models. We follow other MI research (Zhang et al., 2020; Nguyen et al., 2023; Struppek et al., 2022; Peng et al., 2022) to train defense models. We use 11 architectures for the target model to assess its resistance to MI attacks using various experimental configurations. The details are summarized in Tab. 2. We train target models with the same hyper-parameter (a_h) for all low-resolution data set-ups. In addition, for high-resolution data, we use two value for hyper-parameter $a_h = 0.4$ and $a_h = 0.8$ across all setups. This allows us to demonstrate MIDRE's effectiveness in achieving the optimal trade-off between utility and privacy with consistent hyper-parameter.

Comparison Methods. We compare the performance of our model against no defending method (NoDef) and five defense methods, including NLS (Negative Label Smoothing)(Struppek et al., 2024), TL-DMI (Ho et al., 2024), MI-RAD (MI-resilient architecture designs) (Koh et al., 2024), BiDO (Peng et al., 2022), and MID (Wang et al., 2021). As for MI-RAD (Koh et al., 2024), we compare our results to Removal of Last Stage Skip-Connection (RoLSS), Skip-Connection Scaling Factor (SSF), Two-Stage Training Scheme (TTS).

We establish a baseline (NoDef) by training the target model from scratch without incorporating any MI defense strategy. According to NLS, TL-DMI, MI-RAD, we follow their setup and evaluation to compare with MIDRE. We then carefully tune the hyper-parameters of each method to achieve optimal performance.

Evaluation Metrics. MI defenses typically involve a trade-off between the model's utility and its resistance

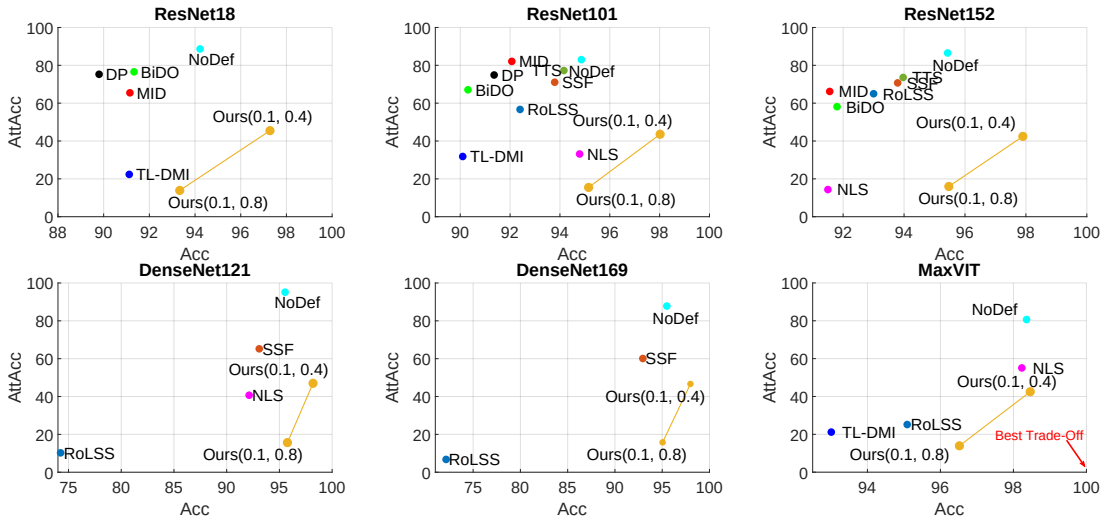


Figure 5: We evaluate PPA attack (Struppek et al., 2022) on our proposed method, NoDef, MID (Wang et al., 2021), BiDO (Peng et al., 2022), NLS (Struppek et al., 2024), and TL-DMI (Ho et al., 2024). Target models are trained on $\mathcal{D}_{priv}$ = Facescrub with 6 architectures. The results show that our method archives the best trade-of between utility and privacy among state-of-the-art defenses.

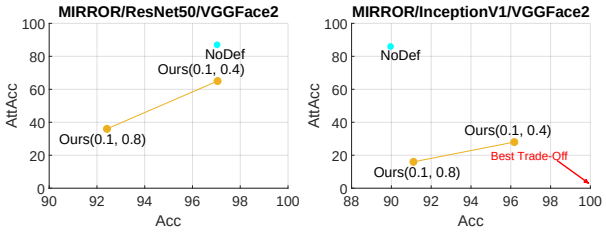


Figure 6: We evaluate MIRROR attack (An et al., 2022) on VGGFace2 dataset. The results show that our method archives the best trade-of between utility and privacy than NoDef model.

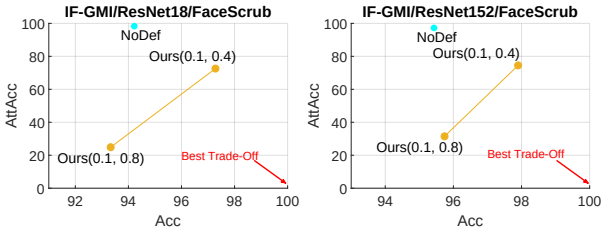


Figure 7: Results of IF-GMI (Qiu et al., 2024) attack on Facescrub dataset. Here, we use T = ResNet18/ ResNet152, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ, image resolution = 224×224 images.

to model inversion attacks. In this section, we evaluate these defenses using two key metrics: Natural Accuracy (Acc ↑) to evaluate the model utility and Attack accuracy (AttAcc ↓) and to evaluate the model privacy. We further show other evaluation metrics, including K-Nearest Neighbor Distance, δ_{eval}, δ_{face} (Struppek et al., 2022), complement these results with qualitative results and a user study in Supp.

4.2 Comparison against SOTA MI Defenses

We compare the model accuracy and attack accuracy of defense models in 6 architectures using attack method PPA (Struppek et al., 2022) in Fig. 5. All the target models are trained on Facescrub dataset. Interestingly, we are the first to observe that our defense models achieve higher natural accuracy but lower attack acuracy than no defense model for larger image sizes (224×224). With small masking areas (Ours(0.1,0.4)), our proposed method consistently achieves the lowest attack accuracy among all defense models while its natural accuracy is higher than NoDef, BiDO, MID, and DP models. For example, using ResNet101, our model reduces attack accuracy by 39.42% compared to NoDef while achieving the model accuracy is higher than NoDef model 3.16%. MaxVIT, a recent advanced architecture, has very high attack accuracy (80.66%). Our defense mechanism significantly enhances its robustness, lowering attack accuracy to 42.5% without compromising model performance. By increasing the masking areas (Ours(0.1,0.8)), they achieve a significant reduction in attack accuracy while maintaining high natural accuracy, outperforming other strong defense methods like NLS and TL-DMI. Specially, *our attack accuracies are below 20% for all architectures*. This



Figure 8: Reconstructed image obtained from IF-GMI attack with T = ResNet-18, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ. Overall, our reconstructed images have less similarity to private images compared to those from the no-defense model, suggesting the efficiency of our proposed defense MIDRE.

represents the best utility-privacy trade-off among all evaluated defense models, demonstrating our method's effectiveness in mitigating model inversion attacks.

We further show the effectiveness of our proposed method using two attacks MIRROR (An et al., 2022) and IF-GMI (Qiu et al., 2024). Regarding the MIRROR attack, we compare the results of MIDRE and the NoDef model using $\mathcal{D}_{priv}$ = VGGFace2 (see Fig. 6). Our defense reduces the attack accuracy by 22% and 70% without harming model utility, where the target model T = ResNet50/InceptionV1. Results of IF-GMI attack are shown in Fig. 7. The results show that MIDRE reduces the attack accuracy by more than 22%.

For low resolution attacks, we evaluate against four MI attacks, including GMI (Zhang et al., 2020), KedMI (Chen et al., 2021), LOMMA (Nguyen et al., 2023) with two variances (LOMMA+GMI and LOMMA+KedMI), and PLGMI (Yuan et al., 2023) We follow the same setup and compare with NLS in Tab. 3. In addition, we also use NoDef baseline in NLS paper to compare and estimate Δ in Tab. 3.

Table 3: We evaluate multiple SOTA MI attacks on 64×64 images, comparing their performance under NoDef, NLS, and our MIDRE defense. T = VGG16, D_{priv} = CelebA, D_{pub} = CelebA.

Attack	Defense	Acc ↑	AttAcc ↓	Δ ↑
LOMMA + GMI	NoDef	85.74	53.64 ± 4.64	-
	NLS	80.02	39.16 ± 4.25	2.53
	MIDRE	79.85	**26.62 ± 1.93**	**4.59**
LOMMA + KedMI	NoDef	85.74	72.96 ± 1.92	-
	NLS	80.02	63.60 ± 1.37	1.64
	MIDRE	79.85	**41.82 ± 1.24**	**5.29**
PLGMI	NoDef	85.74	71.00 ± 3.31	-
	NLS	80.02	72.00 ± 2.50	-0.17
	MIDRE	79.85	**66.60 ± 2.94**	**0.75**

Overall, our proposed method, MIDRE, achieves significant improvements in security for 64×64 setups compared to SOTA MI defenses. MIDRE achieves this by demonstrably reducing top-1 attack accuracy while maintaining natural accuracy on par with other leading MI defenses. More results of other attacks can be found in Supp.

We show the comparison on qualitative results in Fig. B.2. Images are reconstructed using IF-GMI atack (Qiu et al., 2024), T = ResNet-18, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ. Overall, our reconstructed images have less similarity to private images compared to those from the no-defense model, demonstrating the effectiveness of our defense. More reconstructed images are included in Supp and our project page.

The experiment results demonstrate that our defense model has a small impact on model utility while significantly enhancing the model's robustness against state-of-the-art MI attacks. Moreover, we are the first to report a substantial improvement in model utility among all existing defenses.

4.3 Ablation study on data augmentation type.

To evaluate the effectiveness of our method compared to other data augmentation-based defenses, we compare

Table 4: We report the PPA attack on images with resolution 224×224. T = ResNet18, D_{priv} = Facescrub, D_{pub} = FFHQ to target models trained with different data augmentation. MIDRE consistently achieves the best balance between utility and privacy, significantly degrading MI attack performance while maintaining competitive natural accuracy.

