Neural Computing and Applications (2023) 35:10909–10927
https://doi.org/10.1007/s00521-023-08274-w

ORIGINAL ARTICLE

# Deep adaptive hiding network for image hiding using attentive frequency extraction and gradual depth extraction

Le Zhang[1] · Yao Lu[1] · Jinxing Li[1] · Fanglin Chen[1] · Guangming Lu[1,3] · David Zhang[2]

Received: 22 June 2022 / Accepted: 6 January 2023 / Published online: 28 January 2023

**Abstract**
Image hiding secures information security in multimedia communication. Existing deep image hiding methods usually process the secret and cover information at first, and then fuse such entire processed information. This complete and rough fusion pipeline, however, severely hinders the quality improvement of the stego and revealed secret images. This paper proposes a deep image hiding architecture, named Deep Adaptive Hiding Network (DAH-Net), to gradually extract and fuse the necessary secret and cover information at the frequency and the depth (layer) extents. Specifically, we propose the Attentive Frequency Extraction method for the DAH-Net to adaptively extract the necessary secret and cover information at the frequency level. The Gradual Depth Extraction method is further proposed for the DAH-Net to gradually extract and fuse the attentive frequency secret and cover information at the depth (layer) level of the deep image hiding network. Extensive experiment results demonstrate the proposed DAH-Net is more universal and achieves state-of-the-art performances in image hiding, watermarking, and photographic steganography.

**Keywords** Image hiding · Deep adaptive hiding network · Attentive frequency extraction · Gradual depth extraction

## 1 Introduction

Image hiding (or image steganography) achieves covert communication between the sender and receiver without being detected by third parties [1]. The sender hides secret images in cover images imperceptibly and transfers the generated stego images to the receiver for revealing secret images. For the sake of invisibility and security, the stego images should have high image-quality and high similarity with the cover images. Meanwhile, the quality of the revealed secret images directly affects whether the secret information can be transmitted to the receiver accurately. Similarly, The process of generating watermarking also encodes secret information within image; however, watermarking is usually utilized to prove image ownership as a form of copyright protection [2]. The difference between watermarking and steganography is that image hiding pays more attention to the invisibility and security of communication to ensure the effective implementation of covert communication, and watermarking is more focused on robustness to prove the image ownership in the copyright protection field.

✉ Guangming Lu
luguangm@hit.edu.cn

Le Zhang
zhangle408@gmail.com

Yao Lu
luyao2021@hit.edu.cn

Jinxing Li
lijinxing158@hit.edu.cn

Fanglin Chen
chenfanglin@hit.edu.cn

David Zhang
davidzhang@cuhk.edu.cn

[1] Department of Computer Science, Harbin Institute of Technology Shenzhen, Xiaqing Street, Shenzhen 518071, Guangdong, China

[2] School of Data Science, Chinese University of Hong Kong Shenzhen, Longxiang Street, Shenzhen 518172, Guangdong, China

[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, Shenzhen, China

Traditional information hiding methods [3–6] successfully hide a small amount of binary information into cover images and generate high quality stego images. The steganography capacity of such methods, however, is limited, hardly meeting the requirements of a large capacity for image hiding [7]. Convolution Neural Network (CNN) has shown excellent performances on various computer vision tasks, such as image recognition [8–11], object detection [12, 13], image denoising [14–16], and semantic segmentation [17–19]. Due to the powerful representations of CNN, Deep-Stego [20, 21] is the first successful implementation of hiding a color image in another color image using CNN. Deep-Stego first uses a network to preprocess secret images and then encodes the preprocessed secret images and cover images using a hiding network to generate the stego images. Then, the revealed secret images are generated using the revealing network from the stego images. Different from the traditional information hiding methods [22–24], the Deep-Stego network can effectively encode the secret images and predict the hiding positions of the secret information within the cover information. The existing deep hiding pipeline including Deep-Stego falls into one meta-architecture category termed Dependent Deep Hiding (DDH) in [7]. The meta-architecture is seen in Fig. 1a. DDH mainly includes the hiding and the revealing networks. The hiding process can be formulated as the following equation:

$$C' = 1 \times f(S \oplus C), \tag{1}$$

where $f(\cdot)$ represents the transformation of the hiding network, and $\oplus$ indicates concatenation operation at the channel axis. $S$, $C$, and $C'$ denote the secret, cover, and stego images, respectively. The Universal Deep Hiding (UDH) method, a more general deep hiding framework, is proposed for image hiding, watermarking, and photographic steganography in [7]. UDH first uses a hiding network to encode the secret images, and then directly adds the cover and encoded secret images to generate the final stego images. This hiding process can be formulated as follows:

