HideMIA: Hidden Wavelet Mining for Privacy-Enhancing Medical Image Analysis

Xun Lin*
Beihang University
linxun@buaa.edu.cn

Yi Yu*
Nanyang Technological University
yuyi0010@e.ntu.edu.sg

Zitong Yu†
Great Bay University
yuzitong@gbu.edu.cn

Ruohan Meng
Nanjing University of Information Science and Technology
ruohanmeng.melody@gmail.com

Jiale Zhou
Beihang University
zhoujiale@buaa.edu.cn

Ajian Liu
Institute of Automation, Chinese Academy of Sciences
ajian.liu@ia.ac.cn

Yizhong Liu
Beihang University
liuyizhong@buaa.edu.cn

Shuai Wang
Beihang University
wangshuai@buaa.edu.cn

Wenzhong Tang
Beihang University
tangwenzhong@buaa.edu.cn

Zhen Lei
MAIS, CASIA
School of Artificial Intelligence, UCAS
CAIR, HKISI, CAS
zlei@nlpr.ia.ac.cn

Alex Kot
Rapid-Rich Object Search Lab (ROSE),
Nanyang Technological University
eackot@ntu.edu.sg

Abstract

Despite the advancements that deep learning has brought to medical image analysis (MIA), protecting the privacy of images remains a challenge. In a client-server MIA framework, especially after deployment, patients' private medical images can be easily captured by attackers from the transmission channel or malicious third-party servers. Previous MIA privacy-enhancing methods, whether based on distortion or homomorphic encryption, expose the fact that the transmitted images are medical images or transform the images into semantic-lacking noise. This tends to alert attackers, thereby falling into a *cat-and-mouse game* of theft and protection. To address this issue, we propose a covert MIA framework based on deep image hiding, namely HideMIA, which secures medical images by embedding them within natural cover images that are unlikely to raise suspicion. By directly analyzing the hidden medical images in the steganographic domain, HideMIA makes it difficult for attackers to notice the presence of medical images. Specifically, we propose the Mixture-of-Difference-Convolutions (MoDC) and Asymmetric Wavelet Attention (AsyWA) to enable HideMIA to conduct fine-grained analysis on each wavelet sub-band within the steganographic domain, mining features that are specific to medical images. Moreover, to reduce resource consumption on client devices, we design function-aligned knowledge distillation to obtain a lightweight hiding network, namely LightIH. Extensive experiments on six medical datasets demonstrate that our HideMIA achieves superior MIA performance and protective imperceptibility on medical image segmentation and classification.

CCS Concepts

• **Computing methodologies** → **Computer vision problems**.

Keywords

Image hiding, medical image analysis, privacy-enhancing

ACM Reference Format:
Xun Lin, Yi Yu, Zitong Yu, Ruohan Meng, Jiale Zhou, Ajian Liu, Yizhong Liu, Shuai Wang, Wenzhong Tang, Zhen Lei, and Alex Kot. 2024. HideMIA: Hidden Wavelet Mining for Privacy-Enhancing Medical Image Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680806

1 Introduction

Deep learning technologies have rapidly advanced in Medical Image Analysis (MIA) and have become an integral part of modern medical diagnostics [19]. These technologies significantly improve the accuracy and efficiency of clinical diagnosis, *e.g.* early screening for major diseases such as cancer [31]. Despite the impressive performance of deep neural networks, there are increasing concerns about the security issues [30, 39, 49, 50, 56–58] associated with artificial intelligence. Among these security issues, the protection of patient privacy remains a severe challenge [38]. From a legal and ethical perspective, patients' medical images must be rigorously protected, ensuring the security and privacy of these images during their storage, transmission, and analysis processes

*Both authors contributed equally to this paper. †Corresponding author.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680806

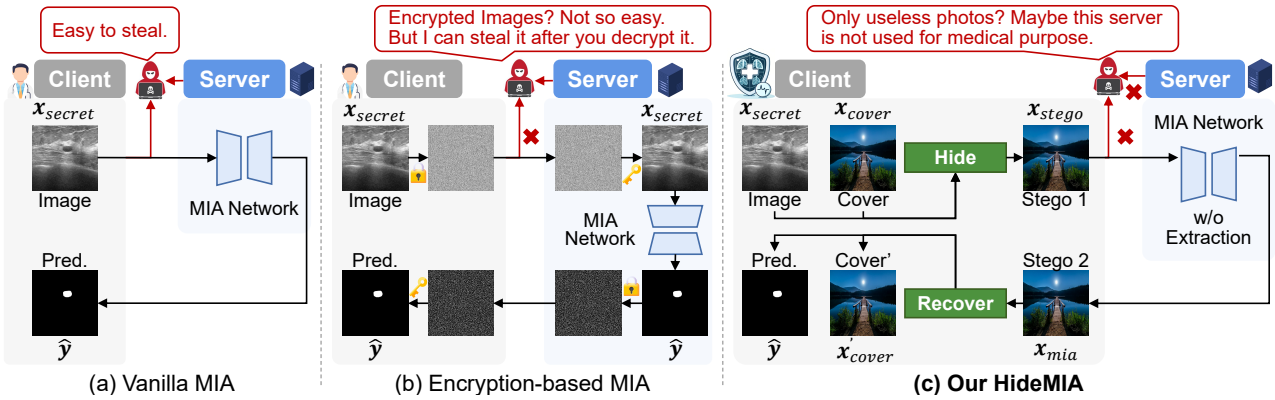

Figure 1: Different client-server frameworks for MIA, including (a) vanilla, (b) encryption-based, and (c) the proposed HideMIA framework. Our HideMIA makes the attacker hardly notice the existence of medical images.

[18]. Some medical institutions have adopted methods such as data anonymization or pseudonymization [3] to mitigate the risk of data leakage. However, real-world scenarios indicate that these methods are not robust against re-identification attacks [19]. As illustrated in Fig. 1(a), there is a risk of leakage, especially during the transmission of images from a client to a server [2]. Although techniques like non-homomorphic image encryption can prevent image leakage during transmission (see Fig. 1(b)), they are not robust to attacks from malicious servers [20]. Once an attacker controls the server by system vulnerabilities or embedded malicious backdoors (as a third-party server provider), the decrypted images on the server-side are also at the risk of being captured.

Currently, most image privacy enhancement technologies focus on the training phase, for instance, by enabling different institutions to collaboratively train models without the direct exchange of image data through federated learning [11], or by generating unlearnable examples to prevent unauthorized model training [30]. These methods effectively enhance the privacy protection of training data, yet image protection after the model deployment remains a challenge. Some studies focus on protecting images after deployment through distortion-based [9, 20, 40] methods to modify detailed information of the original images. Although these methods somewhat protect image privacy, attackers can easily notice that the transmitted images are medical images. To further enhance security, recent works propose encoding-based [21, 45] and homomorphic-encryption-based [55, 63] methods to protect medical images. However, these methods significantly alter the distribution of the images, easily alerting attackers to the fact that the images are protected by specific techniques, thus falling into a *cat-and-mouse game* of encryption and decryption. Among these methods, methods based on homomorphic encryption are time-consuming, making them impractical for many real-world applications.

Recent progress in deep image hiding (DIH) networks [12, 17, 26], which are proposed to conceal a secret image within a cover image for covert image transmission, has inspired us to propose a novel framework to perform **Hidde**n **M**edical **I**mage **A**nalysis (**HideMIA**) based on DIH. As shown in Fig. 1(c), HideMIA hides a medical image within a natural photo, allowing for direct MIA in the steganographic domain without the image extraction and decryption on the server-side. However, directly performing MIA in the steganographic domain is challenging because the medical-specific features within stego images are subtle, and the conspicuous information from the cover images introduces noise to the MIA.

Considering that DIH networks tend to hide the information of the secret image in high-frequency parts of the cover images [17], we design the spectral-aware **M**ixture-**o**f-**D**ifference-**C**onvolutions (**MoDC**) and **Asy**mmetric **W**avelet **A**ttention (**AsyWA**) to solve the problems above. These two modules are integrated with HideMIA's server-side MIA network. They can perform band-adaptive analysis on stego images in the steganographic domain (targeting different bands after discrete wavelet transform [14]). Specifically, within MoDC, inspired by [23, 29, 46, 59], we introduce various novel pixel difference convolutions to extract features of the hidden medical images in a fine-grained manner. Considering that the content and form of hidden information may vary across different bands, we bring the idea of the Mixture-of-Experts (MoE) [24, 62] to these convolutions, enabling adaptive combinations of these difference convolutions for each frequency band. Meanwhile, we propose inter-band and inner-band cross-attention with AsyWA to enable a global perception of concealed medical images across spatial and spectral dimensions. Considering that most of the semantic information in the cover belongs to the low-frequency bands, we design an asymmetric interaction constraint for inter-band cross-attention to ensure global perception while preventing cover-specific features in the low-frequency band from hindering the extraction of medical-related features in the high-frequency bands.