Attack	Defense	Acc ↑	AttAcc ↓
PPA	NoDef	94.22	88.67
	MIDRE	97.28	**48.16**
	CutMix	98.74	67.12
	Random Cropping	92.24	74.22
	Gaussian Blur	97.57	87.12

Table 5: The combination MIDRE with existing defense BiDO and NLS. The combine models significantly reduces attack accuracy compared to individual defenses. We denote "OP" for Δ if the accuracy of the defense model outperforms that of the NoDef model.

Setup	**Defense**	**Acc (↑)**	**AttAcc (↓)**	Δ(↑)
Setup 1	NoDef	95.43	86.51	-
	NLS	91.50	13.94	18.47
	MIDRE	95.47	15.97	OP
	MIDRE + NLS	93.69	3.75	47.65
Setup 2	NoDef	86.90	81.80 ± 1.44	-
	BiDO	79.85	63.00 ± 2.08	2.67
	MIDRE	79.85	43.07 ± 1.99	5.49
	MIDRE + BiDO	82.15	39.00 ± 1.30	9.01

MIDRE with models trained using CutMix, random cropping, and Gaussian blur. These augmentations are commonly employed to enhance model generalization. Here, we use the setup: attack method = PPA, T = ResNet18, D_{priv} = Facescrub, D_{pub} = FFHQ, image resolution = 224×224.

The results, summarized in Table 4, indicate that while alternative augmentations provide some level of protection, *MIDRE consistently achieves the best balance between utility and privacy, significantly degrading MI attack performance while maintaining competitive natural accuracy.*

4.4 Combination with Existing Defenses

Since MIDRE improves defense effectiveness from the training data perspective, our proposed method can be combined with other defense mechanism from the training objective perspective such as BiDO (Peng et al., 2022) and NLS (Struppek et al., 2024). We use 2 setups: **Setup 1**: T = ResNet152, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ, Attack method = PPA, image size = 224 × 224. **Setup 2**: T = VGG16, $\mathcal{D}_{priv}/\mathcal{D}_{pub}$ = CelebA, Attack method = LOMMA + KedMI, image size = 64 × 64. We use a_e = [0.1,0.4] and a_e = [0.1,0.8] for setup 1 and setup 2 to train MIDRE and during attack.

The results (see Tab. 5) demonstrate the effectiveness of combining MIDRE with either NLS or BiDO for enhancing defense against MI attacks, as our MIDRE takes a data-centric perspective for defense, complementary to existing defenses. In both experimental setups, the combination models demonstrate a substantial reduction in attack accuracy compared to using MIDRE or the other defenses independently. In particular, in setup 1, the combination of MIDRE and Negative LS achieves a remarkable 3.75% attack accuracy when using the state-of-the-art PPA attack while preserving model utility. For Setup 2, MIDRE + BiDO improves the natural accuracy of the model by 2.3% while reducing the attack accuracy by 4.07% and 24% compared to MIDRE and BiDO, respectively. **This shows our effectiveness of combining MIDRE and existing defense for a better defense.** The combination ability of MIDRE supports that it examines a distinct aspect of the system by focusing on data input, setting it apart from other existing approaches to defend against model inversion attacks.

5 Conclusion

We propose a novel approach to MI Defense via Random Erasing (MIDRE). We conduct an analysis to demonstrate that RE possess two crucial properties to degrade MI attack while the impact on model utility is small. Furthermore, our features space analysis shows that model trained with RE-private images following MIDRE leads to a discrepancy between the features of MI-reconstructed images and that of private images, resulting in reducing of attack accuracy. Experiments validate that our approach achieves outstanding performance in balancing model privacy and utility. The results consistently demonstrate the superiority of our method over existing defenses across various MI attacks, network architectures, and attack configurations. The code and additional results can be found in the Supplementary section.

Acknowledgements

This research is supported by Temasek Laboratories, Singapore University of Technology and Design. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); The Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority. This research is also part-funded by the European Union (Horizon Europe 2021-2027 Framework Programme Grant Agreement number 10107245. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them) and by the Engineering and Physical Sciences Research Council under grant number EP/X029174/1.

References

Inception resnet (v1) models in pytorch. https://github.com/timesler/ facenet-pytorch.

Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022.

Mir Mohammad Azad, Apoorva Ganapathy, Siddhartha Vadlamudi, and Harish Paruchuri. Medical diagnosis using deep learning techniques: a research survey. *Annals of the Romanian Society for Cell Biology*, 25(6):5591–5600, 2021.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-enriched distributional model inversion attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16178–16187, 2021.

Yu Cheng, Jian Zhao, Zhecan Wang, Yan Xu, Karlekar Jayashree, Shengmei Shen, and Jiashi Feng. Know you at one glance: A compact vector representation for low-shot learning. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 1924–1932, 2017.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

E Dataset. Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer*. Citeseer, 2011.

Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8599–8603. IEEE, 2013.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pp. 1–19. Springer, 2008.

Xueluan Gong, Ziyao Wang, Shuaike Li, Yanjiao Chen, and Qian Wang. A gan-based defense framework against model inversion attacks. *IEEE Transactions on Information Forensics and Security*, 2023.

Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Computer vision and image understanding*, 189:102805, 2019.

Gyojin Han, Jaehyun Choi, Haeil Lee, and Junmo Kim. Reinforcement learning-based black-box model inversion attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20504–20513, 2023.

J Harikrishnan, Arya Sudarsan, Aravind Sadashiv, and Remya AS Ajai. Vision-face recognition attendance monitoring system for surveillance using deep learning technology and computer vision. In *2019 international conference on vision towards emerging trends in communication and networking (ViTECoN)*, pp. 1–5. IEEE, 2019.

Tufail Sajjad Shah Hashmi, Nazeef Ul Haq, Muhammad Moazam Fraz, and Muhammad Shahzad. Application of deep learning for weapons detection in surveillance videos. In *2021 international conference on digital futures and transformative technologies (ICoDT2)*, pp. 1–6. IEEE, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sy-Tuyen Ho, Koh Jun Hao, Keshigeyan Chandrasegaran, Ngoc-Bao Nguyen, and Ngai-Man Cheung. Model inversion robustness: Can transfer learning help? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12183–12193, 2024.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15045–15053, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Jun Hao Koh, Sy-Tuyen Ho, Ngoc-Bao Nguyen, and Ngai-man Cheung. On the vulnerability of skip connections to model inversion attacks. In *European Conference on Computer Vision*, 2024.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pp. 471–478. IEEE, 2018.

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165, 2019.

Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*, pp. 343–347. IEEE, 2014.

Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-thinking model inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2023.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

Niall O'Mahony, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Daniel Riordan, and Joseph Walsh. Deep learning vs. traditional computer vision. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*, pp. 128–144. Springer, 2020.

Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral dependency optimization: Defending against model-inversion attacks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1358–1367, 2022.

Xiong Peng, Bo Han, Feng Liu, Tongliang Liu, and Mingyuan Zhou. Pseudo-private data guided model inversion attacks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 33338–33375. Curran Associates, Inc., 2024.

Gege Qi, YueFeng Chen, Xiaofeng Mao, Binyuan Hui, Xiaodan Li, Rong Zhang, and Hui Xue. Model inversion attack via dynamic memory learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5614–5622, 2023.

Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A closer look at gan priors: Exploiting intermediate features for enhanced model inversion attacks. In *Proceedings of European Conference on Computer Vision*, 2024.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correira, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. In *International Conference on Machine Learning*, pp. 20522–20545. PMLR, 2022.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Be careful what you smooth for: Label smoothing can be a privacy shield but also a catalyst for model inversion attacks. In *The Twelfth International Conference on Learning Representations*, 2024.

Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1): 23–34, 2004.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479. Springer, 2022.

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11666–11673, 2021.

Dayong Ye, Sheng Shen, Tianqing Zhu, Bo Liu, and Wanlei Zhou. One parameter defense—defending against data inference attacks via differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:1466–1480, 2022.

Xiaojian Yuan, Kejiang Chen, Jie Zhang, Weiming Zhang, Nenghai Yu, and Yang Zhang. Pseudo label-guided model inversion attack via conditional generative adversarial network. *AAAI 2023*, 2023.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Xiaokang Zhou, Xuesong Xu, Wei Liang, Zhi Zeng, and Zheng Yan. Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart iot. *IEEE Internet of Things Journal*, 8(16):12588–12596, 2021.

Supplementary material

In this supplementary material, we provide additional experiments, analysis, ablation study, and details that are required to reproduce our results. These are not included in the main paper due to space limitations.

Our code and additional results are available at: `https://ngoc-nguyen-0.github.io/MIDRE/`

Contents

A Details on Experimental Setup **16**
- A.1 Evaluation Method . . . 16
- A.2 Dataset . . . 17
- A.3 Train the Defense model using Random Erasing . . . 18
- A.4 Hyper-parameters for Model Inversion Attack . . . 18
- A.5 Hyper-parameters for MIDRE . . . 19

B Additional Experimental Results **19**
- B.1 Experiments on low resolution images . . . 19
- B.2 Additional results . . . 20
- B.3 User Study . . . 23
- B.4 Qualitative Results . . . 24

C Additional analysis of privacy effect of MIDRE **24**
- C.1 Feature space analysis of Random Erasing's defense effectiveness . . . 24
- C.2 Importance of partial erasure and random location for privacy-utility trade-off . . . 26

D Ablation Study **28**
- D.1 Ablation study on alternative masking strategies . . . 28
- D.2 Ablation study on Masking Values. . . . 29
- D.3 Ablation study on Area Ratio. . . . 29
- D.4 Ablation study on Aspect Ratio. . . . 30
- D.5 Adaptive attack . . . 30
- D.6 The effectiveness of substitute pixels generated by inpainting for MIDRE. . . . 31

E Discussion **31**
- E.1 Broader Impacts . . . 31
- E.2 Limitation . . . 31

F Experiments Compute Resources **32**

G Related Work **32**
- G.1 Model Inversion Attacks . . . 32
- G.2 Model Inversion Defenses . . . 32

A Details on Experimental Setup

A.1 Evaluation Method

To evaluate the attack, existing methods Zhang et al. (2020); Chen et al. (2021); Nguyen et al. (2023); Struppek et al. (2024); An et al. (2022) train an evaluation model E that has a distinct architecture and is trained on the private dataset $\mathcal{D}_{priv}$. Similar to human inspection practices (Zhang et al., 2020), the evaluation model acts as a human proxy for assessing the quality of information leaked through MI attacks. We report the details of the evaluation models in the Tab. A.1. All the evaluation models are provided by Chen et al. (2021); Struppek et al. (2022); An et al. (2022).