$$C' = 1 \times f(S) + 1 \times C. \tag{2}$$

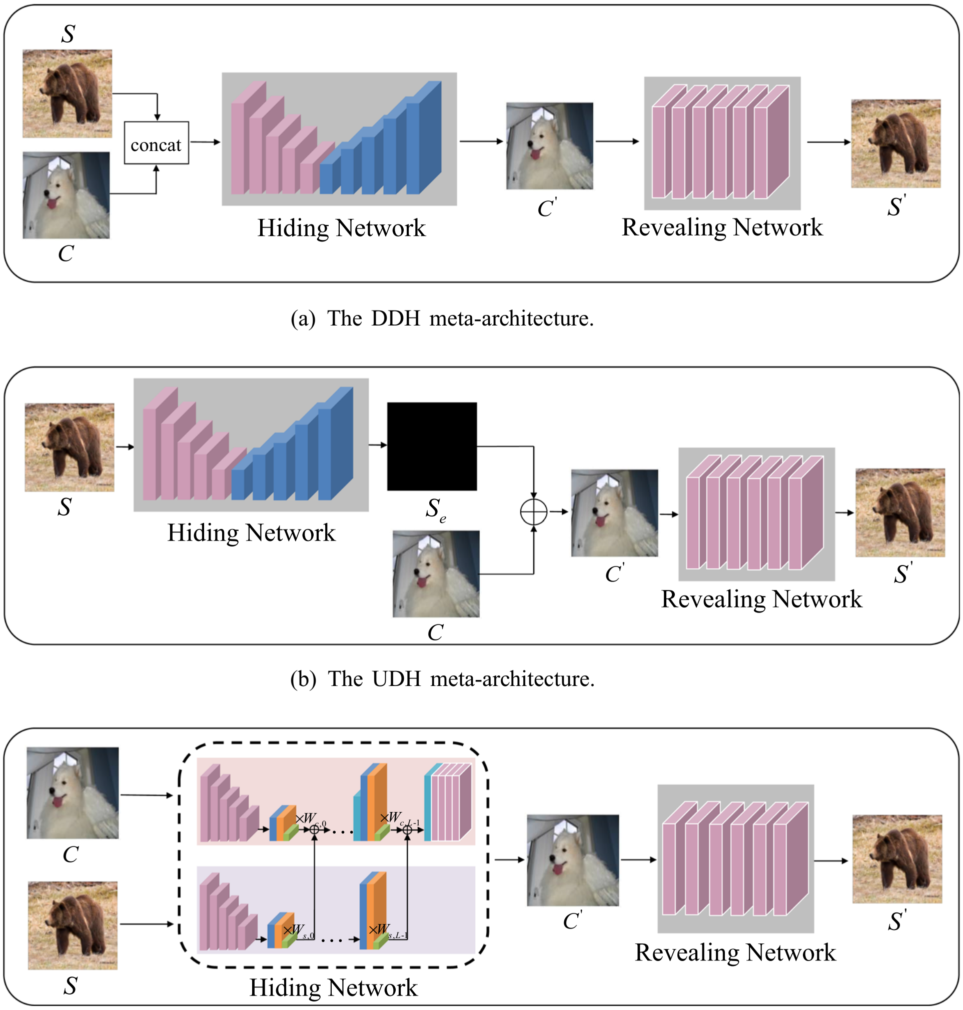
The meta-architecture of UDH is seen in Fig. 1b. In the hiding process, the stego images are produced by directly adding the encoded secret and cover images. Therefore, compared to DDH, it is easy to keep a small value of the cover average pixel discrepancy and the high quality of such stego images [7].

Because the DDH and UDH methods both fuse all of the information of the cover and secret images, additionally entire cover images may not reveal the secret images satisfactorily. Moreover, these methods directly fuse the cover and secret information at the end of the encoding network, leading to the rough fusion. This also causes a large difference between the stego and cover images and such difference is very similar to the secret images, further leading to a severely dangerous transmission in the steganography. Due to these problems, we argue only extracting and fusing the necessary information of secret and cover images may reveal high-quality secret images. Furthermore, owing to different frequency information contained in images and the hierarchical structure of CNN, we can extract and fuse the secret and cover information at the frequency and depth (layer) levels in the deep image hiding network, generating much smaller and safer differences between the stego and cover images as well as the revealed secret images with much higher quality.

Inspired by the above motivations, we propose Deep Adaptive Hiding Network (DAH-Net) using Attentive Frequency Extraction (AFE) and Gradual Depth Extraction (GDE). Specifically, we propose AFE to adaptively extract the necessary secret and cover information within the frequency domain. In detail, the proposed AFE method uses the preset DCT kernels to transform the secret and cover images to the frequency domain. Then, Cross Sharing Attention (CSA) module is proposed in AFE to produce the weights of extracted information. CSA first predicts the attentive weights at the channel, spatial, and frequency extents, respectively. Then, such weights are crossly shared at the mutual extents to generate the final attentive weights. Secondly, GDE is further proposed for the DAH-Net. Because CNN is a hierarchical structure, different layers can retrieve the characteristics with different extraction extents. Therefore, the proposed GDE can gradually extract and fuse the secret and cover information with the increasing depths in the hiding network. Integrating the proposed AFE with GDE methods in our DAH-Net, only the necessary information of secret and cover images is adaptively extracted and the fusion of such extracted information is much finer, producing satisfactory stego and revealed secret images.

In short, the contributions of this paper can be summarized as follows:

- We propose the DAH-Net to adaptively and gradually extract and fuse only the necessary information of secret and cover images at both the frequency and depth extents for image hiding. Such an information hiding pipeline can significantly improve the quality of stego and revealed secret images.
- We propose the AFE method in the DAH-Net to adaptively extract the necessary information of secret and cover images using the DCT filters and CSA. Such extraction information is produced at the frequency



(a) The DDH meta-architecture.