Moreover, existing DIH networks are not resource-friendly for clients with limited computing resources, *e.g.,* medical imaging devices. Therefore, motivated by [33], we propose the function-aligned knowledge distillation to obtain a **light**weight D**IH** network, namely **LightIH**. Unlike the feature-aligned knowledge distillations, our function-aligned distillation ensures that the student network learns sensitive features that influence analysis performance and imperceptibility during the distillation process. We employ the widely-used INN-based and spectrum-aware DIH network, HiNet, which is sensitive to manipulations in the steganographic domain [8], as the teacher network. To the best of our knowledge, this is the first work to develop a lightweight DIH network through knowledge distillation.

Our contributions can be summarized as follows:

• We propose a privacy-enhancing client-server MIA framework, namely HideMIA, to covertly analyze medical images without raising the suspicions of attackers.

• We design MoDC and AsyWA for wavelet-transformed stego images to ensure that HideMIA can accurately perform MIA in the steganographic domain without interference from conspicuous information in the covers.

• We develop LightIH, which, through the proposed function-aligned distillation, reduces the resource consumption of the DIH network while enhancing the accuracy of the MIA process.

• Extensive experiments on three image segmentation datasets and three image classification datasets demonstrate the effectiveness and imperceptibility of HideMIA.

2 Related Works

2.1 Privacy Enhancement for Medical Images

Distortion-based Methods. To prevent threats from transmission channels and malicious servers, privacy-preserving client-server frameworks are proposed for medical image analysis. Kim et al. [22] protect the privacy of medical images in the segmentation pipeline by adding the images sent from the client with reference images. Kim et al. [20], Packhäuser et al. [40] propose deformation generators to produce pseudo-random non-linear deformation for images from the client, which allows both distortion and recovery of the sensitive information of medical images and segmentation results. These distortion-based methods can defend against reconstruction and re-identification attacks, however, they still let attackers recognize that the transmitted images are medical.

Encoding-based and Encryption-based Methods. To further enhance privacy, encoding-based and encryption-based are proposed to change the content distribution of medical images. [21] design an encoder to remove identity-related information from medical images and propose a discriminator to identify ROI. Shiri et al. [45] adopt a learnable auto-encoder that employs convolution operations for the sparse transformation of medical images and adds pseudo-random noise to further obfuscate them. The effectiveness of homomorphic encryption is discussed in [55, 64] to further improve privacy protection for distributed medical image segmentation. Although encoding-based and encryption-based methods achieve better protective performance, these methods alter the image into feature maps or noise that are difficult for humans to understand. This may easily alert attackers and prompt them to design more threatening attacks.

2.2 Deep Image Hiding

Image hiding aims to covertly conceal a secret image within a cover image and enables the extraction of the hidden secret image. Baluja [4], Hayes, and Danezis [13] are the first to propose DIH networks based on the encoder-decoder structure. Liu et al. [36] improve the encoder based on U-Net and discrete wavelet transformation (DWT) [14] to embed the secret image, further enhancing the reversibility. Recent progress on invertible neural networks (INNs) in various image-to-image tasks [51, 52] has inspired the application of INNs in image hiding. Lu et al. [37] design INN-based image hiding by modeling the embedding and extraction processes as the forward and inverse operations in affine transformations. To enhance reversibility and imperceptibility, Deng et al. [8, 12, 17] input the wavelet-transformed cover image and secret image into an INN-based image hiding network, thereby embedding the secret into the high-frequency components of the cover. Xu et al. [53] propose a conditional normalizing flow to model the distribution of the redundant high-frequency component conditioned on the cover images, enhancing robustness against distortion.

3 HideMIA

3.1 Overall Framework

Our HideMIA consists of image hiding network $\mathcal{H}(\cdot,\cdot)$, medical image analysis network $\mathcal{M}(\cdot)$, and image recovering networks $\mathcal{R}(\cdot)$. The image hiding and image recovering networks are deployed on the client-side, while the medical image analysis network operates on the server. Let $\boldsymbol{x}_{secret}$ denote the image captured by medical imaging devices, and $\boldsymbol{x}_{cover}$ represent the cover image (**which can be any natural image**) adopted to conceal $\boldsymbol{x}_{secret}$ within. The process $\boldsymbol{x}_{stego} = \mathcal{H}(\boldsymbol{x}_{secret}, \boldsymbol{x}_{cover})$ yields a stego image, wherein $\boldsymbol{x}_{secret}$ is embedded into $\boldsymbol{x}_{cover}$. After the concealment, $\boldsymbol{x}_{stego}$ is transmitted from the client to the server, where it is directly fed into the MIA network: $\boldsymbol{x}_{mia} = \mathcal{M}(\boldsymbol{x}_{stego})$. The MIA network analyzes the concealed medical image within the steganographic domain. Upon returning $\boldsymbol{x}_{mia}$ to the client, the recovery network extracts the hidden analysis results as $\hat{\boldsymbol{y}} = \mathcal{R}(\boldsymbol{x}_{mia})$. In this pipeline, both $\boldsymbol{x}_{mia}$ and $\boldsymbol{x}_{stego}$ are easily captured by attackers, thus we require $\boldsymbol{x}_{mia}$ and $\boldsymbol{x}_{stego}$ remain visually indistinguishable from $\boldsymbol{x}_{cover}$ to ensure covert MIA without raising the attackers' suspicion.

3.2 MIA in the Steganographic Domain

We observe that existing MIA methods struggle to conduct precise and covert analysis within the steganographic domain. They often produce numerous false alarms and struggle to maintain visual consistency between $\boldsymbol{x}_{mia}$ and $\boldsymbol{x}_{stego}$, primarily due to the presence of conspicuous information in the cover image. Consequently, we propose the Mixture-of-Difference-Convolutions (MoDC) and Asymmetric Wavelet Attention (AsyWA) to perform fine-grained and cover-agnostic MIA in the steganographic domain. In our MIA network, $\boldsymbol{x}_{stego}$ is decomposed into four wavelet sub-bands using DWT: $\boldsymbol{x}^{LL}_{stego}, \boldsymbol{x}^{LH}_{stego}, \boldsymbol{x}^{HL}_{stego}, \boldsymbol{x}^{HH}_{stego} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$. These represent low (L) and high (H) frequencies across horizontal and vertical directions, detailing $\boldsymbol{x}_{stego}$'s frequency spectrum diversely. We utilize U-Net [44] as the backbone for each sub-network, with the output dimension segmentation heads set to 3. Both MoDC and AsyWA are integrated within each sub-network. The outputs from these sub-networks, namely $\boldsymbol{x}^{LL}_{mia}, \boldsymbol{x}^{LH}_{mia}, \boldsymbol{x}^{HL}_{mia}, \boldsymbol{x}^{HH}_{mia} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$, are combined using Inverse Wavelet Transform (IWT) to compose $\boldsymbol{x}_{mia}$, and then returned to the client.

Besides, we notice the guidance of original $\boldsymbol{x}^{LL}_{stego}$ can significantly improve the visual similarity between $\boldsymbol{x}_{mia}$ and $\boldsymbol{x}_{stego}$, thereby enhancing the HideMIA's imperceptibility. Specifically, since most of the semantic information in the cover, rather than the secret, is distributed within the LL band, we allow the LL sub-network to learn the residual of $\boldsymbol{x}^{LL}_{stego}$ as follows:

$$\boldsymbol{x}_{mia} = \mathrm{IWT}(\boldsymbol{x}^{LL}_{stego} + \boldsymbol{x}^{LL}_{mia},\ \boldsymbol{x}^{LH}_{mia},\ \boldsymbol{x}^{HL}_{mia},\ \boldsymbol{x}^{HH}_{mia}). \tag{1}$$

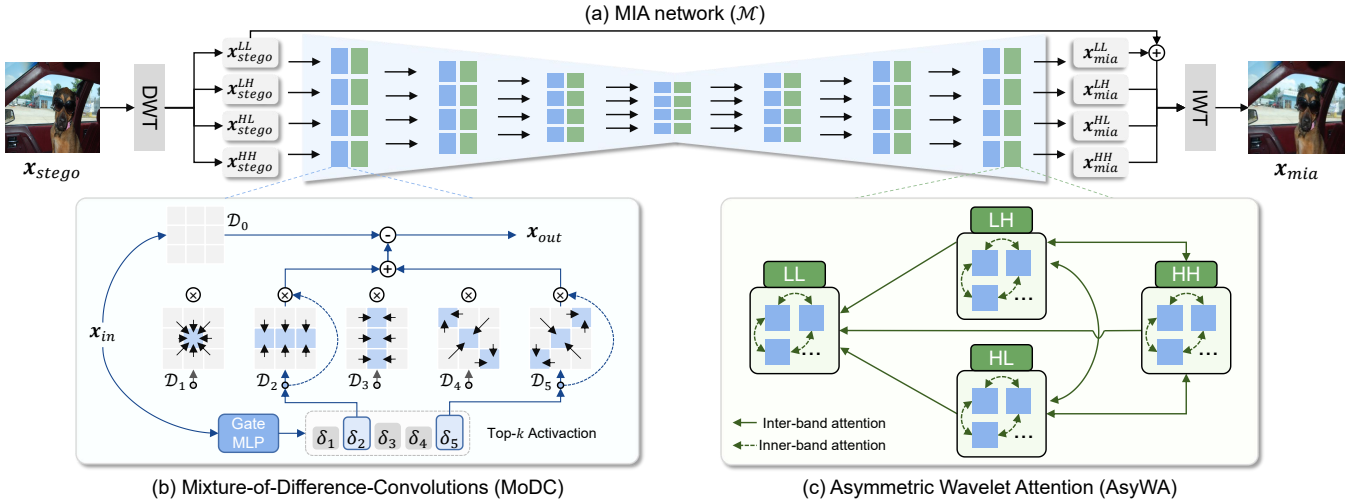

Figure 2: Illustration of (a) the server-side MIA network of the proposed HideMIA, (b) MoDC, and (c) AsyWA.