We evaluate defense methods using the following metrics:

Table A.1: Details of evaluation model E in all the experimental setup. All the evaluation models are provided by Chen et al. (2021); Struppek et al. (2022); An et al. (2022).

Attack	T	$\mathcal{D}_{priv}$	$\mathcal{D}_{pub}$	Resolution	E	E's accuracy
GMI (Zhang et al., 2020) KedMI (Chen et al., 2021) LOMMA (Nguyen et al., 2023) PLGMI (Yuan et al., 2023) RLBMI (Han et al., 2023) BREPMI (Kahla et al., 2022)	VGG16 (Simonyan & Zisserman, 2014) IR152 (He et al., 2016) FaceNet64 (Cheng et al., 2017)	CelebA	CelebA/FFHQ	64×64	FaceNet112	95.80
PPA (Struppek et al., 2022)	ResNet18 (He et al., 2016) ResNet101 (He et al., 2016) ResNet152 (He et al., 2016) DenseNet121 (Huang et al., 2017) DenseNet169 (Huang et al., 2017) MaxVIT (Tu et al., 2022)	Facescrub	FFHQ	224×224	Inception-V3	96.20%
	ResneSt101	Stanford Dogs	AFHQ Dogs		Inception-V3	79.79%
MIRROR (An et al., 2022)	Inception-V1 (Inc)	VGGFace2	FFHQ	160×160	ResNet50	99.88%
	ResNet50 (He et al., 2016)			224×224	Inception-V1	99.65%
IF-GMI (Qiu et al., 2024)	ResNet18 ResNet152	Facescrub	FFHQ	224×224	Inception-V3	96.20%

- **Natural Accuracy (Acc ↑)**. This metric measures the accuracy of the defended model on a private test set, reflecting its performance on unseen data. Higher natural accuracy indicates better performance of the primary task.
- **Attack accuracy (AttAcc ↓) Zhang et al. (2020).** This metric measures the percentage of successful attacks, where success is defined as the ability to reconstruct private information from the model's outputs. Lower attack accuracy indicates a more robust defense. Following existing works (Zhang et al., 2020; Chen et al., 2021; Nguyen et al., 2023; Struppek et al., 2022), we utilize a separate evaluation model E to evaluate the inverted images. Higher attack accuracy on the evaluation model signifies a more effective attack, implying a weaker defense.
- **K-Nearest Neighbor Distance (KNN Dist ↑) Chen et al. (2021).** KNN distance measures the similarity between a reconstructed image of a specific identity and their private images. This is calculated using the L_2 norm in the feature space extracted from the penultimate layer of the evaluation model. In MI defense, a higher KNN Dist value indicates a greater degree of robustness against model inversion (MI) attacks and a lower quality of attacking samples on that model.
- δ_{eval} **and** δ_{face}**Struppek et al. (2022).** We also use δ_{eval} and δ_{face} metrics from (Struppek et al., 2022) to quantify the quality of inverted images generated by PPA. These two metrics are the same concept as KNN Dist, but different in the model to produce a feature to calculate distance. δ_{face} use pretrained FaceNet (Schroff et al., 2015) as model to extract penultimate features, while δ_{eval} uses evaluation model for PPA attack.
- **Trade-off value (Δ ↑) Ho et al. (2024).** To quantify the trade-off between model utility (natural accuracy) and attack performance (attack accuracy), we follow previous work and let NoDef model and defended model be f_n and f_d respectively, we compute $\Delta = \frac{AttAcc_{f_n} - AttAcc_{f_d}}{Acc_{f_n} - Acc_{f_d}}$. This metric calculates the ratio between the decrease in attack accuracy and the decrease in natural accuracy when applying an MI attack to a model without defenses (NoDef) and defense models. We remark that this metric is used when defense models have lower natural accuracy compared to the no-defense model. A higher Δ value indicates a more favorable trade-off.

A.2 Dataset

We use three datasets including CelebA (Liu et al., 2015), Facescrub (Ng & Winkler, 2014), and Stanford Dogs (Dataset, 2011) as private training data and use two datasets including FFHQ (Karras et al., 2019) and AFHQ Dogs(Choi et al., 2020) as public dataset.

The CelebA dataset (Liu et al., 2015) is an extensive compilation of facial photographs, encompassing more than 200,000 images that represent 10,177 distinct persons. For MI task, we follow (Zhang et al., 2020;

Algorithm 1 Train the Defense model using Random Erasing

Input: Private training data $\mathcal{D}_{priv} = \{(x_i, y_i)\}_{i=1}^{N}$, model T_θ, a maximum masking area portion a_h.
Output: The MIDRE-trained model T_θ.
Initialize $t \leftarrow 0$
while $t < t_{RE}$ **do**
  Sample a mini-batch $\mathcal{D}_b$ with size b from $\mathcal{D}_{priv}$
  $\mathcal{D}_{RE} = \{\}$
  while (x, y) in $\mathcal{D}_b$ **do**
    $\tilde{x} = x$
    Randomly select a_e within the range $[0.1, a_h]$
    $\tilde{x} = RE(x, a_e)$ ▷ *This is following the procedure discussed in Sec. 2.2*
    $\mathcal{D}_{mask} \leftarrow (\tilde{x}, y)$
  end while
  Compute $\mathcal{L}(\theta) = \frac{1}{b} \sum^{\mathcal{D}_{RE}} \ell(T_\theta(\tilde{x_i}), y_i)$
  Backward Propagation $\theta \leftarrow \theta - \alpha \nabla \mathcal{L}(\theta)$
end while

Chen et al., 2021; Nguyen et al., 2023) to divide CelebA into private dataset and public dataset. There is no overlap between private and public dataset. All the images are resized to 64×64 pixels.

Facescrub (Ng & Winkler, 2014) consists of a comprehensive collection of 106836 photographs showcasing 530 renowned male and female celebrities. Each individual is represented by an average of around 200 images, all possessing diversity of resolution. Following PPA (Struppek et al., 2022), we resize the image to 224×224 for training target models.

The FFHQ dataset comprises 70,000 PNG images of superior quality, each possessing a resolution of 1024x1024 pixels. FFHQ is used as a public dataset to train GANs using during attacks (Zhang et al., 2020; Chen et al., 2021; Struppek et al., 2022).

Stanford dogs (Dataset, 2011) contains more than 20,000 images encompassing 120 different dogs. AFHQ Dogs (Choi et al., 2020) contain around 5,000 dog images in high resolution. Follow (Struppek et al., 2022), we use Stanford dogs dataset as private dataset while AFHQ Dogs as the public dataset.

VGGFace2 (Cao et al., 2018) is a large-scale face recognition dataset designed for robust face recognition tasks. It consists of images that are automatically downloaded from Google Image Search, capturing a wide range of variations in factors such as pose, age, illumination, ethnicity, and profession. The diversity of the dataset makes it suitable for training and evaluating face recognition models across different conditions and demographics. It contains more than 3.3 milions images for 9000 identities.

A.3 Train the Defense model using Random Erasing

We depict our method in Algorithm 1.

A.4 Hyper-parameters for Model Inversion Attack

In the case of GMI(Zhang et al., 2020), KedMI(Chen et al., 2021), and PLG-MI(Yuan et al., 2023), BREPMI(Kahla et al., 2022), our approach is primarily based on the referenced publication outlining the corresponding attack. However, in certain specific scenarios, we adhere to the BiDO study due to its distinct model inversion attack configuration in comparison to the original paper. The LOMMA(Nguyen et al., 2023) approach involves adhering to the optimal configuration of the method, which encompasses three augmented model architectures: EfficientNetB0, EfficientNetB1, and EfficientNetB2. We adopt exactly the same experimental configuration, including the relevant hyper-parameters, as described in the referenced paper. We also follow PPA and MIRROR paper's configuration (Struppek et al., 2022; An et al., 2022) for our MI attack setups.

A.5 Hyper-parameters for MIDRE

Our method only requires a hyper-parameter a_h, which is 0.4 for all low-resolution setups. According to high-resolution setups, we use $a_h = 0.4$ and $a_h = 0.8$ as two setups for our defense.

B Additional Experimental Results

B.1 Experiments on low resolution images

We evaluate our method against existing Model Inversion defenses. We follow the experiment setup in BiDO (Peng et al., 2022) and report the results on the standard setup using T = VGG16 and $\mathcal{D}_{priv}$ = CelebA in Tab. B.2. We evaluate against six MI attacks, including GMI (Zhang et al., 2020), KedMI (Chen et al., 2021), LOMMA (Nguyen et al., 2023) with two variances (LOMMA+GMI and LOMMA+KedMI), PLGMI (Yuan et al., 2023), and a black-box attack, BREPMI (Kahla et al., 2022).

Overall, our proposed method, MIDRE, achieves significant improvements in security for 64×64 setups compared to SOTA MI defenses. MIDRE achieves this by demonstrably reducing top-1 attack accuracy while maintaining natural accuracy on par with other leading MI defenses. Specifically, compared to BiDO, MIDRE offers a substantial 43.74% decrease in top-1 attack accuracy with sacrificing only 7.05% in natural accuracy (measured using the KedMI attack method). Notably, while BiDO achieves similar natural accuracy to MIDRE, it suffers from a significantly higher top-1 attack accuracy (8.84% higher than MIDRE).

Table B.2: We report the MI attacks under multiple SOTA MI attacks on images with resolution 64×64. We compare the performance of these attacks against existing defenses including NoDef, BiDO, MID and our method. T = VGG16, D_{priv} = CelebA, D_{pub} = CelebA.