(b) The UDH meta-architecture.

(c) The DAH-Net meta-architecture.

**Fig. 1** The meta-architectures of DDH, UDH, and DAH-Net. S and C denote the secret and cover images, respectively. $C'$ and $S'$ indicate the stego and revealed secret images, respectively. The pink block indicates the convolution layer with the kernel size of 4×4, and the blue block indicates the transposed convolution layer with the kernel size of 4×4. The pink block with a white line indicates the convolution layer with the kernel size of 3×3. In DDH and UDH, the hiding network is a U-Net, and the revealing network consists of several convolution layers. There are two sub-networks in the hiding network of DAH-Net. The yellow block denotes the DCT layer. The green block denotes the AFE block, and the sky-blue block indicates the fusing layer with the kernel size of 1×1

extent, leading to much finer fusion in the subsequent steps.

- We further propose the GDE method to extract and fuse the secret and cover information at the depth (layer) extent of our DAH-Net. Such gradual fusion can generate more satisfactory stego and revealed secret images.
- Extensive experiment results show that the proposed DAH-Net achieve excellent performances on image hiding, watermarking, and photographic steganography, compared to other methods.

The rest of this paper is organized as follows. Section 2 briefly reviews related works of deep image hiding as well as the theory of Discrete Cosine Transform (DCT) and harmonic neural network. Section 3 presents the proposed DAH-Net, AFE, and GDE. Section 4 reports the experiment results for image hiding, watermarking, and photographic steganography. Finally, Sect. 5 presents the conclusions of this work.

## 2 Related works

### 2.1 Image hiding

The existing image hiding approaches fall broadly into two categories, *i.e.*, traditional image hiding [2, 22, 24–28], and deep image hiding methods. Based on the classification of secret information, the existing deep image hiding approaches can mainly be divided into two categories, *i.e.*, low-capacity binary information hiding and large-capacity image information hiding. Our work focuses on large-capacity image information hiding and aims to hide single and multiple secret images in a cover image.

Deep-Stego first achieves hiding the color secret images in the color cover images [20, 21] using CNN. Such a method first introduces a preprocess network to preprocess the secret images and then encodes the concatenation of the preprocessed secret information and cover images using the hiding network. The revealed secret images are produced using another revealing network.

The existing deep hiding pipeline including Deep-Stego falls into one meta-architecture category termed Dependent Deep Hiding (DDH) in [7]. Different from Deep-Stego, the input of the hiding network in DDH is the original secret and cover images. UDH is proposed in [7] for image hiding, watermarking, and photographic steganography. UDH suggests that the success of deep image hiding contributes to the frequency discrepancy between the encoded secret and cover images. UDH utilizes a hiding network to encode the secret images, and then adds such encoded secret and original cover images to directly generate the stego images. The UDH method can not only hide multiple secret images in a cover image but also achieve image hiding across gray and color images. As illustrated in [7], UDH achieves better image hiding performance than DDH. Considering information hiding and revealing processes are a pair of inverse problems, the Invertible Steganography Network (ISN) [29] employs the Invertible Neural Network (INN) [30] for image hiding. Similarly, HiNet [31] achieves single image hiding using INNs in the wavelet domain. Due to the gradient explosion problem, the methods based on INN are difficult to train. This phenomenon is especially serious when multiple secret images are hidden.

This paper proposes the DAH-Net for extensive deep image hiding tasks. Different from the traditional deep image hiding methods, our method extract and fuse the only necessary secret and cover information. Furthermore, the proposed DAH-Net can adaptively and gradually fuse the secret and cover information at both the frequency and depth extents. Integrating these mechanisms, our DAH-Net can generate higher quality stego and revealed secret images for single and multiple image hiding.

### 2.2 Discrete cosine transform

Discrete Cosine Transform (DCT) is commonly used to compress images and videos in JPEG and MPEG formats by removing the high-frequency coefficients within images [32]. DCT-II is commonly used to decompose the image $\boldsymbol{X}$ ($\boldsymbol{X} \in \mathbb{R}^{H\times W}$) to its spatial frequency spectrum by the orthogonal transformation method as follows:

$$\boldsymbol{Y}_{u,v} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\sqrt{\frac{\alpha_u}{H}}\sqrt{\frac{\alpha_v}{W}} \boldsymbol{X}_{i,j}\cos\left(\frac{\pi}{H}\left(i+\frac{1}{2}\right)u\right)\cos\left(\frac{\pi}{W}\left(j+\frac{1}{2}\right)v\right), \tag{3}$$

where the normalized factors $\alpha_u$ and $\alpha_v$ in the basis functions are set to 1, when $u$=0 and $v$=0, and in otherwise cases, these factors are set to 2 to ensure orthonormality. $\boldsymbol{Y}_{u,v}$ indicates the DCT coefficient for $\boldsymbol{X}$ using the orthogonal transformation at $u^{th}$ and $v^{th}$ frequency in the horizontal and vertical orientations, respectively.