This operation has little impact on MIA performance and significantly improves HideMIA's imperceptibility (even feasibility). In the following sections, we provide detailed descriptions of the proposed MoDC and AsyWA.

Mixture-of-Difference-Convolutions. To ensure imperceptibility and recoverability, DIH networks tend to conceal different components of the secret image within different frequency bands of the cover images in different manners [25]. This makes the thorough analysis of the concealed information within each band using the same type of convolution operation challenging. Drawing inspiration from Central Difference Convolution (CDC) in detail-required vision tasks (*e.g.,* face anti-spoofing [59]), we propose four novel directional difference convolution operations, *i.e.,* horizontal, vertical, left diagonal, and right diagonal, to adapt to different bands generated by DWT, accommodating the high- and low-frequency bands in horizontal and vertical orientations.

Additionally, inspired by the idea of Mixture-of-Experts (MoE) [32, 62], we design a novel structure based on the newly introduced directional difference convolution operations, namely MoDC. As shown in Fig. 2(b), MoDC enables the network to adaptively combine the differential convolution operations that are most apt for each frequency band. Specifically, the vanilla convolution $\mathcal{D}_0$ and the aforementioned five difference convolutions, *i.e.,* $\mathcal{D}_1$ (central), $\mathcal{D}_2$ (horizontal), $\mathcal{D}_3$ (vertical), $\mathcal{D}_4$ (left diagonal), and $\mathcal{D}_5$ (right diagonal), can be described as follows:

$$
\begin{aligned}
\mathcal{D}_0(r_x, r_y) &= \sum_{(\Delta r_x, \Delta r_y)\in\mathcal{R}} w(\Delta r_x, \Delta r_y)\cdot \boldsymbol{x}_{in}(r_x - \Delta r_x, r_y - \Delta r_y),\\
\mathcal{D}_1(r_x, r_y) &= \sum_{(\Delta r_x, \Delta r_y)\in\mathcal{R}} w(\Delta r_x, \Delta r_y)\cdot \boldsymbol{x}_{in}(r_x, r_y),\\
\mathcal{D}_2(r_x, r_y) &= \sum_{(\Delta r_x, \Delta r_y)\in\mathcal{R}} w(\Delta r_x, \Delta r_y)\cdot \boldsymbol{x}_{in}(\Delta r_x, r_y),\\
\mathcal{D}_3(r_x, r_y) &= \sum_{(\Delta r_x, \Delta r_y)\in\mathcal{R}} w(\Delta r_x, \Delta r_y)\cdot \boldsymbol{x}_{in}(r_x, \Delta r_y),\\
\mathcal{D}_4(r_x, r_y) &= \sum_{(\Delta r_x, \Delta r_y)\in\mathcal{R}} w(\Delta r_x, \Delta r_y)\cdot \boldsymbol{x}_{in}(\Delta r_x + \Delta r_y, \Delta r_x + \Delta r_y),\\
\mathcal{D}_5(r_x, r_y) &= \sum_{(\Delta r_x, \Delta r_y)\in\mathcal{R}} w(\Delta r_x, \Delta r_y)\cdot \boldsymbol{x}_{in}(\Delta r_x - \Delta r_y, \Delta r_y - \Delta r_x),
\end{aligned}
\tag{2}
$$

where $\mathcal{R} = \{(1,1),(0,1),\cdots,(-1,0),(-1,-1)\}$ is the local respective field of the trainable 3×3 vanilla convolution kernel w, and r_x, r_y denote the current position of the kernel conducting on $\boldsymbol{x}_{in}$ and $\boldsymbol{x}_{out}$. The Gate Multi-Layer Perception (MLP) [48] in Fig. 2 is composed of a sequence of layers: pooling, linear, GELU activation, another linear, and finally Softmax. Gate MLP receives the input feature x_{in} and outputs the weights $\delta \in \mathbb{R}^5$ of five difference convolutions. MoDC then selects $P = \text{top}_k(\delta)$ kernels for activation and normalize their weights. This step is designed to prevent kernels that are not suitable for covert information analysis in the current band. Only the most appropriate difference kernels are conducted on the input features $\boldsymbol{x}_{in}$. Subsequently, the outputs of the activated difference convolutions are weighted by their corresponding weights δ_i, summed, and then subtracted from the output of the vanilla convolution kernel $\mathcal{D}_0$, completing the spectrum-aware difference convolution. MoDC is formulated as follows:

$$
\boldsymbol{x}_{out} = \mathcal{D}_0(\boldsymbol{x}_{in}) - \lambda \cdot \sum_{i\in S} \frac{\delta_i}{\sum_{j\in P}\delta_j}\cdot \mathcal{D}_i(\boldsymbol{x}_{in}), \tag{3}
$$

where λ is used to trade-off between the intensity of vanilla convolution and difference convolutions.

Asymmetric Wavelet Attention. As mentioned above, different components of medical images are concealed in varying forms within different frequency bands of the cover image. Therefore, enabling inter-band interactions to acquire a band-wise global perception of the concealed medical images is important. Previous works [12, 17] on DIH validate that the secret image is primarily concealed within the high-frequency parts of an image, *e.g.* within the non-LL bands in wavelet-transformed images. Consequently, during inter-band interactions, the semantic information within the low-frequency band can easily interfere with the fine-grained analysis of other bands.

To solve this problem, we propose the AsyWA, which allows inter-band cross-attention [6] among the features of different frequency bands, achieving global-spectrum perception. Bidirectional feature interactions are permitted among the high-frequency bands (*i.e.,* LH, HL, and HH), but the LL band can only receive cues from

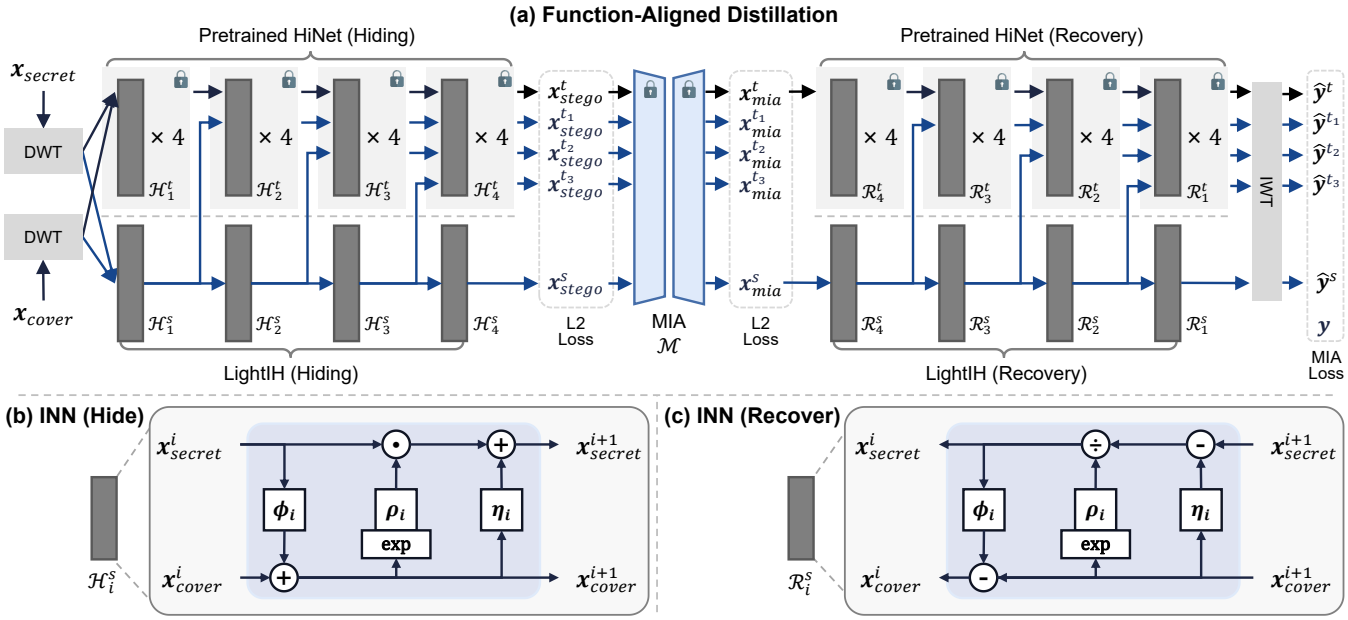

Figure 3: Illustration of LightIH. (a) Function-aligned distillation. (b) Invertible neural network for image hiding. (c) Invertible neural network for image recovery.