Attack	Defense	Acc ↑	AttAcc ↓	Δ ↑	KNN Dist ↑
LOMMA + GMI	NoDef	86.90	74.53 ± 5.65	-	1312.93
	MID	79.16	54.53 ± 4.35	2.58	1348.21
	BiDO	79.85	53.73 ± 4.99	2.95	1422.75
	MIDRE	79.85	**31.93 ± 5.10**	**6.04**	**1590.12**
LOMMA + KedMI	NoDef	86.90	81.80 ± 1.44	-	1211.45
	MID	79.16	67.20 ± 1.59	1.89	1249.18
	BiDO	79.85	63.00 ± 2.08	2.67	1345.94
	MIDRE	79.85	**43.07 ± 1.99**	**5.49**	**1503.89**
PLGMI	NoDef	86.90	97.47 ± 1.68	-	1149.67
	MID	79.16	93.00 ± 1.90	0.58	1111.61
	BiDO	79.85	92.40 ± 1.74	0.72	1228.36
	MIDRE	79.85	**66.60 ± 2.94**	**4.38**	**1475.76**
GMI	NoDef	86.90	20.07 ± 5.46	-	1679.18
	MID	79.16	20.93 ± 3.12	-0.11	1698.50
	BiDO	79.85	6.13 ± 2.98	1.98	1927.11
	MIDRE	79.85	**3.20 ± 2.15**	**2.39**	**2020.49**
KedMI	NoDef	86.90	78.47 ± 4.60	-	1289.46
	MID	79.16	53.33 ± 4.97	3.25	1364.02
	BiDO	79.85	43.53 ± 4.00	4.96	1494.35
	MIDRE	79.85	**34.73 ± 4.15**	**6.20**	**1620.66**
BREPMI	NoDef	86.90	57.40 ± 4.92	-	1376.94
	MID	79.16	39.20 ± 4.19	2.35	1458.61
	BiDO	79.85	37.40 ± 3.66	2.84	1500.45
	MIDRE	79.85	**21.73 ± 2.99**	**5.06**	**1611.78**

Table B.3: Results of IF-GMI(Qiu et al., 2024) attack on Facescrub dataset. Here, we use T = ResNet18/ ResNet152, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ, image resolution = 224×224 images, attack method = IF-GMI.

Architecture	Defense	Acc ↑	AttAcc ↓	δ_{eval} ↑	δ_{face} ↑	FID ↑
ResNet18	NoDef	94.22	98.30	110.04	0.647	40.239
	MIDRE (0.1, 0.4)	97.28	72.58	122.03	0.698	39.7238
	MIDRE (0.1, 0.8)	93.33	**24.85**	**171.48**	**0.966**	**41.325**
ResNet152	NoDef	95.43	97.24	115.76	0.633	**45.703**
	MIDRE (0.1, 0.4)	97.90	74.50	133.22	0.662	40.669
	MIDRE (0.1, 0.8)	95.74	**31.43**	**150.89**	**0.847**	40.388

B.2 Additional results

We further show the effectiveness of our proposed method on a wide range of target model architectures including IR152, FaceNet64, DenseNet-169, ResNeSt-101, and MaxVIT. The results are shown in Tab. B.4, B.5, and Tab.B.6 and B.7 (for comparison with TL-DMI) for 64×64 images and in Tab.B.9 and B.10 for 224×224 images.

The experiment results consistently demonstrate the effectiveness of our proposed method. For example, with T = IR152, we sacrifice only 6.25% in natural accuracy, but the attack accuracies drop significantly, from 22.07% (PLGMI attack) to 40% (LOMMA + GMI attack). Similarly, when T = FaceNet64, natural accuracy decreases by 6.94%, while the attack accuracies drop significantly, from 24.47% (PLGMI attack) to 45% (LOMMA attack). We report the results of additional setup in Tab. B.8. In particular, we use attack method = PLGMI, T = VGG16/IR152/FaceNet64, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ. In addition to measuring attack accuracy, we incorporate KNN distance to demonstrate the efficacy of our strategy across different evaluation scenarios. The specifics of KNN distance can be found in Sec. A.1.

For high resolution images, interestingly, with $\mathcal{D}_{priv}$ = Facescrub, we see a slight increase in natural accuracy (1.95%) along with a significant reduction in attack accuracy of around 40%. These results consistently show that MIDRE significantly reduces the impact of MI attacks. We report detailed results of PPA attack on our method compared to SOTA defense including MID, DP, BiDO, TL-DMI, NLS and RoLSS, SSF, TTS. the results are presented in Tab. B.9 and B.10. We also use δ_{eval} and δ_{face}, with details in Sec. A.1 to evaluate quality of PPA inverted images.

Table B.4: Additional results on 64×64 images. We use T = IR152. The target models are trained on $\mathcal{D}_{priv}$ = CelebA and $\mathcal{D}_{pub}$ = CelebA. The results conclusively show that our defense model is effective compared to NoDef models.

Attack	Defense	Acc ↑	AttAcc ↓	KNN Dist ↑
GMI	NoDef	91.16	32.40 ± 4.88	1587.28
	MIDRE	84.91	**7.87 ± 3.30**	**1888.47**
KedMI	NoDef	91.16	78.93 ± 5.15	1262.44
	MIDRE	84.91	**40.07 ± 4.99**	**1548.16**
LOMMA + GMI	NoDef	91.16	80.93 ± 4.56	1253.03
	MIDRE	84.91	**40.93 ± 6.11**	**1559.88**
LOMMA + KedMI	NoDef	91.16	90.87 ± 1.31	1116.90
	MIDRE	84.91	**52.13 ± 1.81**	**1481.70**
PLGMI	NoDef	91.16	99.47 ± 0.93	1021.42
	MIDRE	84.91	**77.40 ± 4.79**	**1470.46**

Table B.5: Additional results on 64×64 images. We use T = FaceNet64. The target models are trained on $\mathcal{D}_{priv}$ = CelebA and $\mathcal{D}_{pub}$ = CelebA. The results conclusively show that our defense model is effective compared to NoDef models.

Attack	Defense	Acc ↑	AttAcc ↓	KNN Dist ↑
GMI	NoDef	88.50	29.60 ± 5.43	1607.86
	MIDRE	81.56	**6.73 ± 3.42**	**1908.19**
KedMI	NoDef	88.50	81.67 ± 2.63	1270.71
	MIDRE	81.56	**36.33 ± 6.06**	**1545.93**
LOMMA + GMI	NoDef	88.50	83.33 ± 3.40	1259.61
	MIDRE	81.56	**37.60 ± 3.74**	**1570.85**
LOMMA + KedMI	NoDef	88.50	90.87 ± 1.31	1116.90
	MIDRE	81.56	**54.33 ± 1.44**	**1456.84**
PLGMI	NoDef	88.50	99.47 ± 0.69	1091.51
	MIDRE	81.56	**75.00 ± 4.30**	**1509.78**

Table B.6: Additional results compared with TL-DMI on 64×64 images. We use T = IR152. The target models are trained on $\mathcal{D}_{priv}$ = CelebA and $\mathcal{D}_{pub}$ = CelebA. The results conclusively show that our defense model is effective.

Attack	Defense	Acc ↑	AttAcc ↓	Δ ↑	KNN Dist↑
GMI	NoDef	91.16	32.40 ± 4.88	-	1587.28
	TL-DMI	86.70	**8.93 ± 3.73**	5.26	**1819.00**
	MIDRE	87.94	11.07 ± 3.60	**6.62**	1813.11
KedMI	NoDef	91.16	78.93 ± 5.15	-	1262.44
	TL-DMI	86.70	64.60 ± 4.93	3.21	1333.00
	MIDRE	87.94	**46.67 ± 5.45**	**10.02**	**1455.88**
LOMMA + GMI	NoDef	91.16	80.93 ± 4.56	-	1253.03
	TL-DMI	86.70	**41.87 ± 5.37**	8.76	**1551.00**
	MIDRE	87.94	49.40 ± 6.30	**9.79**	1497.50
LOMMA + KedMI	NoDef	91.16	90.87 ± 1.31	-	1116.90
	TL-DMI	86.70	77.73 ± 1.57	2.95	1305.00
	MIDRE	87.94	**62.93 ± 2.15**	**8.68**	**1551.00**

Table B.7: Additional results compared with TL-DMI on 64×64 images. We use T = FaceNet64. The target models are trained on $\mathcal{D}_{priv}$ = CelebA and $\mathcal{D}_{pub}$ = CelebA. The results conclusively show that our defense model is effective.

Attack	Defense	Acc ↑	AttAcc ↓	Δ ↑	KNN Dist ↑
GMI	NoDef	88.50	29.60 ± 5.43	-	1607.86
	TL-DMI	83.41	15.73 ± 4.58	2.72	1752.00
	MIDRE	85.74	**7.47 ± 2.59**	**8.02**	**1898.29**
KedMI	NoDef	88.50	81.67 ± 2.63	-	1270.71
	TL-DMI	83.41	73.40 ± 4.10	1.62	1265.00
	MIDRE	85.74	**42.93 ± 5.22**	**14.04**	**1512.52**
LOMMA + GMI	NoDef	88.50	83.33 ± 3.40	-	1259.61
	TL-DMI	83.41	43.67 ± 5.60	7.79	**1616.00**
	MIDRE	85.74	**43.33 ± 6.02**	**14.49**	1550.77
LOMMA + KedMI	NoDef	88.50	90.87 ± 1.31	-	1116.90
	TL-DMI	83.41	79.60 ± 1.78	2.21	**1345.00**
	MIDRE	85.74	**58.07 +/- 1.78**	**11.88**	1386.67

Table B.8: We report the PLGMI attacks on images with resolution 64×64. We compare to NoDef and BiDO methods. T = VGG16, IR152 and FaceNet64, D_{pub} = FFHQ. We remark that BiDO only releases their implementation and pretrained model in the setup of T = VGG16.