### 2.3 Harmonic neural network

Harmonic Neural Network (HNN) [33] proposes harmonic convolution to extract different frequency information of the input feature. There are two layers in the harmonic convolution, *i.e.*, the DCT layer and the fusing layer. The DCT layer extracts different frequency information of the input feature $\boldsymbol{X}$ ($\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$) using the preset feature extractors (DCT filters) $\boldsymbol{W}_{\psi}$ with a window-based strategy. When the kernel size of the DCT filters is set to $k$, the number of the DCT filters $n$ is $k^2$, and the size of DCT filters is "$n\times k\times k$". Then, the fusing layer learns the weights of different frequency feature channels using a point-wise convolution layer with the combinational filter $\boldsymbol{W}_{\phi}$ ($\boldsymbol{W}_{\phi} \in \mathbb{R}^{M\times Cn\times 1\times 1}$) and produces the output feature $\boldsymbol{Y}$ ($\boldsymbol{Y} \in \mathbb{R}^{M\times H'\times W'}$). The transformation process of $j^{th}$ ($j \in \{0, 1, \ldots, M-1\}$) output channel can be formulated as follows:

$$\boldsymbol{Y}_j = \sum_{i=0}^{N-1}\sum_{f=0}^{k^2-1}\boldsymbol{W}_{\phi}^{i,j,f}\boldsymbol{W}_{\psi}^{f}\otimes\boldsymbol{X}_i, \tag{4}$$

where $\otimes$ represents the 2D convolution operation.

## 3 Method

### 3.1 DAH-net

We propose the DAH-Net using AFE and GDE to adaptively and gradually extract and fuse only the essential secret and cover information. This process can be formulated as follows:

$$C^{'} = W_s \times f_s(S) + W_c \times f_c(C), \tag{5}$$

where $f_s$ and $f_c$ indicate the transformations of the hiding network for the secret and cover images, respectively. $W_c$ and $W_s$ indicate the weights of cover and secret images to be fused, respectively. $S$, $C$, and $C^{'}$ denote the secret, cover, and stego images, respectively.

The meta-architecture of the proposed DAH-Net is shown in Fig. 1c. The DAH-Net consists of two networks, *i.e.*, the hiding and revealing networks. There are two sub-networks in the hiding network to extract the secret and cover information, respectively. Each of the sub-network is an architecture of U-Net [34] including the encoder and decoder stages. Using two sub-networks in the hiding network, the proposed DAH-Net separately encodes the secret and cover images and then gradually extracts and

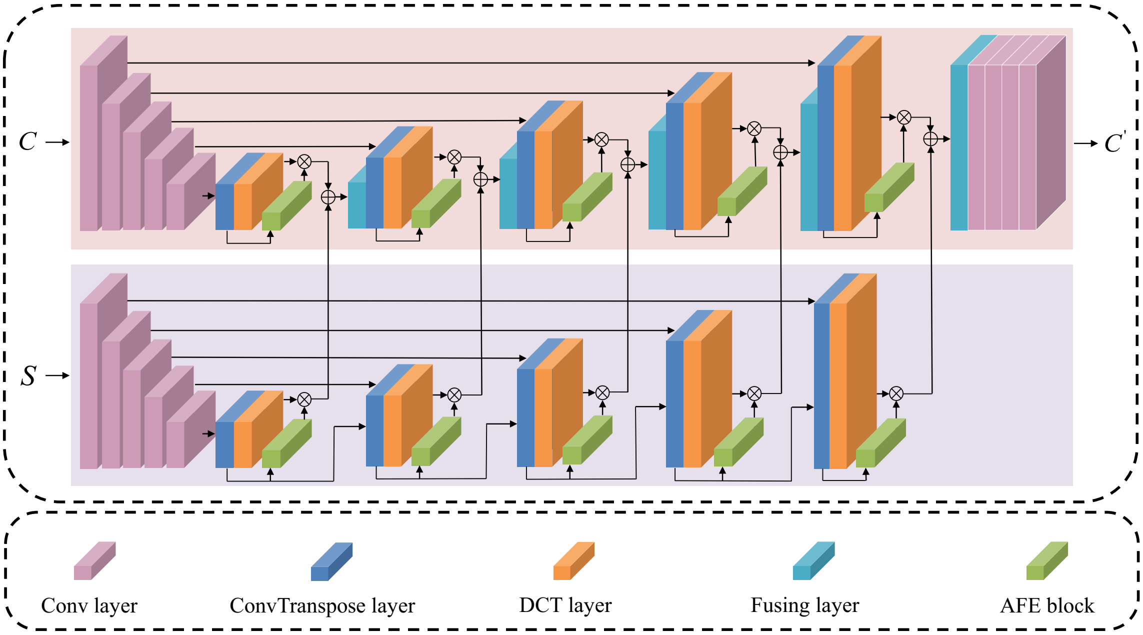


**Fig. 2** The architecture of the hiding network in the proposed DAH-Net. S, C, and $C'$ denote the secret, cover, and stego images, respectively. The pink block indicates the convolution layer with the kernel size of 4×4, and the pink block with a white line denotes the convolution layer with the kernel size of 3×3. The fusing layer is the convolution layer with the kernel size of 1×1