the high-frequency bands. AsyWA prevents the cover-specific semantic features mined from the low-frequency spectrum during MIA from misleading the extraction of the medical-specific fine-grained information. Additionally, considering the significance of spatial correlations between patches within the same band, we also integrate inner-band attention within AsyWA. Both our inter-patch and inner-band attention mechanisms are implemented based on multi-head attention [35]. AsyWA with inter- and inner-band attention can be formulated as follows:

$$\mathrm{MHA}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{Q(\boldsymbol{x}_2)K(\boldsymbol{x}_1)^\top}{\sqrt{d}} V(\boldsymbol{x}_1), \tag{4}$$

$$\boldsymbol{x}_{out}^{LL} = \boldsymbol{x}_{in}^{LL} + \textstyle\sum_{b\in B} \mathrm{MHA}(\boldsymbol{x}_{in}^{LL}, \boldsymbol{x}_{in}^{b}), \tag{5}$$

$$\boldsymbol{x}_{out}^{h} = \boldsymbol{x}_{in}^{h} + \textstyle\sum_{h'\in H} \mathrm{MHA}(\boldsymbol{x}_{in}^{h}, \boldsymbol{x}_{in}^{h'}), \quad \forall h \in H, \tag{6}$$

where $B = \{LL, HL, LH, HH\}$ denotes the band set of DWT, $H = \{HL, LH, HH\}$ represents the high-frequency band set, d denotes the number of pixels in the input feature, Q, K, V are the linear projections corresponding to the *query*, *key*, and *value*, respectively.

3.3 Lightweight DIH Network

To make the $\mathcal{H}$ and $\mathcal{R}$ resource-friendly for clients, we perform knowledge distillation [60] on the state-of-the-art (SOTA) DIH network, HiNet [17], to obtain a lightweight network, namely LightIH. Please note that HiNet also integrates DWT to improve stego imperceptibility. Existing feature distillation methods for vision tasks commonly [10] use distance such as L2 to align the features at corresponding stages between the student and teacher networks [27]. However, within the HideMIA framework, which includes multiple parts, *i.e.,* $\mathcal{M}$, $\mathcal{H}$, and $\mathcal{R}$, changing the intermediate features in different directions by the same L2 distance at early stages can lead to big differences in the final predictions [34].

To this end, we propose a function-aligned distillation strategy, where we encourage feature alignment solely from the perspective of DIH and MIA performance. This means that the features or outputs from corresponding stages of the teacher and student networks, once fed into the latter part of the same network, should yield closely similar results. Our function-aligned distillation encourages the student to focus more on the sensitive directions concluded by the teacher that significantly impact HideMIA's performance, rather than overly emphasizing the replication of the teacher's intermediate outputs. To ensure HideMIA possesses better MIA capabilities and imperceptibility, we combine the distillation loss, visual consistency loss, and MIA loss to supervise LightIH.

As illustrated in Fig. 3, we adopt the pretrained HiNet (16 layers) as the teacher network. Our student network LightIH consists of 4 INN layers (a quarter of HiNet). We conduct the function-aligned distillation across all phases of HideMIA (hiding, MIA, and recovery). Specifically, both the teacher and student networks are divided into four stages. In the teacher network, each stage consists of 4 INNs, denoted as $\mathcal{H}_i^t$ for hiding and $\mathcal{R}_i^t$ for recovery, while in the student network, each stage includes only one INN ($\mathcal{H}_i^s$ and $\mathcal{R}_i^s$). Given that the two DIH backbone are INN-based, the convolution layer weights (ϕ_i, ρ_i, and η_i) in $\mathcal{H}_i$ and $\mathcal{R}_i$ are shared. As shown in Figs. 3(b)-(c), only the sequence and operations applied to the inputs differ. Taking the hiding stage as an example, we feed the middle output from the student's $\mathcal{H}_i^s$ into the teacher's $\mathcal{H}_{i+1}^t$ and pass it through the subsequent blocks of the teacher to obtain $\boldsymbol{x}_{stego}^{t_i}$. As formulated in Fig. 3(a) and Eq. (7), these stego images and $\boldsymbol{x}_{stego}^{s}$ are then supervised with the teacher network's original output $\boldsymbol{x}_{stego}^{t}$ to guide the function alignment of image hiding.

Similarly, after passing through the MIA network, we use a comparable function-aligned supervision (see Eq. (7)). In the recovery

Algorithm 1: Training Process of HideMIA

Input : Teacher DIH networks $\mathcal{H}^t$,$\mathcal{R}^t$, student DIH networks $\mathcal{H}^s$,$\mathcal{R}^s$, MIA network $\mathcal{M}$, training dataset $\mathcal{T}$, epoch numbers of stages 1, 2, and 3: e_1, e_2, e_3, respectively, and learning rates $\ell_r^1, \ell_r^2, \ell_r^3$.

for $j \leftarrow 1$ *to* $e_1 + e_2 + e_3$ **do**
  for $\{\boldsymbol{x}_{cover}, \boldsymbol{x}_{secret}, \boldsymbol{y}\} \in \mathcal{T}$ **do**
    $\boldsymbol{x}_{stego}^t \leftarrow \mathcal{H}^t(\boldsymbol{x}_{cover}, \boldsymbol{x}_{secret})$;
    $\boldsymbol{x}_{mia}^t \leftarrow \mathcal{M}(\boldsymbol{x}_{stego}^t)$;
    $\hat{\boldsymbol{y}}^t \leftarrow \mathcal{R}^t(\boldsymbol{x}_{mia}^t)$;
    if $j \le e_1$ **then**
      # Stage 1: optimizing $\mathcal{M}$
      $\mathcal{L}_{total} \leftarrow \beta_1 \cdot \mathcal{L}_{mia}(\hat{\boldsymbol{y}}^t, \boldsymbol{y}) + \beta_2 \cdot \mathcal{L}_{percept}(\boldsymbol{x}_{mia}^t, \boldsymbol{x}_{cover})$
      $\theta_{\mathcal{M}} \leftarrow \theta_{\mathcal{M}} - \ell_r^1 \cdot \nabla_{\theta_{\mathcal{M}}} \mathcal{L}_{total}$
    else if $j \ge e_1$ **then**
      # Stage 2: using $\mathcal{H}^t$ and $\mathcal{R}^t$ to distill $\mathcal{H}^s$ and $\mathcal{R}^s$
      $\boldsymbol{x}_{stego}^s, \{\boldsymbol{x}_{stego}^{t_i}\}_{i=1}^3 \leftarrow \mathcal{H}^s(\boldsymbol{x}_{cover}, \boldsymbol{x}_{secret}, \mathcal{H}^t)$;
      $\boldsymbol{x}_{mia}^s, \{\boldsymbol{x}_{mia}^{t_i}\}_{i=1}^3 \leftarrow \mathcal{M}(\boldsymbol{x}_{stego}^s, \{\boldsymbol{x}_{stego}^{t_i}\}_{i=1}^3)$;
      $\hat{\boldsymbol{y}}^s, \{\hat{\boldsymbol{y}}^{t_i}\}_{i=1}^3 \leftarrow \mathcal{R}^s(\boldsymbol{x}_{mia}^s, \{\boldsymbol{x}_{stego}^{t_i}\}_{i=1}^3, \mathcal{R}^t)$;
      $\mathcal{L}_{dist}^{hide} \leftarrow \ell_2(\boldsymbol{x}_{stego}^s, \boldsymbol{x}_{stego}^t) + \sum_{i=1}^3 \ell_2(\boldsymbol{x}_{stego}^{t_i}, \boldsymbol{x}_{stego}^t)$
      $\mathcal{L}_{dist}^{mia} \leftarrow \ell_2(\boldsymbol{x}_{mia}^s, \boldsymbol{x}_{mia}^t) + \sum_{i=1}^3 \ell_2(\boldsymbol{x}_{mia}^{t_i}, \boldsymbol{x}_{mia}^t)$
      $\mathcal{L}_{dist}^{recover} \leftarrow \mathcal{L}_{mia}(\hat{\boldsymbol{y}}^s, \boldsymbol{y}) + \sum_{i=1}^3 \mathcal{L}_{mia}(\hat{\boldsymbol{y}}^{t_i}, \boldsymbol{y})$
      $\theta_{\mathcal{H}^s} \leftarrow \theta_{\mathcal{H}^s} - \ell_r^2 \cdot \nabla_{\theta_{\mathcal{H}^s}} (\beta_1 \cdot \mathcal{L}_{dist}^{hide} + \beta_2 \cdot \mathcal{L}_{dist}^{mia})$
      $\theta_{\mathcal{R}^s} \leftarrow \theta_{\mathcal{R}^s} - \ell_r^2 \cdot \nabla_{\theta_{\mathcal{R}^s}} (\beta_3 \cdot \mathcal{L}_{dist}^{recover})$
      if $e_2 \le j \le e_3$ **then**
        # Stage 3: optimizing $\mathcal{M}$, $\mathcal{H}^s$, and $\mathcal{R}^s$
        $\theta_{\mathcal{H}^s} \leftarrow \theta_{\mathcal{H}^s} - \ell_r^3 \cdot \nabla_{\theta_{\mathcal{H}^s}} (\beta_1 \cdot \mathcal{L}_{dist}^{hide} + \beta_2 \cdot \mathcal{L}_{dist}^{mia})$
        $\theta_{\mathcal{M}} \leftarrow \theta_{\mathcal{M}} - \ell_r^3 \cdot \nabla_{\theta_{\mathcal{M}}} (\beta_2 \cdot \mathcal{L}_{dist}^{mia} + \beta_3 \cdot \mathcal{L}_{dist}^{recover})$
        $\theta_{\mathcal{R}^s} \leftarrow \theta_{\mathcal{R}^s} - \ell_r^3 \cdot \nabla_{\theta_{\mathcal{R}^s}} (\beta_3 \cdot \mathcal{L}_{dist}^{recover})$
      end
    end
  end
end