Architecture	Defense	Acc ↑	AttAcc ↓	Δ ↑	KNN Dist ↑
VGG16	NoDef	86.90	81.80 ± 2.74	-	1323.27
	BiDO	79.85	60.93 ± 3.99	2.96	1440.16
	MIDRE	79.85	**36.07 ± 4.76**	**6.49**	**1654.41**
IR152	NoDef	91.16	96.60 ± 2.11	-	1187.37
	MIDRE	84.91	**54.02 ± 4.86**	**6.81**	**1579.28**
FaceNet64	NoDef	88.50	95.00 ± 2.56	-	1250.90
	MIDRE	81.56	**51.60 ± 3.61**	**6.25**	**1501.85**

Table B.9: We report the PPA MI attacks on images with resolution 224×224. We compare the performance of these attacks against existing defenses including NoDef, MID, DP, BiDO NLS, TLDMI, and MI-RAD variances. D_{priv} = Facescrub D_{pub} = FFHQ, Arhchitecture is Resnet18, ResNet152 and ResNet101. We denote "NA" for δ_{face} and δ_{eval} if these numbers are not available in the official paper Ho et al. (2024); Koh et al. (2024); Struppek et al. (2024). We denote "OP" for Δ if the accuracy of the defense model outperforms that of the NoDef model.

Architecture	Defense	Acc ↑	AttAcc ↓	δ_{eval} ↑	δ_{face} ↑	Δ ↑
ResNet18	NoDef	94.22	88.67	123.85	0.74	-
	MID	91.15	65.47	137.75	0.87	7.56
	DP	89.80	75.26	130.41	0.82	3.03
	BiDO	91.33	76.56	127.86	0.75	4.54
	TL-DMI	91.12	22.36	NA	NA	21.39
	MIDRE(0.1, 0.4)	**97.28**	48.16	131.72	0.80	OP
	MIDRE(0.1,0.8)	93.33	**13.89**	**154.79**	**0.97**	**84.02**
ResNet152	NoDef	95.43	86.51	113.03	0.73	-
	MID	91.56	66.18	137.18	0.86	5.25
	BiDO	91.80	58.14	147.28	0.87	7.82
	NLS	91.50	**14.34**	NA	**1.23**	18.36
	RoLSS	93.00	64.98	NA	NA	8.86
	SSF	93.79	70.71	NA	NA	9.63
	TTS	93.97	73.59	NA	NA	8.85
	MIDRE(0.1,0.4)	**97.90**	42.44	139.66	0.82	OP
	MIDRE(0.1,0.8)	95.47	**15.97**	**155.61**	0.95	OP
ResNet101	NoDef	94.86	83.00	128.60	0.76	-
	MID	92.70	82.08	122.96	0.76	0.43
	DP	91.36	74.88	131.38	0.82	2.32
	BiDO	90.31	67.07	139.15	0.84	3.50
	TL-DMI	90.10	31.82	NA	NA	10.75
	NLS(-0.05)	94.79	33.14	130.94	0.90	712.29
	RoLSS	92.40	58.68	NA	NA	9.89
	SSF	93.79	71.06	NA	NA	11.16
	TTS	94.16	77.26	NA	NA	8.20
	MIDRE(0.1,0.4)	**98.02**	43.58	139.01	0.81	OP
	MIDRE(0.1,0.8)	95.15	**15.47**	**155.80**	**0.96**	OP

Table B.10: We report the PPA MI attacks on images with resolution 224×224. We compare the performance of these attacks against existing defenses including NoDef, MID, DP, BiDO NLS, TLDMI, and MI-RAD variances. D_{priv} = Facescrub D_{pub} = FFHQ, Arhchitecture is DenseNet169, DenseNet121, ResneSt101, and MaxVIT. We denote "NA" for δ_{face} and δ_{eval} if these numbers are not available in the official paper Ho et al. (2024); Koh et al. (2024); Struppek et al. (2024). We denote "OP" for Δ if the accuracy of the defense model outperforms that of the NoDef model.

Architecture	Defense	Acc ↑	AttAcc ↓	δ_{eval} ↑	δ_{face} ↑	Δ ↑
DenseNet169	NoDef	95.49	87.80	124.74	0.77	-
	RoLSS	72.14	6.77	NA	NA	3.47
	SSF	92.95	60.99	NA	NA	10.56
	MIDRE(0.1,0.4)	**97.99**	46.67	136.18	0.81	NA
	MIDRE(0.1,0.8)	95.04	**15.78**	**154.96**	**0.95**	**160.04**
DenseNet121	NoDef	95.54	95.13	116.14	0.68	-
	NLS(-0.05)	92.13	40.69	179.53	**0.97**	15.96
	RoLSS	74.25	10.24	NA	NA	3.99
	SSF	93.09	65.21	NA	NA	12.21
	MIDRE(0.1,0.4)	**98.19**	46.98	134.86	0.81	OP
	MIDRE (0.1,0.8)	95.76	15.66	**154.62**	0.96	OP
ResneSt101	NoDef	95.38	84.27	129.18	0.81	-
	NLS(-0.05)	88.82	13.23	**172.73**	**1.10**	10.01
	MIDRE(0.1,0.4)	**98.11**	45.43	137.78	0.80	NA
	MIDRE(0.1,0.8)	95.09	15.54	156.44	0.96	**237.00**
MaxVIT	NoDef	98.36	80.66	110.69	0.69	-
	TL-DMI	93.01	21.17	NA	NA	10.59
	NLS(-0.05)	98.23	55.09	127.68	0.81	63.93
	RoLSS	95.09	25.17	NA	NA	15.68
	MIDRE(0.1,0.4)	**98.46**	42.50	133.61	0.81	OP
	MIDRE(0.1,0.8)	96.52	13.92	**155.31**	**0.96**	31.63

B.3 User Study

In addition to attack accuracy measured by the evaluation model, we conduct a user study to further validate the attack's effectiveness. Overall, we conduct two setups for user study with low-resolution images and high-resolution images. Our interface for user study is illustrated in Fig. B.1.

In the low-resolution setup, we compare our proposed method and BiDO (Peng et al., 2022). For fair comparison, we use the same setup as BiDO: T = VGG16, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = CelebA and use the pre-trained model of BiDO to generate their images. We use the attack method PLG-MI to generate the inverted images and randomly select one image for each identity for overall 150 first identities. We upload it to Amazon Mechanical Turk and designate three individuals to vote on two of our model's and BiDO's reconstructed images, for a total of 450 votes. Participants were asked to select one of 4 options: BiDO, MIDRE, none, or both, for each image pair. Each pair was rated by three different users.

In the high-resolution setup, we compare MIDRE and TL-DMI (Ho et al., 2024), which is a state-of-the-art MI defense. We use the setup: T = ResNet101, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ, attack method = PPA. For every defense, we create inverted images for each of the 530 classes, then select one image for each class. Finally, we upload them to Amazon Mechanical Turk and follow the same procedure as low-resolution images setup.

Comparing BiDO and our proposed MIDRE: According to the results, 221 users voted in favour of BiDO, 108 in favour of our approach, 119 in favour of neither, and 2 in favour of both. It suggests that the reconstructed image quality from our model is not as good as the reconstructed image quality from BiDO, therefore **our proposed defense is more effective**. Our results are presented in Tab. B.11.

Comparing SOTA TL-DMI and our proposed MIDRE: According to the results in Tab. B.12, 509 users chose images inverted from our model, while 537 users voted in favor of TL-DMI. This suggests that the inverted images from our models are of lower quality than those from TL-DMI. In addition, there are 522 people voted for none of the two images is similar with the original image, meanwhile only 22 users chose that both images are similar to the real image.

According to the final results of both settings, MIDRE is a better defense mechanism against MI than SOTA BiDO and TL-DMI, which is in line with the findings of other evaluation metrics.

Table B.11: We report results for an user study that was performed with Amazon Mechanical Turk. Reconstructed samples of PLG-MI/VGG16/CelebA/CelebA with first 150 classes. The study asked users for inputs regarding the similarity between a private training image and the reconstructed image from BiDO trained model and our trained model. Less number of reconstructed images from our defensed model are selected by users, suggesting our defense is more effective.

Defense	Num of samples selected by users as more similar to private data
BiDO	221
Ours	**108**
Both	119
None	2

Table B.12: We report results for an user study that was performed with Amazon Mechanical Turk. Reconstructed samples of PPA/ResNet101/FaceScrub/FFHQ with all 530 classes. The study asked users for inputs regarding the similarity between a private training image and the reconstructed image from TL-DMI trained model and our trained model. Less number of reconstructed images from our defensed model are selected by users, suggesting our defense is more effective.

Defense	Num of samples selected by users as more similar to private data
TL-DMI	537
Ours	**509**
Both	522
None	22

B.4 Qualitative Results

We provide inversion results from the recent IF-GMI attack in Fig. B.2 (T = ResNet-18) and Fig. B.3 (T = ResNet-152) and $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ. These results further demonstrate the effectiveness of our proposed method.

C Additional analysis of privacy effect of MIDRE

C.1 Feature space analysis of Random Erasing's defense effectiveness

In addition to the visualization of feature space analysis in Sec. 3.2 (main paper), we provide more visualization in other setup: T = ResNet-152 (Simonyan & Zisserman, 2014), D_{priv} = Facecrub (Ng & Winkler, 2014), D_{pub} = FFHQ (Karras et al., 2019), attack method = PPA (Struppek et al., 2022). We observe **Property P1: Model trained with RE-private images following our MIDRE leads to a discrepancy between the features of MI-reconstructed images and that of private images**, resulting in degrading of attack accuracy.