fuses the necessary information of the secret and cover images from different frequencies and increasing depths at the decoding stage of the sub-network. There are 5 convolution layers and 5 transposed convolution layers in each sub-network, and the kernel size and stride are set to 4 and 2, respectively. The input and output channel numbers of these layers are set to $\{3, 64, 128, 256, 512, 512, 1024, 512, 256, 128\}$ and $\{64, 128, 256, 512, 512, 256, 128, 64, 3\}$, respectively. The input and output channel numbers of the DCT layer are the same and are set to $\{512, 256, 128, 64, 3\}$, respectively. After the last extraction and fusion of the secret and cover information, 4 convolution layers are used to further fuse the secret and cover information. The input and output channel numbers of these convolution layers are set to $\{3, 64, 128, 64\}$ and $\{64, 128, 64, 3\}$, respectively. A normalization layer and a ReLU layer follow the first three convolution layers, and a sigmoid layer is attached after the last convolution layer. The details of the hiding network in the proposed DAH-Net are seen in Fig. 2. The secret images can be revealed by the revealing network with high quality, and the revealing network is trained with the hiding network. In our work, the revealing network is stacked by six convolution layers and a sigmoid layer, and the kernel size and stride of the convolution layers are 3 and 1, respectively. The input and output channels of the convolution layer in the revealing network are $\{3, 64, 128, 256, 128, 64\}$ and $\{64, 128, 256, 128, 64, 3\}$, respectively. The details of the revealing network are described in Fig. 1c. The revealing process also can be formulated as the following equation:

$$S' = f_R(C'), \tag{6}$$

where $S'$ indicates the revealed secret images. $f_R$ denotes the revealing process. The optimization goal is to minimize the loss function defined as the following equation:

$$L(S, S', C, C') = \|C' - C\| + \beta\|S' - S\|, \tag{7}$$

where following [20] and [7] the revealing weight $\beta$ is set to 0.75 in our work. The process of DAH-Net is shown in Algorithm 1.

Algorithm 1 : The algorithm of DAH-Net

```
1: Initialize: The number of secret images: $n_s$ and cover images: $n_c$;
       Spatial kernel size of the DCT layer: $k \times k$, and the number
   of the DCT filter: $n = k^2$;
       Depth of the decoder for extracting and fusing information:
   $L$;
       Parameters in the hiding and revealing networks:
   $\boldsymbol{\theta} = \{\boldsymbol{\theta}_H^C, \boldsymbol{\theta}_H^S, \boldsymbol{\theta}_R\}$;
       Attentive frequency weights $W_f$ of the secret and cover
   images;
       Reduction ratio $r$ in AFE;
       Revealing weight $\beta$ in loss function.
2: Repeat:
3:     Input of the hiding network in DAH-Net: the cover
   images C and the secret images S ;
```

Algorithm 1 : The algorithm of DAH-Net

```
4:      Extract different frequency information of the secret and
    cover images based on Eqn. (12);
5:      Generate the channel-spatial weight $W_t$ of secret and cover
    images based on Eqns. (12) - (19);
6:      Generate the cross share attention weight $W$ of secret and
    cover information based on Eqns. (9) - (11);
7:      Gradually fuse the extracted secret and cover information,
    and generate the stego images $C'$ based on Eqn. (21);
8:      Generate the revealed secret images $S'$ based on Eqn. (6);
9:      Compute loss based on Eqn. (7), update the attentive
    frequency weights $W_f$ and $\boldsymbol{\theta}$.
10: Until convergence
```

## 3.2 Attentive frequency extraction

To obtain the stego and revealed secret images with superior image quality, Attentive Frequency Extraction (AFE) is proposed to adaptively extract and fuse only the relatively essential part of the secret and cover information.

The extracted essential part of cover information ensures the quality and security of the stego images, and the extracted essential secret information ensures the secret images could be revealed by the revealing network with high quality. We further propose Cross Sharing Attention (CSA) module in AFE to produce the weights of extracted information to be fused. The structure of AFE is shown in Fig. 3 in detail.

### 3.2.1 Entire attentive frequency extraction structure

Given the input $\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{C-1}] \in \mathbb{R}^{C\times H\times W}$ ($\boldsymbol{x}_i \in \mathbb{R}^{H\times W}$, $i \in \{0, 1, \cdots, C-1\}$), the transformation of DCT layer with the kernel size of $k$ can be formulated as follows:

$$\boldsymbol{Y} = \sum_{f=0}^{k^2-1} \boldsymbol{W}_{\psi}^{f} \otimes \boldsymbol{X}, \tag{8}$$

where the output $\boldsymbol{Y} = [\boldsymbol{y}_0, \boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{(Cn-1)}] \in \mathbb{R}^{Cn\times H\times W}$ ($\boldsymbol{y}_i \in \mathbb{R}^{H\times W}$, $i \in \{0, 1, \cdots, Cn-1\}$). $\boldsymbol{W}_{\psi}$ indicates the DCT filters with the size of "$n \times k \times k$".

In order to keep the consistency with the kernel size of the regular convolution, the kernel size of the DCT layer $k$ is set to 3. $n$ is the number of DCT filters, and $n = k^2 = 9$. The output of the DCT layer can also be written as $\boldsymbol{Y} = [\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{C-1}]$, and $\boldsymbol{Y}_i = [\boldsymbol{y}_i^0, \boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^{n-1}]$, $i \in \{0, 1, \ldots, C-1\}$.