phase, to maximize the MIA performance while distilling, we directly use the MIA label $\boldsymbol{y}$ for supervision. This approach aims to ensure that the student network not only mimics the teacher's functionality closely but also enhances its capability for medical image analysis through direct guidance from the true labels.

$$\begin{aligned}
\mathcal{L}_{dist}^{hide} &= \ell_2(\boldsymbol{x}_{stego}^s, \boldsymbol{x}_{stego}^t) + \textstyle\sum_{i=1}^3 \ell_2(\boldsymbol{x}_{stego}^{t_i}, \boldsymbol{x}_{stego}^t),\\
\mathcal{L}_{dist}^{mia} &= \ell_2(\boldsymbol{x}_{mia}^s, \boldsymbol{x}_{mia}^t) + \textstyle\sum_{i=1}^3 \ell_2(\boldsymbol{x}_{mia}^{t_i}, \boldsymbol{x}_{mia}^t),\\
\mathcal{L}_{dist}^{recover} &= \mathcal{L}_{mia}(\hat{\boldsymbol{y}}^s, \boldsymbol{y}) + \textstyle\sum_{i=1}^3 \mathcal{L}_{mia}(\hat{\boldsymbol{y}}^{t_i}, \boldsymbol{y}),\\
\mathcal{L}_{dist} &= \alpha_1 \cdot \mathcal{L}_{dist}^{hide} + \alpha_2 \cdot \mathcal{L}_{dist}^{mia} + \alpha_3 \cdot \mathcal{L}_{dist}^{recover},
\end{aligned} \tag{7}$$

where $\alpha_1, \alpha_2, \alpha_3$ are used to trade-off among these loss functions.

3.4 Training Process

We use a multi-stage training strategy to train HideMIA (see Algorithm 1). In the first stage, we freeze the pretrained DIH network and train the MIA network with $\mathcal{L}_{total}$. In the second stage, we freeze the pretrained DIH network and MIA network to prune the LightIH with $\mathcal{L}_{dist}$. Finally, we use $\mathcal{L}_{total}$ to supervise the fine-tuning of the whole HideMIA, *i.e.,* MIA network and LightIH. Note that $\mathcal{L}_{dice}$ is not included for the classification task.

$$\begin{aligned}
\mathcal{L}_{mia} = \mathcal{L}_{ce}(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \mathcal{L}_{dice}(\hat{\boldsymbol{y}}, \boldsymbol{y}),\ \mathcal{L}_{percept} = \ell_2(\boldsymbol{x}_{mia}, \boldsymbol{x}_{cover}),\\
\mathcal{L}_{total} = \beta_1 \cdot \mathcal{L}_{mia} + \beta_2 \cdot \mathcal{L}_{percept},
\end{aligned} \tag{8}$$

where β_1, β_2 are used to trade-off between $\mathcal{L}_{mia}$ and $\mathcal{L}_{percept}$.

4 Experiments and Results

4.1 Experimental Setup

Datasets. For comprehensive comparisons, we use six widely used MIA datasets, including three segmentation datasets and three classification datasets. It should be noted that all selected datasets are acquired with different imaging devices and capture distinct subjects. These datasets are BUSI (breast ultrasound tumor segmentation with 612 images) [1], Kvasir-SEG (endoscopic polyp segmentation with 1,000 images) [16], ChildDental (tooth X-ray image segmentation with 2,489 images) [5, 15], SIPaKMed [42] (pathological cervical cell classification with 1,510 images), DermaMNIST [54] (dermatoscopic skin lesion classification with 10,015 images), and ChestCT [41] (chest computed tomography image classification with 1,000 images). For all experiments in this work, we follow the official splitting configuration of the datasets. For those without an official configuration (BUSI, Kvasir-SEG, and SIPaKMed), we randomly divide the data into training and testing sets at a ratio of 8:2. We randomly sample 512 images of MS COCO [28] as the cover image dataset. There is no overlap between the training cover images and testing cover images.

Competing Methods. To the best of our knowledge, no method is proposed to perform MIA in the steganographic domain. To ensure a comprehensive and fair comparison, we select three widely used DIH methods (*i.e.,* DeepStega [4], HiDDeN [65], and HiNet [17]) and four famous or SOTA medical image segmentation networks, *i.e.,* U-Net [44], TransUNet [7], XNet [61], and CMUNeXt [47]. Then we pair each DIH method with each segmentation network, forming twelve competing methods. For classification tasks, we also employ a pair of "segmentation network + DIH network" for comparison. This is due to the architecture of our HideMIA, where the MIA network is designed with an encoder-decoder structure. Employing segmentation networks, which also use this encoder-decoder structure, rather than encoder-only structured classification networks, allows fairer comparisons.

Evaluation Metrics. Dice Similarity Coefficient (DSC) and Average Surface Distance (ASD) are adopted for medical image segmentation tasks [43]. The medical image classification performance is evaluated by Accuracy (Acc) and Area Under receiver operating characteristic Curve (AUC). We calculate the average Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) between $\boldsymbol{x}_{mia}$ and $\boldsymbol{x}_{cover}$ to assess the imperceptibility. We randomly sample five cover images to calculate the mean values of these metrics.

Implementation Details. The size of the images is standardized to 224×224. We use the Adam optimizer with learning rates $\ell_r^1, \ell_r^2, \ell_r^3$ of 5×10^{-4}, 1×10^{-4}, 1×10^{-5}, respectively. The batch size is 8. During training, all cover images $\boldsymbol{x}_{stego}$ are randomly selected. In the loss functions, *i.e.,* Eqs. (7)-(8), the weights $\alpha_1, \alpha_2, \alpha_3, \beta_1$, and β_2 are respectively set to 800, 800, 4, 1, and 200. We set λ of MoDC shown in Eq. (3) to 0.5. For the secret images, $\boldsymbol{x}_{secret}$, hiding is performed across three channels, and recovery for $\boldsymbol{x}_{res}$ is also conducted across three channels. During segmentation, the recovered $\boldsymbol{x}_{res}$ has its three channels averaged followed by a sigmoid activation to obtain the final binary map. For classification tasks, a classification head is appended after the recovered results to obtain the predictions.

Table 1: Comparisons of covert medical image segmentation and classification performance across six datasets. DSC (%) and ASD are reported for segmentation datasets, while Acc (%) and AUC (%) are reported for classification datasets.