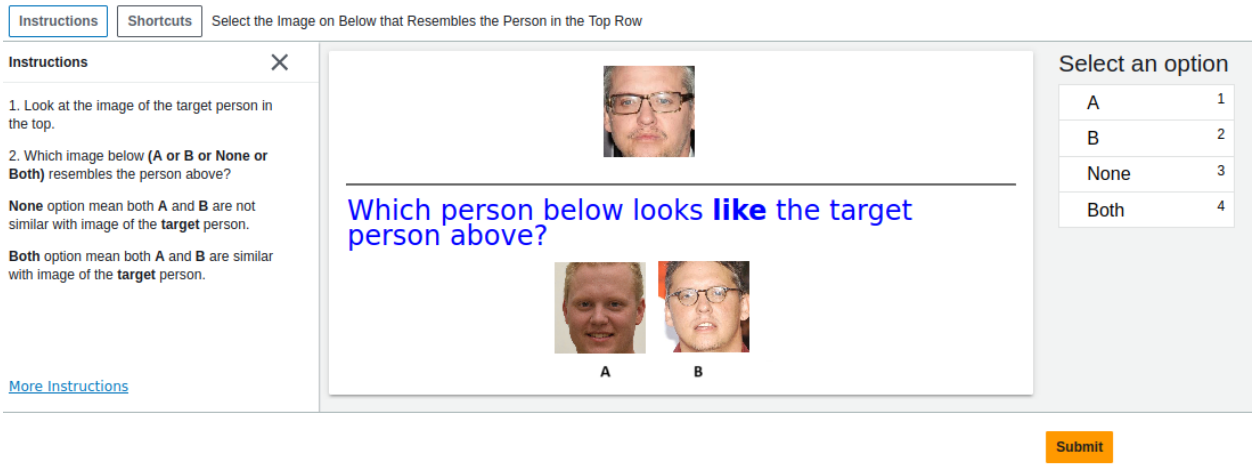


Figure B.1: Our Amazon Mechanical Turk (MTurk) interface for user study with model inversion attacking samples. Participants were asked to select one of 4 options: A, B, none, or both, for each image pair where A and B are the inverted images of our defense and other defense model. Each pair was rated by three different users.

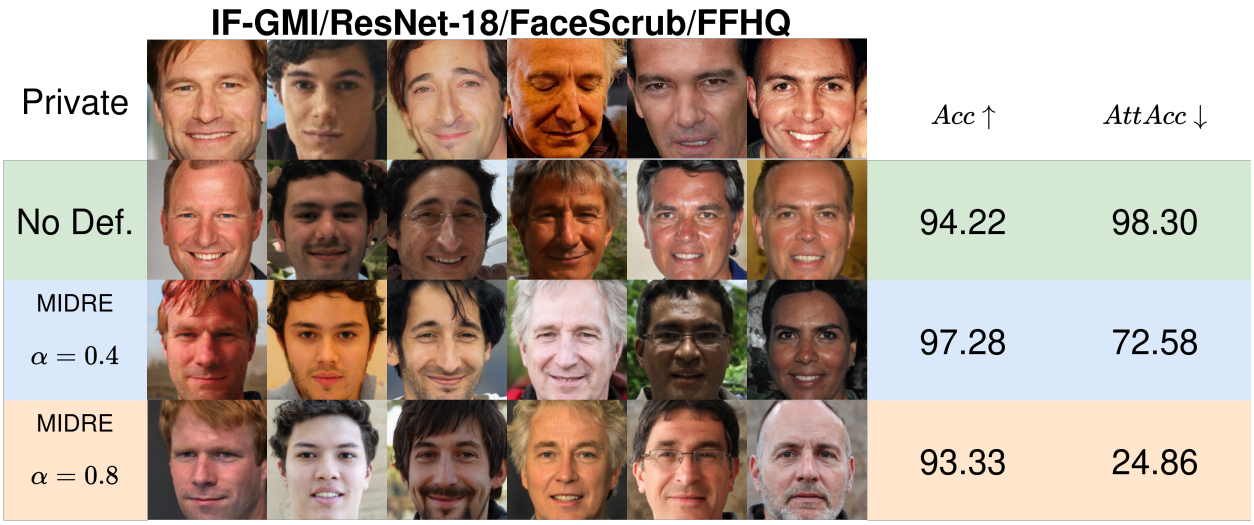


Figure B.2: Reconstructed image obtained from IF-GMI attack with T = ResNet-18, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ. The quality of the reconstructed image obtained from the attack on the model trained by MIDRE is comparatively worse when compared to that from NoDef method, suggesting the efficiency of our proposed defense MIDRE.

We use the following notation: f_{train}, f_{priv}, f_{RE}, and f_{recon} represent the features of training images, private images, RE-private images, and MI-reconstructed images, respectively. To extract these features, we first train the target model without any defense (NoDef) and another target model with our MIDRE. Then, we pass images into these models to obtain the penultimate layer activations. Specifically, we input private images into the models to obtain f_{priv}. Next, we apply RE to private images, pass these RE-private images into the models to obtain f_{RE}. We also perform MI attacks to obtain reconstructed images from NoDef model (resp. MIDRE model), and then feed them into the NoDef model (resp. MIDRE model) to obtain f_{recon}. Then, we visualize penultimate layer activations $f_{priv}, f_{RE}, f_{recon}$ by both NoDef and our MIDRE model. We use $a_e = 0.4$ to train MIDRE and to generate RE-private images. Additionally, we



Figure B.3: Reconstructed image obtained from IF-GMI attack with T = ResNet-152, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ. The quality of the reconstructed image obtained from the attack on the model trained by MIDRE is comparatively worse when compared to that from NoDef method, suggesting the efficiency of our proposed defense MIDRE.

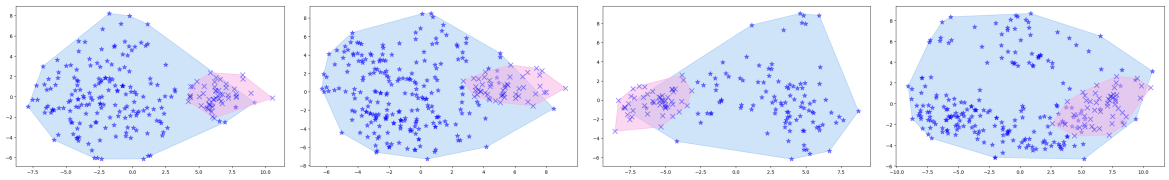
visualize the convex hull of these features. For visualization, we employ PCA to reduce the dimension of the feature space.

The visualization in Fig. C.4 shows the same trend as in Sec. 3.2. Specially, we observe the mismatch in feature space of MIDRE. Under MIDRE target model, f_{RE}^{MIDRE} and f_{priv}^{MIDRE} have partial overlaps, but they are not identical. Meanwhile, f_{recon}^{MIDRE} tend to match with f_{RE}^{MIDRE} (RE-private images are training data for MIDRE, and follows the discussion above). Therefore, f_{recon}^{MIDRE} do not replicate f_{priv}^{MIDRE}, significantly degrading the MI attack. Furthermore, Fig. C.5 shows that the mismatch between f_{RE}^{MIDRE} and f_{priv}^{MIDRE} does not cause the reduction of model utility. This is because the private images remain distinct from other classes and distant from other classification regions, even when their representations are partially overlapped with RE-private images (the training data).

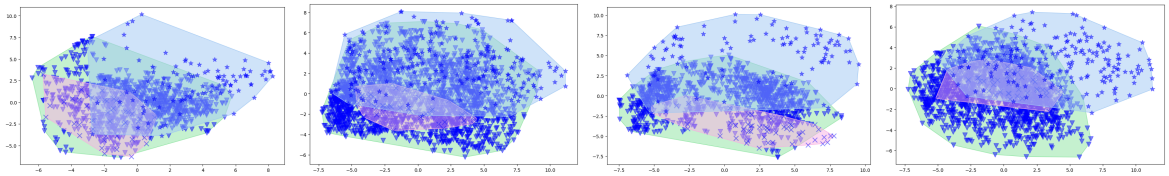
C.2 Importance of partial erasure and random location for privacy-utility trade-off

In this section, we analyse two properties of Random Erasing that are: **Property P2: Partial Erasure**, and **Property P3: Random Location**. In the addition of the setup in Sec. 3.2, we report a new experiment using the following setup: We use T = MaxVIT, D_{priv} = Facecrub (Ng & Winkler, 2014), D_{pub} = FFHQ (Karras et al., 2019), attack method = PPA (Struppek et al., 2022).

To evaluate the effectiveness of **Partial Erasure** and **Random Location**, we conduct experiments on three schemes: **Entire Erasing (EE), Fixed Erasing (FE)**, and **Random Erasing (RE)**. These schemes are compared against a No Defense baseline, which is trained for 100 epochs without any defense. In Entire Erasing (EE) scheme, we progressively reduce the number of training epochs to simulate different levels of pixel concealment. Specifically, we train the model for 50, 60, 70, 80, 90, and 100 epochs, corresponding to 50%, 40%, 30%, 20%, 10%, and 0% pixel concealment, respectively. For Fixed Erasing (FE), a fixed location within each image is erased throughout the entire training process. However, the erased location varied across different images. For Random Erasing (RE), the location of erased areas is randomly selected for each image and training iteration. We train the RE model for 100 epochs with different values of the erasure ratio, a_e = 0.5, 0.4, 0.3, 0.2, 0.1 corresponding to 50%, 40%, 30 %, 20%, and 10% pixel concealment, respectively.