This indicates that different frequency features $\boldsymbol{Y}_i$ are produced by the convolution operation between $\boldsymbol{X}_i$ and the DCT filters $\boldsymbol{W}_{\psi}$. In another word, $\{\boldsymbol{y}_i^0, \boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^{n_f-1}\}$ share the same channel and spatial information of $\boldsymbol{X}_i$. Meanwhile, $\{\boldsymbol{y}_0^f, \boldsymbol{y}_1^f, \ldots, \boldsymbol{y}_{C-1}^f\}$ ($f \in \{0, 1, \ldots, n-1\}$) are all produced from $f^{th}$ DCT filter, and these channels share the same frequency information. Therefore, we introduce the

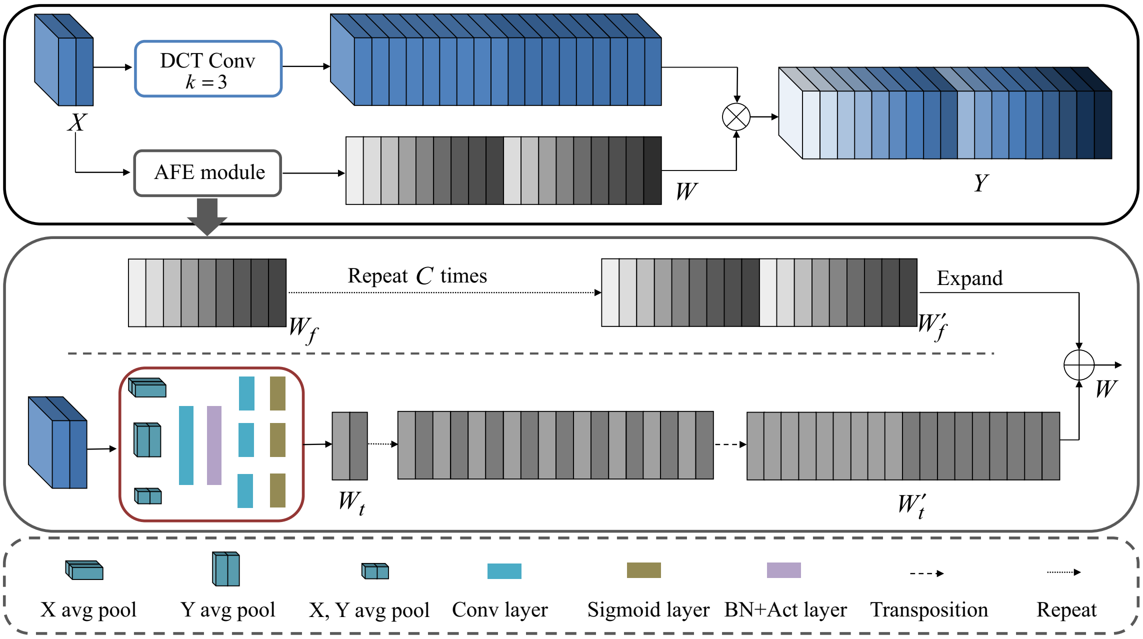


**Fig. 3** The structure of the proposed AFE. The number of the input feature channels $C = 2$, and the number of the DCT kernel $n = 9$. The output channels of the DCT layer are 18. The pooling kernel of the pooling layer at the height, width, and spatial dimensions are (H, 1), (W, 1), and (H, W), respectively

CSA method in the proposed AFE to predict the channel, spatial, and frequency weights, and then crossly share these weights at the mutual frequency and channel-spatial extents to generate the final attentions.

### 3.2.2 Cross sharing attention

The weights obtained by CSA $W \in \mathbb{R}^{Cn\times H\times W}$ can be formulated as follows:

$$W = f_e(W'_f \to W'_t) + W'_t, \tag{9}$$

where $W'_f \in \mathbb{R}^{Cn\times 1\times 1}$ indicates the final attentive frequency weights, and $W'_t \in \mathbb{R}^{Cn\times H\times W}$ denotes the final attentive channel-spatial weights. $f_e(\cdot \to \cdot)$ indicates the expanding operation to keep the size of the left side consistent with that of the right side. The channels produced from the same DCT filter have the same frequency information. Therefore, CSA is proposed in AFE to share the same attentive frequency weights among such channels. Thus, we initialize the original frequency weights $W_f = [w_f^0, w_f^1, \ldots, w_f^{n-1}] \in \mathbb{R}^{n\times 1\times 1}$ ($w_f^i \in \mathbb{R}^{1\times 1}$, $i \in \{0, 1, \cdots, n-1\}$) before training the DAH-Net, and update $W_f$ iteratively in the training process. Then, $W_f$ is repeated for $C$ times to formulate the final shared weights $W'_f$:

$$W'_f = f_r^C(W_f), \tag{10}$$

where $W'_f = [W_{f,0}, W_{f,1}, \ldots, W_{f,C-1}]$, and $W_{f,0} = W_{f,1} = \ldots = W_{f,C-1} = W_f = [w_f^0, w_f^1, \ldots, w_f^{n-1}]$. $f_r^C$ indicates the repeat operation with $C$ times on the channel dimension. These adaptive frequency weights can extract the essential frequency information with less computation.