Task →		Segmentation						Classification					
Dataset →		BUSI		Kvasir-SEG		ChildDental		SIPaKMeD		DermaMNIST		ChestCT	
MIA ↓	Hiding ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	DSC ↑	ASD ↓	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑
U-Net	-	72.85	6.58	81.83	4.30	91.15	0.15	-	-	-	-	-	-
ResNet50	-	-	-	-	-	-	-	67.23	84.35	73.10	91.20	58.41	85.37
U-Net	HiNet	60.55	13.76	64.82	9.47	77.62	1.18	44.63	73.74	67.08	82.72	31.75	69.60
TransUNet	HiNet	63.92	10.61	60.29	11.5	74.35	1.42	49.72	81.07	68.33	77.10	38.10	55.04
XNet	HiNet	59.74	13.43	61.60	10.88	74.66	1.49	44.41	75.07	69.23	86.46	31.43	69.43
CMUNeXt	HiNet	64.36	11.04	68.26	9.15	75.23	1.78	47.46	79.93	67.48	86.47	33.65	65.19
U-Net	DeepStega	66.77	9.45	60.38	8.42	69.84	2.11	43.05	79.03	66.18	77.97	33.33	54.44
TransUNet	DeepStega	62.16	11.95	58.93	13.11	71.54	1.43	47.91	83.92	65.59	85.50	34.29	64.93
XNet	DeepStega	65.44	9.65	63.71	11.53	71.12	1.57	56.72	85.99	66.38	82.70	35.24	66.05
CMUNeXt	DeepStega	61.41	12.93	65.23	10.22	76.08	1.80	52.54	76.57	67.23	85.40	33.02	69.92
U-Net	HiDDeN	61.79	10.38	65.92	8.67	74.90	2.65	38.98	75.96	67.13	80.33	30.16	65.32
TransUNet	HiDDeN	61.68	10.54	61.82	12.83	77.39	1.73	49.94	82.65	68.28	78.97	29.84	66.08
XNet	HiDDeN	63.45	12.09	63.83	11.88	70.07	1.49	44.18	74.31	66.88	72.94	31.11	65.41
CMUNeXt	HiDDeN	63.29	12.95	64.51	11.98	78.59	1.1	42.37	73.79	67.68	81.88	37.14	54.24
HideMIA (Ours)		**74.11**	**6.76**	**77.56**	**5.79**	**89.45**	**0.19**	**63.28**	**87.25**	**78.10**	**92.73**	**51.75**	**72.72**

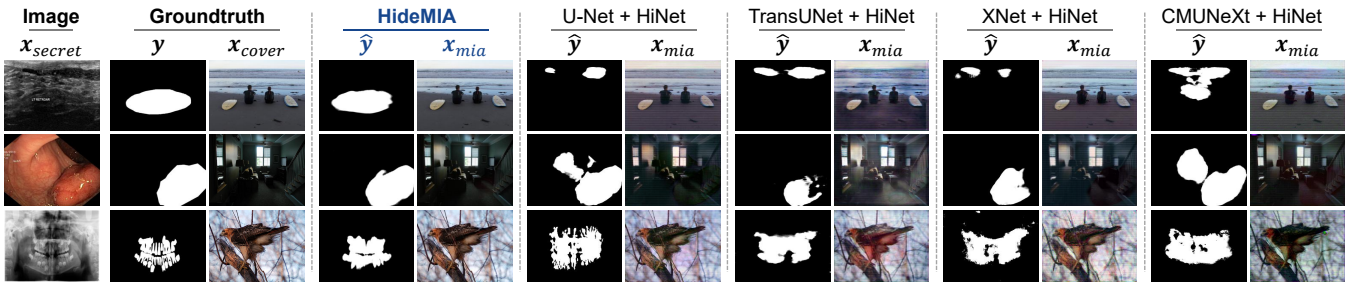

Figure 4: Visual comparisons for covert medical image segmentation on BUSI, Kvasir-SEG, and ChildDental.

Table 2: Comparisons of imperceptibility of the covert MIA. We report average PSNR and SSIM between x_{mia} and x_{cover} on all segmentation datasets and classification datasets.

Method		Segmentation		Classification	
MIA	Hiding	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
U-Net	HiNet	18.58	0.2953	19.29	0.3460
TransUNet	HiNet	17.35	0.3251	20.30	0.3694
XNet	HiNet	18.69	0.2904	17.42	0.2473
CMUNeXt	HiNet	18.72	0.2916	19.77	0.3850
U-Net	DeepStega	18.07	0.3150	21.49	0.3759
TransUNet	DeepStega	17.38	0.3653	15.85	0.2721
XNet	DeepStega	18.11	0.4131	19.57	0.3561
CMUNeXt	DeepStega	18.07	0.3678	20.04	0.3697
U-Net	HiDDeN	17.08	0.2865	17.03	0.2980
TransUNet	HiDDeN	16.83	0.2217	16.43	0.2759
XNet	HiDDeN	16.60	0.2387	16.30	0.2706
CMUNeXt	HiDDeN	18.13	0.2746	13.48	0.4532
HideMIA (Ours)		**33.01**	**0.8556**	**36.84**	**0.8963**

4.2 Comparison Results

Segmentation and Classification Results. As shown in Table 1 and Fig. 4, HideMIA achieves the highest segmentation and classification performance across three datasets compared to other covert MIA methods, respectively. On average, its segmentation performance surpasses that of the second-place *CMUNeXt + HiNet* by 11.09% in DSC and is lower by 3.07 in ASD. When compared to the vanilla framework (without privacy-enhancing) utilizing U-Net, HideMIA's average performance only shows a slight decrease of 1.57% in DSC, while ASD increases by 0.57. Meanwhile, its classification performance surpasses that of the second-place *XNet + DeepStega* by 11.60% in Acc and 5.98% in AUC. When compared to the vanilla framework (without privacy-enhancing) utilizing U-Net, HideMIA's average performance only shows a slight decrease of 1.87% in Acc and 2.74% in AUC. This indicates HideMIA's effectiveness in maintaining high MIA performance (both for segmentation and classification) while enhancing privacy protection. We note that XNet is also based on wavelet transforms, similar to HideMIA, underperforms due to its lack of adaptive analysis across different bands, unlike HideMIA. Additionally, despite TransUNet, XNet, and CMUNeXt having incremental designs over U-Net, their performance does not improve across all datasets. This inconsistency can be attributed to the fact that their incremental designs are not effective in the steganographic domain.

Imperceptiblity. Results in Table 2 demonstrate that our HideMIA outperforms competing methods in PSNR and SSIM for both classification and segmentation tasks, with a significant improvement over other methods. As shown in Fig. 4, x_{mia} generated by HideMIA has the highest visual similarity to x_{cover}. Compared to x_{cover}, there are significant color modifications and obvious horizontal striping artifacts in x_{mia} generated by other methods. The poor imperceptibility of competing methods' x_{mia} is attributed to the absence of low-frequency guidance from x_{stego}. Results in Table 3 demonstrate that removing x_{stego}^{LL} in Eq. (1) significantly decreases PSNR and SSIM, yet the MIA performance remains almost unaffected. The imperceptibility of x_{stego} will be discussed in Sec. 4.3.

Table 3: Ablation results on the LL-band residual (x^{LL}_{stego}) in Eq. (1) over BUSI.

Residual of x^{LL}_{stego} →	w/	w/o
PSNR ↑	**33.84**	17.41
SSIM (%) ↑	**83.94**	26.15
DSC (%) ↑	74.11	**74.39**
ASD ↓	**6.76**	6.89

Table 4: Ablation results on different numbers of activated difference convolutions within MoDC over BUSI.

Activated Number k	DSC (%) ↑	ASD ↓
0 (Vanilla Conv.)	70.77	9.91
1	72.75	9.12
2	**74.11**	**6.76**
3	73.47	7.01
4	71.88	7.34
5	71.37	7.36

Table 5: Ablation results on utilizing different difference convolutions within MoDC over BUSI. We fix the activated number k to 2.

Candidate Set of DCs	DSC (%) ↑	ASD ↓
$\{\mathcal{D}_1, \mathcal{D}_2\}$	70.47	8.91
$\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3\}$	71.51	8.73
$\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4\}$	73.80	6.89
$\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5\}$	**74.11**	**6.76**

Table 6: Ablation results on AsyWA. We compare different variations of AsyWA over BUSI.

Variation of AsyWA	DSC (%) ↑	ASD ↓
w/o attention	68.94	10.14
Inner-band	69.91	8.22
Inter-band (Symmetry)	71.35	7.39
Inter-band (Asymmetric)	73.38	6.75
Inner- & Inter-band (Symmetry)	72.32	9.36
Inner- & Inter-band (Asymmetric)	74.11	6.76

Table 7: Ablation results on LightIH. We calculate PSNR and SSIM between x_{stego} and x_{cover} to evaluate hiding performance over BUSI.