(a) NoDef, T = ResNet152, $a_e = 0.4$, AttAcc = 86.51%, $f^{NoDef}_{train} = f^{NoDef}_{priv}$



(b) MIDRE, T = ResNet152, $a_e = 0.4$, AttAcc = 15.97%, $f^{MIDRE}_{train} = f^{MIDRE}_{RE}$

Figure C.4: **Feature space analysis to show that, under MIDRE, f^{MIDRE}_{recon} and f^{MIDRE}_{priv} has a discrepancy, degrading MI attack.** We visualize penultimate layer activations of private images ($\star$ f_{priv}), RE-private images ($\blacktriangledown$ f_{RE}), and MI-reconstructed images ($\times$ f_{recon}) generated by both (a) NoDef and (b) our MIDRE model. We also visualize the convex hull for private images, RE-private images, and MI-reconstructed images. In (a), f^{NoDef}_{recon} closely resemble f^{NoDef}_{priv}, consistent with high attack accuracy. In (b), private images and RE-private images share some similarity but they are not identical, with partial overlap between f^{MIDRE}_{priv} and f^{MIDRE}_{RE}. Importantly, f^{MIDRE}_{recon} closely resembles f^{MIDRE}_{RE} as RE-private is the training data for MIDRE. This results in **a reduced overlap between f^{MIDRE}_{recon} and f^{MIDRE}_{priv}, explaining that MI does not accurately capture the private image features under MIDRE.**

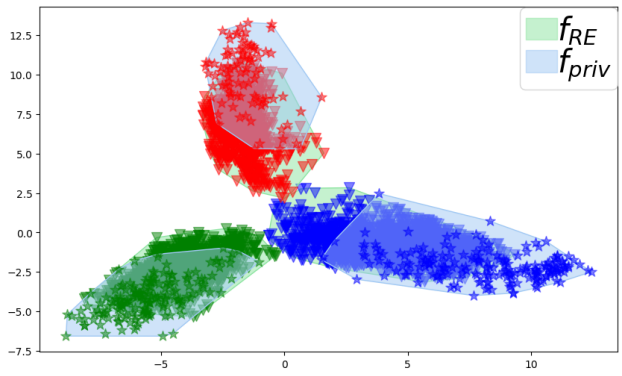


Figure C.5: **MIDRE target model achieves high accuracy despite partial overlap of f^{MIDRE}_{RE} and f^{MIDRE}_{priv}** using the target model T = ResNet152. We visualize the penultimate layer activations of RE-private images and private images for three identities. While f^{MIDRE}_{RE} and f^{MIDRE}_{priv} do not completely overlap, the model can still classify private images with high accuracy. This is because the private images remain distinct from other classes and distant from other classification regions, even when their representations are partially shared with RE-private images (the training data). We remark that RE randomly erases different regions from the images in different iterations, preventing the model to learn shortcut features and forcing the model to learn intr insic features and become more generalizable beyond training data.

We report the results in Tab. C.13. The results exhibit the same trend as outlined in Section 3.2 of the main paper. Specifically, **Property P2** demonstrates the privacy effect in defending against MI attacks, where partial erasure (fixed or random) proves more effective than entire erasure (reducing epochs) despite identical pixel concealment percentages. **Property P3** validates the recovery of model utility, evidenced by

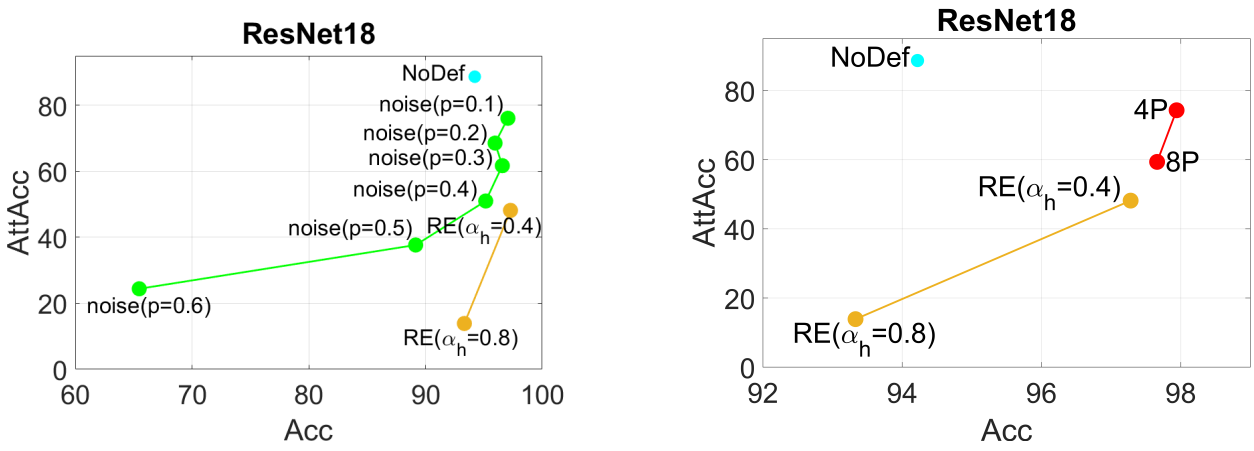
the enhanced accuracy of RE models while archiving lower attack accuracy than FE models across varying erased portion ratios a_e.

Table C.13: We compare three different techniques for pixel concealment, to reduce the amount of private information presented to the model during training. Here, we use T = MaxVIT, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ, attack method = PPA. The results show that simply reducing epochs as in "Entire Erasure" is insufficient for degrading attack performance. Meanwhile, RE improves model utility while degrading attack accuracy effectively.

Concealment	**Partial Erasure**					**Entire Erasing**	
		Random Erasing		**Fixed Erasing**			
	a_e	**Acc (↑)**	**AttAcc (↓)**	**Acc (↑)**	**AttAcc (↓)**	**Acc (↑)**	**AttAcc (↓)**
0%	0	98.36	80.66	98.36	80.66	98.36	80.66
10%	0.1	**98.73**	**70.92**	98.59	74.81	97.93	82.93
20%	0.2	**98.61**	**56.93**	98.11	57.74	98.00	82.17
30%	0.3	**98.35**	**42.10**	97.90	43.40	98.04	83.49
40%	0.4	**98.06**	**28.21**	97.31	31.06	97.95	83.18
50%	0.5	**98.73**	16.34	94.67	**16.11**	98.03	84.29

D Ablation Study

D.1 Ablation study on alternative masking strategies



(a) Masking Random Pixels

(b) Masking Multiple Patches

Figure D.6: We compare our RE masking strategy with two alternative masking approaches: (a) *Masking Random Pixels*, and (b) *Masking Multiple Patches*. In (a), we randomly mask a proportion of pixels from 10% ($p = 0.1$) to 60% ($p = 0.6$). In (b), we randomly mask either 4 small patches, denoted as 4P ($\alpha_e \in [0.025, 0.1]$), or 8 small patches, denoted as 8P ($\alpha_e \in [0.0125, 0.1]$). We evaluate these strategies using the PPA attack method with T = ResNet18, $\mathcal{D}_{\text{priv}}$ = FaceScrub, and $\mathcal{D}_{\text{pub}}$ = FFHQ. The results demonstrate that our RE masking strategy achieves a better privacy-utility trade-off compared to both Masking Random Pixels and Masking Multiple Patches.

In this section, we conduct experiments using alternative masking strategies. In addition to the traditional random erasing method, we explore two additional approaches: (1) masking random pixels, and (2) masking multiple patches.

- **Masking random pixels:** Instead of masking a square region as in our proposed Random Erasing (RE) method, we apply masking at the pixel level. For example, we randomly mask 10% of the

image pixels by replacing them with random values. In our experiments, we train the target model with varying levels of random pixel masking, ranging from 10% ($p = 0.1$) to 60% ($p = 0.6$).

- **Masking multiple patches:** Instead of masking a single large square region, we apply multiple smaller masks to the image. In our experiments, we randomly mask each training image with either 4 small patches (4P) or 8 small patches (8P). To ensure a fair comparison with MIDRE, we adjust the patch sizes accordingly. For 4P, we set $\alpha_e \in [0.025, 0.1]$, so that the total area of the four small patches is approximately equivalent to that of MIDRE with $\alpha_e \in [0.1, 0.4]$) (RE($\alpha_h = 0.4$)). Similarly, for 8P, we use $\alpha_e \in [0.0125, 0.1]$, making the total masked area comparable to MIDRE with $\alpha_e \in [0.1, 0.8]$ (RE($\alpha_h = 0.8$)).

We summarize the results of the two alternative masking strategies in Fig. D.6.

- **Masking Random Pixels:** We clearly observe that this method performs better than the baseline (NoDef) in terms of reducing attack accuracy. However, it is less effective than our proposed MIDRE in both lowering attack accuracy and preserving natural accuracy.
- **Masking Multiple Patches:** Although distributing the masking across multiple smaller regions provides some privacy benefits, masking a single large region—as done in our approach—still achieves a better utility-privacy trade-off.

D.2 Ablation study on Masking Values.

In this section, we examine the effect of masking value to MIDRE performance. We select attack method = PLGMI (Yuan et al., 2023), T = FaceNet64, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ. We set a_e = (0.2,0.2). Similar to (Zhong et al., 2020), we investigate four types of masking values: 0, 1, a random value, and the mean value. In case of random value, we randomly select it within a range (0,1). The mean value uses the ImageNet dataset's mean pixel values ([0.485, 0.456, 0.406]).

Tab. D.14 demonstrates that the mean value offers the best balance between robustness against MI attacks and maintaining natural image accuracy. Consequently, we adopt the Imagenet mean pixel values for masking in MIDRE.

Table D.14: The effect of different masking value. We use attack method = PLGMI (Yuan et al., 2023), T = FaceNet64, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ. Overall, mean value achieves the best balance between robustness against MI attacks and maintaining natural image accuracy.

Masking value	Acc ↑	AttAcc ↓	Δ ↑	Ranking
NoDef	88.50	95.00 ± 2.56	-	-
0	83.72	69.20 ± 2.64	5.40	3
1	83.68	70.00 ± 3.18	5.18	4
random	80.76	51.87 ± 4.43	5.57	2
mean	85.14	68.87 ± 3.97	**7.78**	1

D.3 Ablation study on Area Ratio.

In MIDRE, the area ratio a_e controls the portion of an image masked to prevent MI attacks. This experiment investigates the impact of different a_e values on MIDRE's performance. In particular, a_e is randomly selected within the range (0.1, a_h), guaranting that at least 10% of the image is always masked. We select three values for a_h: 0.3, 0.4, and 0.5. Similar to the previous ablation study, we employ attack method = PLGMI (Yuan et al., 2023), T = FaceNet64, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ. The masking process uses the ImageNet mean pixel values.

The results in Tab. D.15 indicate that increasing a_h strengthens MIDRE's defense against MI attacks, but this comes at the cost of reduced natural accuracy. To achieve a balance between robustness and natural accuracy, we opt $a_h = 0.4$ in MIDRE.