As mentioned before, $\{\boldsymbol{y}_i^0, \boldsymbol{y}_i^1, \ldots, \boldsymbol{y}_i^{n-1}\}$ ($i \in \{0, 1, \ldots, C-1\}$) obtained from the same channels of the DCT layer input $\boldsymbol{x}_i$ share the same channel and spatial information. Thus, the channel and spatial attentive weights are shared on the frequency channels produced from the same input channel through the DCT convolution layer. This sharing process can be formulated as the following equation:

$$W'_t = (f_r^n(W_t))^{T(1,2)}, \tag{11}$$

where $W_t = [w_{t,0}, w_{t,1}, \ldots, w_{t,C-1}] \in \mathbb{R}^{C\times H\times W}$ ($w_{t,i} \in \mathbb{R}^{H\times W}$, $i \in \{0, 1, \cdots, C-1\}$) is the channel-spatial weights generated from the input $\boldsymbol{X}$. $W'_t = [W_{t,0}, W_{t,1}, \cdots, W_{t,C-1}]$, and $W_{t,i} = [w_{t,i}^0, w_{t,i}^1, \cdots, w_{t,i}^{n-1}] \in \mathbb{R}^{n\times H\times W}$ ($i \in \{0, 1, \cdots, C-1\}$). In addition, $w_{t,i}^j = w_{t,i}$ $j \in \{0, 1, \cdots, n-1\}$.$f_r^n$ indicates the repeat operation with $n$ times on the channel dimension. $T^{(1,2)}$ indicates the transpose operation between the first and second axes.

To generate both the channel-wise and location attentions in $W_t$, we first use three pooling kernels to encode each feature map. The pooling kernel sizes on the width, height, and channel axes are (H, 1), (1, W), and (H, W), respectively. Given the input $\boldsymbol{X}$, the pooling processes with these three pooling kernels can be formulated as the following three equations:

$$\zeta_c^h(h) = \frac{1}{W}\sum_{i=0}^{W-1} \boldsymbol{x}_c(i, h), \tag{12}$$

$$\zeta_c^w(w) = \frac{1}{H}\sum_{j=0}^{H-1} \boldsymbol{x}_c(j, w), \tag{13}$$

$$\zeta_c^{h,w}(c) = \frac{1}{H \times W}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1} \boldsymbol{x}_c(i, j), \tag{14}$$

where $\zeta_c^h$, $\zeta_c^w$, and $\zeta_c^{h,w}$ are the output of the pooling operations using pooling kernels ($H$, 1), (1, $W$), and ($H$, $W$), respectively.

To save computation, these three pooling outputs are concatenated on the height dimension and fed to a shared point-wise convolution layer $f_p$ with kernel size $1 \times 1$. This projection can be formulated as follows:

$$\ell = \delta(f_p[\zeta^h \oplus \zeta^w \oplus \zeta^{h,w}]), \tag{15}$$

where $\ell \in \mathbb{R}^{C/r\times(H+W+1)}$, and $r$ is the reduction ratio, $r \in (0, 1]$. $\oplus$ denotes the concatenation operation. $\delta$ is the nonlinear ReLU activation function.

Then, $\ell^h$, $\ell^w$, and $\ell^{h,w}$ are generated by splitting $\ell$ on the height dimension with size $\{H, W, 1\}$. Another three point-wise convolution layers $f_h, f_w$ and $f_{h,w}$ are applied to process $\ell^h$, $\ell^w$, and $\ell^{h,w}$, respectively. This process can be formulated as the following equations:

$$\eta^h = \sigma(f_h(\ell^h)), \tag{16}$$

$$\eta^w = \sigma(f_w(\ell^w)), \tag{17}$$

$$\eta^{h,w} = \sigma(f_{h,w}(\ell^{h,w})). \tag{18}$$

Finally, the output weights $W_t$ can be written as:

$$w_t(i,j) = \eta_c^{h,w}(i,j) \times \eta_c^h(i) \times \eta_c^w(j). \tag{19}$$

## 3.3 Gradual depth extraction

Because the CNN are hierarchical structures, different characteristics with different extraction extents can be retrieved in different layers. The shallow layers of CNN usually retrieve the detailed information while the structure information is retrieved in the deep layers. This information on depth levels within the features can guide the steganography process. Therefore, the Gradual Depth Extraction (GDE) is proposed to gradually extract and fuse the secret and cover information from different depths of the network. Due to the proposed AFE, only the essential part of the secret and cover information is extracted from the increasing depths in the hiding network. Further, we gradually extract and fuse the essential secret and cover information at the decoder stages. We formulate this process as the following equation:

$$C_l' = W_{c,l} \times G_l\left(\sum_{f=0}^{k^2-1} \boldsymbol{W}_{\psi}^{f} \otimes (C_l)\right) + W_{s,l} \times G_l\left((\sum_{f=0}^{k^2-1} \boldsymbol{W}_{\psi}^{f} \otimes (S_l)\right), \tag{20}$$