DIH Network	FLOPs ↓	Parameters ↓	DSC (%) ↑	PSNR ↑	SSIM ↑
HiNet	152.17G	4.05M	71.45	**45.27**	**0.989**
DeepStega	<u>73.30G</u>	**0.49M**	70.25	31.76	0.812
HiDDeN	87.49G	<u>0.58M</u>	69.80	34.16	0.969
LightIH (Feature-Aligned)	**38.04G**	1.01M	72.06	41.83	0.975
LightIH (Function-Aligned)	**38.04G**	1.01M	**74.11**	<u>43.60</u>	<u>0.982</u>

4.3 Ablation Study

Effectiveness of MoDC. Results in Table 4 reveal that when the activation number is 0 (using vanilla convolution), there is a significant performance decline compared to others. This verifies the effectiveness of the designed difference convolutions in the steganographic domain. Moreover, we observe that the performance peaks when the activation number is 2, indicating that too few activated DCs fail to form convolution operations tailored to the steganographic analysis suitable for the respective band, leading to inadequate MIA feature extraction. Conversely, too many activations introduce inappropriate kernels, introducing noise into the MIA process. Besides, we fix the number of activations to 2, which shows the best MIA performance, and compare the performance under various sets of candidate DCs in Table 5. When the candidate set includes the full set, *i.e.* $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5\}$, the performance is optimal. A candidate DC count of 4 outperforms a count of 3, and similarly, a count of 3 outperforms a count of 2. These results validate the effectiveness of each of the novel DCs we propose.

Effectiveness of AysWA. As shown in Table 6, performance is at its lowest without any attention mechanism, indicating that global perception across spatial and spectral dimensions is crucial for MIA in the steganographic domain. Besides, integrating only inner-band or inter-band attention is less effective than having both, further validating the effectiveness of the proposed cross-spatial and cross-spectral interactions. Meanwhile, we observe that inter-band (Asymmetric) outperforms Inter-band (Symmetric) and Inner- & Inter-band (Asymmetric) outperforms Inner- & Inter-band (Symmetric). This demonstrates that our asymmetric design for the LL band is effective for the HideMIA framework. It can help resist the noisy cover-specific information in the LL band.

Effectiveness of LightIH. We compare LightIH, HiNet, DeepStega, and HiDDeN with resource metrics, *i.e.,* Floating Point Operations (FLOPs) and Parameters, image hiding performance (PSNR and SSIM between $\boldsymbol{x}_{stego}$ and $\boldsymbol{x}_{cover}$), and MIA performance (DSC). As shown in Table 7, since having a quarter of the INN layers compared to HiNet, LightIH exhibits lower FLOPs than other DIH networks. LightIH has fewer parameters than its teacher (HiNet), slightly more than DeepStega and HiDDeN. However, due to our function-aligned distillation strategy, LightIH significantly outperforms DeepStega and HiDDeN in image hiding performance, with a substantial advantage in MIA performance as well. In real-world scenarios, a PSNR over 40 is already very difficult for human eyes to distinguish. Therefore, trading off a slight decrease in imperceptibility for reduced resource consumption and improved MIA performance compared to the teacher network, HiNet, is worthwhile. Besides, we also notice that aligning features directly is not as effective as aligning the DIH and MIA functions in a multi-stage framework like HideMIA.

5 Conclusion

We propose a covert client-server MIA framework that effectively defends against attacks from both the transmission channel and malicious third-party servers. HideMIA is less likely to arouse suspicion among attackers compared to existing privacy-enhancing MIA methods. Comprehensive experiments demonstrate that within HideMIA, our proposed MoDC and AsyWA enable more effective covert analysis directly in the steganographic domain. Additionally, our LightIH obtained by function-aligned knowledge distillation facilitates the deployment of a lightweight DIH network on the client side, making HideMIA more practical. HideMIA achieves the SOTA MIA performance and imperceptibility in medical image segmentation and image classification. It also brings insight into fine-grained analysis in the steganographic domain.

However, as this is the first work to protect the client-server MIA framework through image hiding, there are some limitations that we hope to address in the future: (1) *Better imperceptibility.* As illustrated in Tables 2 and 7, the imperceptibility of $\boldsymbol{x}_{mia}$ is lower than that of $\boldsymbol{x}_{stego}$. After processing through the MIA network, there are usually some stripe-like artifacts (see Fig. 4). Eliminating these artifacts would make attackers more difficult to detect. (2) *Lossless MIA.* Compared to non-privacy-enhancing frameworks, there is still a gap in MIA performance, indicating that the distortion introduced by DIH brings side effects. Designing a nearly lossless covert MIA framework is worth exploring.

Acknowledgement

This research was done at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, and supported in part by the National Natural Science Foundation of China (Grants No. 62272022, 62306061, 62276254, and U23B2054), the National Key Research and Development Program of China (Grants No. 2022YFB3207700 and 210YBXM2024106007), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037), the InnoHK program, and the NTU-PKU Joint Research Institute (sponsored by the Ng Teng Fong Charitable Foundation).

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data in Brief* (2020).

[2] T. Avudaiappan, R. Balasubramanian, S. Pandiyan, S. Murugan, S. Lakshmanaprabu, and K. Shankar. 2018. Medical Image Security Using Dual Encryption with Oppositional Based Optimization Algorithm. *Journal of Medical Systems* 42 (2018), 1–11.

[3] R. Bagai, Nafia Malik, and Murtuza Jadliwala. 2017. Measuring Anonymity of Pseudonymized Data After Probabilistic Background Attacks. *IEEE Transactions on Information Forensics and Security* 12 (2017), 1156–1169.

[4] Shumeet Baluja. 2017. Hiding images in plain sight: Deep steganography. 30 (2017).

[5] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2014. Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration. *IEEE Transactions on Medical Imaging* (2014).

[6] Chun-Fu (Richard) Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In *Proceedings of the IEEE International Conference on Computer Vision*. 347–356.

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* abs/2102.04306 (2021).

[8] Xin Deng, Chao Gao, and Mai Xu. 2023. PIRNet: Privacy-Preserving Image Restoration Network via Wavelet Lifting. In *Proceedings of the IEEE International Conference on Computer Vision*. 22368–22377.

[9] Alex Gaudio, Asim Smailagic, Christos Faloutsos, Shreshta Mohan, Elvin Johnson, Yuhao Liu, Pedro Costa, and Aurélio Campilho. 2023. *DeepFixCX*: Explainable privacy-preserving image compression for medical image analysis. *WIREs Data Mining and Knowledge Discovery* 13, 4 (2023).

[10] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[11] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. 2024. Federated learning for medical image analysis: A survey. *Pattern Recognition* (2024), 110424.

[12] Zhenyu Guan, Junpeng Jing, Xin Deng, Mai Xu, Lai Jiang, Zhou Zhang, and Yipeng Li. 2022. DeepMIH: Deep invertible network for multiple image hiding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 372–390.

[13] Jamie Hayes and George Danezis. 2017. Generating steganographic images via adversarial training. In *Proceedings of the Neural Information Processing Systems*. 1954–1963.

[14] Zhongwei He, Wei Lu, Wei Sun, and Jiwu Huang. 2012. Digital image splicing detection based on Markov features in DCT and DWT domain. *Pattern Recognition* 45, 12 (2012), 4292–4299.

[15] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les R. Folio, Jenifer Siegelman, Fiona M. Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer K. Antani, George R. Thoma, Yi-Xiang J. Wang, Pu-Xuan Lu, and Clement J. McDonald. 2014. Automatic Tuberculosis Screening Using Chest Radiographs. *IEEE Transactions on Medical Imaging* (2014).

[16] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. 2020. Kvasir-SEG: A Segmented Polyp Dataset. In *MMM*.

[17] Junpeng Jing, Xin Deng, Mai Xu, Jianyi Wang, and Zhenyu Guan. 2021. Hinet: Deep image hiding by invertible network. In *Proceedings of the IEEE International Conference on Computer Vision*. 4733–4742.

[18] Georgios Kaissis, M. Makowski, D. Rückert, and R. Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2 (2020), 305–311.

[19] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima Jr, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, et al. 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* 3, 6 (2021), 473–484.

[20] Bach Ngoc Kim, Jose Dolz, Christian Desrosiers, and Pierre-Marc Jodoin. 2021. Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy. In *Proceedings of the British Machine Vision Conference*. 375.

[21] Bach Ngoc Kim, Jose Dolz, Pierre-Marc Jodoin, and Christian Desrosiers. 2021. Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images. *IEEE Transactions on Medical Imaging* 40, 7 (2021), 1737–1749.

[22] Bach Ngoc Kim, Jose Dolz, Pierre-Marc Jodoin, and Christian Desrosiers. 2023. Mixup-Privacy: A Simple yet Effective Approach for Privacy-Preserving Segmentation. In *Proceedings of Information Processing In Medical Imaging*, Vol. 13939. 717–729.

[23] Chenqi Kong, Anwei Luo, Shiqi Wang, Haoliang Li, Anderson Rocha, and Alex C. Kot. 2023. Pixel-Inconsistency Modeling for Image Manipulation Localization. *arXiv* abs/2310.00234 (2023).