Table D.15: The effect of area ratio. We use attack method = PLGMI (Yuan et al., 2023), T = FaceNet64, $\mathcal{D}_{priv}$ = CelebA, $\mathcal{D}_{pub}$ = FFHQ. To achieve a balance between robustness and natural accuracy, we opt a_h = 0.4 in MIDRE.

a_h	Acc ↑	AttAcc ↓	Δ ↑	Ranking
NoDef	88.50	95.00 ± 2.56	-	-
0.3	83.55	65.07 ± 4.02	6.05	2
0.4	81.65	51.60 ± 3.61	**6.34**	1
0.5	78.50	45.40 ± 3.85	4.96	3

D.4 Ablation study on Aspect Ratio.

We perform an ablation study on the aspect ratio of random erasing for model inversion defense. The results presented in Tab. D.16 demonstrate that the influence of aspect ratio on attack accuracy is not as significant as that of area ratio.

Table D.16: We report the LOMMA+KedMI attacks on images with resolution 64×64. T = VGG16, D_{priv} = CelebA, D_{pub} = CelebA with different aspect ratios of RE in MIDRE. We also put NoDef result as a baseline.

Attack	Defense	Acc ↑	AttAcc ↓	Δ ↑	KNN Dist ↑
LOMMA+KedMI	NoDef	86.90	81.80 ± 1.44	-	1211.45
	MIDRE	79.85	43.07 ± 1.99	5.49	1503.89
	MIDRE(aspect ratio = 0.5)	81.32	49.13 ± 1.53	5.85	1424.40
	MIDRE(aspect ratio = 2.0)	81.65	51.87 ± 1.62	5.70	1440.00

D.5 Adaptive attack

We perform adaptive attacks in which the attacker knows the portions of the masking area a_e and uses it during inversion attacks. We use 2 setups: **Setup 1**: T = ResNet152, $\mathcal{D}_{priv}$ = Facescrub, $\mathcal{D}_{pub}$ = FFHQ, Attack method = PPA, image size = 224 × 224. **Setup 2**: T = VGG16, $\mathcal{D}_{priv}/\mathcal{D}_{pub}$ = CelebA, Attack method = LOMMA + KedMI, image size = 64 × 64. We use a_e = [0.1,0.8] and a_e = [0.1,0.4] for setup 1 and setup 2 to train MIDRE and during attack.

Adaptive attacks fail to enhance attack performance in both two experimental setups (See Tab. D.17). This may be due to the dynamic masking positions employed in each attack iteration, hindering the convergence of the inverted images. Overall, even when attackers are fully informed about RE and use this knowledge to design an adaptive MI attack, they still fail to achieve accurate inversion results.

We compare the loss curves of the adaptive and normal attacks in Fig. D.7. The results show that the dynamic masking positions in each iteration cause greater fluctuations in the adaptive attack loss compared to the normal attack. In addition, PPA already incorporates learning rate adjustments during inversion, which do not reduce the loss fluctuations.

Table D.17: We conduct the adaptive attacks where the attacker knows the masking area portions a_e and uses them during inversion attacks. **Adaptive attacks (Adapt.Att) fail to enhance attack performance in both setups.**

Setup	Attack	AttAcc
Setup 1	MIDRE	15.97
	MIDRE (Adapt.Att)	10.50 **(-5.47%)**
Setup 2	MIDRE	43.07
	MIDRE (Adapt.Att)	38.53 **(-4.54%)**

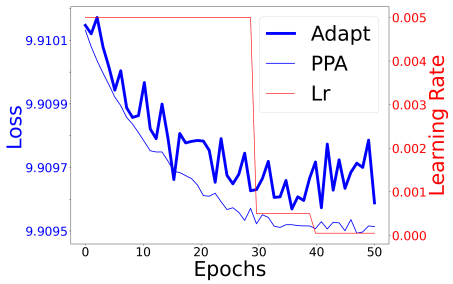


Figure D.7: PPA and PPA(Adapt) loss curves, with learning rate (Lr) adjustment.

D.6 The effectiveness of substitute pixels generated by inpainting for MIDRE.

We incorporated an inpainting method (Telea, 2004) to replace masked values, following the experimental setup described earlier. Our results show that MIDRE (inpainting) modestly improves model accuracy while reducing the attack success rate by 4.34%, which is indicated in Tab. D.18. However, this approach incurs a higher computational cost compared to RE.

Table D.18: We report the LOMMA+KedMI attack on images with resolution 64×64. T = VGG16, D_{priv} = CelebA, D_{pub} = CelebA to target models trained with RE with substitue pixel generate by inpaiting.

Attack	Defense	Acc ↑	AttAcc ↓	KNN Dist ↑
LOMMA+KedMI	NoDef	86.90	81.80 ± 1.44	1211.45
	MIDRE	79.85	**43.07 ± 1.99**	1503.89
	MIDRE (inpainting)	80.42	**38.73 ± 1.27**	**1508.28**

E Discussion

We propose a new defense against MI attacks using Random Erasing (RE) during training. RE reduces private information exposure while significantly lowering MI attack success, with small impact on model accuracy. Our method outperforms existing defenses across 34 experiment setups using 7 SOTA MI attacks, 11 model architectures, 6 datasets, and user study.

E.1 Broader Impacts

Model inversion attacks, a rising privacy threat, have garnered significant attention recently. By studying defenses against these attacks, we can develop best practices for deploying AI models and build robust safeguards for applications, especially those that rely on sensitive training data. Research on model inversion aims to raise awareness of potential privacy vulnerabilities and strengthen the defense.

E.2 Limitation

Firstly, we currently focus on enhancing the robustness of classification models against MI attacks. This is really important because these models are being used more and more in real-life situations where privacy and security are a major concern. In the future, we plan to expand our research scope to encompass MI attacks and defenses for a broader range of machine learning tasks.

Secondly, while our current experiments are comprehensive compared to prior works (Zhang et al., 2020; Chen et al., 2021; Nguyen et al., 2023; Kahla et al., 2022; Struppek et al., 2022; Ho et al., 2024; Koh et al., 2024) which mainly focus on image data, real-world applications often involve diverse private/sensitive training data. Addressing these real-world data complexities through a comprehensive approach will be essential for building robust and trustworthy machine learning systems across various domains.

F Experiments Compute Resources

In order to carry out our experiments, we utilise a workstation equipped with the Ubuntu operating system, an AMD Ryzen CPU, and 4 NVIDIA RTX A5000 GPUs. Furthermore, we utilise a secondary workstation equipped with the Ubuntu operating system, an AMD Ryzen CPU, and two NVIDIA RTX A6000 GPUs.

G Related Work

G.1 Model Inversion Attacks

The GMI (Zhang et al., 2020) is a pioneering approach in model inversion to leverages publicly available data and employs a generative model GAN to invert private datasets. This methodology effectively mitigates the generation of unrealistic data instances. KedMI (Chen et al., 2021) can be considered an enhanced iteration of the GMI model, as it incorporates the transmission of knowledge to the discriminator through the utilization of soft labels. PLGMI (Yuan et al., 2023) is the current leading model inversion method in the field. It leverages pseudo labels derived from public data and the target model. LOMMA (Nguyen et al., 2023) employs an augmented model in order to reduce the model inversion overfitting. The augmented model is trained to distill knowledge from a target model by utilizing public data. During attack, the attackers generate images in order to minimize the identity loss in both the target model and the augmented model. However, it should be noted that the aforementioned four approaches are specifically designed for target models that have been trained on low-resolution data, specifically 64x64 for the CelebA private dataset. Recently, PPA (Struppek et al., 2022), MIRROR (An et al., 2022), and DMMIA (Qi et al., 2023), IF-GMI(Qiu et al., 2024) perform the attack on high resolution images. In addition, Kahla et. al. (Kahla et al., 2022) introduced the BREPMI attack as a form of label-only model inversion attack, where the assault is based on the predicted labels of the target model. Another work is RLBMI (Han et al., 2023), which utilizes a reinforcement learning approach to target a model in a black box scenario.

G.2 Model Inversion Defenses

To defend against MI attacks, differential privacy (DP) (Dwork, 2006; 2008) has been studied in earlier works. Studies in (Dwork, 2006; 2008) have shown that current DP mechanisms do not mitigate MI attacks while maintaining desirable model utility at the same time. More recently, regularizations have been proposed for MI defenses (Wang et al., 2021; Peng et al., 2022; Struppek et al., 2024). (Wang et al., 2021) propose regularization loss to the training objective to limit the dependency between the model inputs and outputs. In BiDO (Peng et al., 2022), they propose regularization to limit the the dependency between the model inputs and latent representations. However, these regularizations conflict with the training loss and harm model utility considerably. To restore the model utility partially, (Peng et al., 2022) propose to add another regularization loss to maximize the dependency between latent representations and the outputs. However, searching for hyperparameters for two regularizations in BiDO requires computationally-expensive. Recently, (Ye et al., 2022) introduced a new approach that utilises differential privacy to protect against model inversion. (Gong et al., 2023) proposed a novel Generative Adversarial Network (GAN)-based approach to counter model inversion attacks. In this paper, we do not conduct experiments to compare to these methods as the code is not available. (Struppek et al., 2024) study the effect of label smoothing regularization on model privacy leakage. Their findings demonstrate that positive label smoothing factors can amplify privacy leakage, whereas negative label smoothing factors mitigate privacy concerns at the cost of a substantial decrease in model utility, resulting in a more favorable utility-privacy trade-off. Recently, (Ho et al., 2024) introduce a novel approach to defending against model inversion attacks by focusing on the model training process. Their proposed Transfer Learning-based Defense against Model Inversion (TL-DMI) aims to restrict the number of layers that encode sensitive information from the private training dataset into the model. As restricting the number of model parameters that encode private information can potentially impact the model's performance. (Koh et al., 2024) study the impact of DNN architecture designs, particularly skip connections, on model inversion attacks. They found that removing skip connections in the last layers can enhance model inversion robustness. However, this approach necessitates searching for optimal skip connection removal and scaling factor combinations, which can be computationally intensive. Both

TL-DMI and MI-RAD experiences difficulty in achieving competitive balance between utility and privacy. We show comparison of several defense approaches with our MIDRE in Fig. 1 (main paper).