[24] Chenqi Kong, Anwei Luo, Song Xia, Yi Yu, Haoliang Li, and Alex C. Kot. 2024. MoE-FFD: Mixture of Experts for Generalized and Parameter-Efficient Face Forgery Detection. *arXiv* abs/2404.08452 (2024).

[25] Akshay Kumar, Rajneesh Rani, and Samayveer Singh. 2023. A survey of recent advances in image steganography. *Secur. Priv.* 6, 3 (2023).

[26] Guobiao Li, Sheng Li, Zicong Luo, Zhenxing Qian, and Xinpeng Zhang. 2024. Purified and Unified Steganographic Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[27] Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen. 2023. When Object Detection Meets Knowledge Distillation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 10555–10579.

[28] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*, Vol. 8693. 740–755.

[29] Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Zitong Yu, Wenzhong Tang, and Alex Kot. 2024. Suppress and Rebalance: Towards Generalized Multi-Modal Face Anti-Spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 211–221.

[30] Xun Lin, Yi Yu, Song Xia, Jue Jiang, Haoran Wang, Zitong Yu, Yizhong Liu, Ying Fu, Shuai Wang, Wenzhong Tang, and Alex Kot. 2024. Safeguarding Medical Image Segmentation Datasets against Unauthorized Training via Contour- and Texture-Aware Perturbations. *arXiv* 2403.14250 (2024).

[31] G. Litjens, Thijs Kooi, B. Bejnordi, A. Setio, F. Ciompi, Mohsen Ghafoorian, J. Laak, B. Ginneken, and C. I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42 (2017), 60–88.

[32] Ajian Liu. 2024. CA-MoEiT: Generalizable Face Anti-spoofing via Dual Cross-Attention and Semi-fixed Mixture-of-Expert. *International Journal of Computer Vision* (2024), 1–14.

[33] Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. 2023. Function-Consistent Feature Distillation. In *Proceedings of the International Conference on Learning Representations*.

[34] Dongyang Liu, Meina Kan, Shiguang Shan, and Xilin Chen. 2023. Function-Consistent Feature Distillation. In *Proceedings of the International Conference on Learning Representations*.

[35] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10770–10780.

[36] Lianshan Liu, Lingzhuang Meng, Yanjun Peng, and Xiaoli Wang. 2021. A data hiding scheme based on U-Net and wavelet transform. *Knowledge-Based Systems* 223 (2021), 107022.

[37] Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L. Rosin. 2021. Large-Capacity Image Steganography Based on Invertible Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10816–10825.

[38] Entao Luo, Md. Zakirul Alam Bhuiyan, Guojun Wang, M. A. Rahman, Jie Wu, and Mohammed Atiquzzaman. 2018. PrivacyProtector: Privacy-Protected Patient Data Collection in IoT-Based Healthcare Systems. *IEEE Communications Magazine* 56 (2018), 163–168.

[39] Ruohan Meng, Chenyu Yi, Yi Yu, Siyuan Yang, Bingquan Shen, and Alex C. Kot. 2024. Semantic Deep Hiding for Robust Unlearnable Examples. *IEEE Transactions on Information Forensics and Security* (2024).

[40] Kai Packhäuser, Sebastian Gündel, Florian Thamm, Felix Denzinger, and Andreas K. Maier. 2023. Deep Learning-Based Anonymization of Chest Radiographs: A Utility-Preserving Measure for Patient Privacy. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*, Vol. 14222. 262–272.

[41] Jasmine pemeena priyadarsini, Ketan Kotecha, G. Rajini, K. Hariharan, K. Raj, K. Ram, V. Indragandhi, V. Subramaniyaswamy, and Sharnil Pandya. 2023. Lung Diseases Detection Using Various Deep Learning Algorithms. *Journal of Healthcare Engineering* 2023 (2023), 1–13.

[42] Marina E. Plissiti, Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, O. Krikoni, and Antonia Charchanti. 2018. Sipakmed: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images. In *Proceedings of the IEEE International Conference on Image Processing*. 3144–3148.

[43] Xiaolei Qu, Jiale Zhou, Jue Jiang, Wenhan Wang, Haoran Wang, Shuai Wang, Wenzhong Tang, and Xun Lin. 2024. EH-former: Regional easy-hard-aware transformer for breast lesion segmentation in ultrasound images. *Information Fusion* 109 (2024), 102430.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention*.

[45] Isaac Shiri, Behrooz Razeghi, Sohrab Ferdowsi, Yazdan Salimi, Deniz Gündüz, Douglas Teodoro, Slava Voloshynovskiy, and Habib Zaidi. 2024. PRIMIS: Privacy-preserving medical image sharing via deep sparsifying transform learning with obfuscation. *Journal of Biomedical Informatics* 150 (2024), 104583.

[46] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. 2021. Pixel Difference Networks for Efficient Edge Detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

[47] Fenghe Tang, Jianrui Ding, Lingtao Wang, Chunping Ning, and S. Kevin Zhou. 2024. CMUNeXt: An Efficient Medical Image Segmentation Network based on Large Kernel and Skip Fusion. In *Proceedings of the IEEE International Symposium on Biomedical Imaging*.

[48] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. MLP-Mixer: An all-MLP Architecture for Vision. In *Proceedings of the Neural Information Processing Systems*. 24261–24272.

[49] Chong Wang, Yi Yu, Lanqing Guo, and Bihan Wen. 2024. Benchmarking Adversarial Robustness of Image Shadow Removal with Shadow-Adaptive Attacks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. 13126–13130.

[50] Song Xia, Yi Yu, Xudong Jiang, and Henghui Ding. 2024. Mitigating the Curse of Dimensionality for Certified Robustness via Dual Randomized Smoothing. In *Proceedings of the International Conference on Learning Representations*.

[51] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. 2020. Invertible image rescaling. In *Proceedings of the European Conference on Computer Vision*. 126–144.

[52] Yazhou Xing, Zian Qian, and Qifeng Chen. 2021. Invertible Image Signal Processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6287–6296.

[53] Youmin Xu, Chong Mou, Yujie Hu, Jingfen Xie, and Jian Zhang. 2022. Robust Invertible Image Steganography. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7865–7874.

[54] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data* 10, 1 (2023), 41.

[55] Ziyuan Yang, Yingyu Chen, Huijie Huangfu, Maosong Ran, Hui Wang, Xiaoxiao Li, and Yi Zhang. 2023. Dynamic Corrected Split Federated Learning With Homomorphic Encryption for U-Shaped Medical Image Networks. *IEEE Journal of Biomedical and Health Informatics* 27, 12 (2023), 5946–5957.

[56] Yi Yu, Yufei Wang, Song Xia, Wenhan Yang, Shijian Lu, Yap-Peng Tan, and Alex C Kot. 2024. Purify Unlearnable Examples via Rate-Constrained Variational Autoencoders. In *Proceedings of the International Conference on Machine Learning (Proceedings of Machine Learning Research)*.

[57] Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-Peng Tan, and Alex C. Kot. 2023. Backdoor Attacks Against Deep Image Compression via Adaptive Frequency Trigger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12250–12259.

[58] Yi Yu, Wenhan Yang, Yap-Peng Tan, and Alex C. Kot. 2022. Towards Robust Rain Removal Against Adversarial Attacks: A Comprehensive Benchmark Analysis and Beyond. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6013–6022.

[59] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. 2020. Searching Central Difference Convolutional Networks for Face Anti-Spoofing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[60] Peng Zhang, Li Su, Liang Li, Bing-Kun Bao, Pamela C. Cosman, Guorong Li, and Qingming Huang. 2019. Training Efficient Saliency Prediction Models with Knowledge Distillation. In *ACM MM*. 512–520.

[61] Yanfeng Zhou, Jiaxing Huang, Chenlong Wang, Le Song, and Ge Yang. 2023. XNet: Wavelet-Based Low and High Frequency Fusion Networks for Fully- and Semi-Supervised Semantic Segmentation of Biomedical Images. In *Proceedings of the IEEE International Conference on Computer Vision*. 21028–21039.

[62] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. 2022. Mixture-of-Experts with Expert Choice Routing. In *Proceedings of the Neural Information Processing Systems*.

[63] Dan Zhu, Hui Zhu, Cheng Huang, Rongxing Lu, Dengguo Feng, and Xuemin Shen. 2024. Efficient and Accurate Cloud-Assisted Medical Pre-Diagnosis With Privacy Preservation. *IEEE Transactions on Dependable and Secure Computing* 21, 2 (2024), 860–875.

[64] Enjun Zhu, Haiyu Feng, Long Chen, Yongqiang Lai, and Senchun Chai. 2024. MP-Net: A Multi-Center Privacy-Preserving Network for Medical Image Segmentation. *IEEE Transactions on Medical Imaging* (2024).

[65] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. HiDDeN: Hiding Data With Deep Networks. In *Proceedings of the European Conference on Computer Vision*, Vol. 11219. 682–697.