where $l \in \{0, 1, \cdots, L-1\}$, $L$ is set to 5. $\boldsymbol{W}_{\psi}$ denotes the DCT filter. $G_l(\cdot)$ indicates extracting information from $l^{th}$ layer in the hiding network. Besides, we gradually fuse the extracted secret and cover information after every transposed convolution layer in the hiding network. The output sizes of the transposed convolution layers are 8×8, 16×16, 32×32, 64×64, 128×128, respectively. The sizes of feature maps are different at various fusion stages. Thus, the GDE also extracts and fuses the secret and cover information from different scales within the pyramidal structure. This multi-level and fine-grained information extraction and fusion method ensures the high quality of both the stego and revealed secret images. In $l^{th}$ fusion layer, the corresponding frequencies of secret images are added to different frequencies of cover images. Eqn. (20) is further formulated as follows:

$$C_l'^{f} = W_{c,l}^{f} \times G_l(C_l^{f}) + W_{s,l}^{f} \times G_l(S_l^{f}), \tag{21}$$

where $l \in \{0, 1, \cdots, L-1\}$, and $f \in \{0, 1, \cdots, n-1\}$. $n$ is the number of DCT filters, and it is set to 9. In summary, the GDE method extracts the essential frequency information of the secret and cover images from different depths in the hiding network. Integrating the proposed AFE with GDE methods, the DAH-Net can extract and fuse only the necessary secret information at both frequency and depth extents. Such extraction and fusion processes are much finer compared to the traditional deep image hiding methods. These proposed methods can drive the DAH-Net to generate more high-quality stego and revealed secret images with significant safety promotion.

# 4 Experiments

In this section, extensive experiments are conducted to prove the effectiveness of the proposed DAH-Net. Section 4.1 introduces datasets and initializations. The quantitative and qualitative comparisons on image hiding, watermarking, and photographic steganography are provided in Sects. 4.1 and 4.2. The effectiveness of GDE and AFE in the proposed DAH-Net is verified in Sect. 4.4.1. Section 4.4.2 further discusses the adaptiveness of attentive frequency weights. The comparisons of security performances are introduced in Sect. 4.4.3. Section 4.4.4 analyzes the robustness of our DAH-Net. Section 4.4.5 provides the effect of kernel size of convolution layer. Section 4.4.6 discusses the limitations of the proposed DAH-Net.

## 4.1 Dataset and initializations

The proposed DAH-Net is compared to state-of-the-art methods on the regular ImageNet datasets [35]. The size of input images is uniformly resized as 128×128, and the mini-batch size B is set to 44. $n_s$ and $n_c$ represent the numbers of secret and cover images, respectively. We randomly select "$n_c \times B \times 2000 + n_s \times B \times 200$" and "$n_c \times B \times 200 + n_s \times B \times 200$" images from ImageNet as the training and testing sets, respectively. The Adam optimizer is used in the training process with $\beta_1$=0.5 and $\beta_2$=0.999. The learning rate starts from 0.001 and is divided by 10 every 30 epochs. The proposed DAH-Net is trained by randomly selecting the cover and secret images from the dataset for 120 epochs. Following with DeepStego [20] and UDH [7], the revealing weight $\beta$ is set to 0.75 in our work. The spatial kernel size of DCT layer is set to "3 × 3", and thus the number of the DCT filters is 9. The depth of the decoder for extracting and fusing information $L$ is set to 5, and the reduction ratio $r$ in AFE is set to 32.

Following UDH [7] and Cycle-GAN [34], the kernel sizes of convolution layer within encoding network and transpose convolution layer within decoding network are both set to 4 × 4. Hence, we follow such common practices to establish our network. Specifically, the kernel size of convolution layer is usually determined by that of transpose convolution layer, to keep the consistently same receptive field in both the encoding and decoding stages to some extent, producing satisfactory performance.

At first, the reason for choosing the kernel size of transpose convolution layer is introduced as follows:

Suppose the input and output of the convolution layer $I$ and $O$ have the sizes of $i \times i$ and $o \times o$, respectively. The output calculation of convolution layer is formulated as follows:

$$o = \left\lfloor \frac{i + 2 \times p - \kappa}{s} + 1 \right\rfloor, \tag{22}$$

where $p$, $\kappa$, and $s$ indicate padding, kernel size, and stride, respectively. In U-Net, after passing through convolution layer, the height and width of the feature map are halved.

Meanwhile, the feature map with a halving image scale can be reconstructed through the transposed convolution layer, and the output calculation of transposed convolution layer is formulated as follows:

$$i = (o - 1) \times s - 2 \times p + \kappa. \tag{23}$$

In order to ensure $i = 2 \times o$, $s$ should be set to 2, and the result of $2p - \kappa$ needs to be equal to 2. Obviously, $\kappa = 3$ or 5 cannot meet the requirements. Thus, the kernel size of the transpose convolution layer is set to $4 \times 4$. Therefore, the kernel sizes of the convolution layer and transpose convolution layer in the proposed hiding network are both set to $4 \times 4$.

## 4.2 Performances of image hiding

Image hiding or image steganography usually hides the secret images within cover images for covert communication. The sender hides the secret images into cover images and generates the stego images for transmission. After transmission, the receiver recovers the secret images from the received stego images to complete the secret communication. In order to ensure that the covert communication will not be easily interrupted or broken by the third parties to a certain extent, the stego images need to have high image-quality and high similarity with the cover images. The quality of the revealed secret images directly affects whether the secret information can be transmitted to the receiver accurately. The image-quality of the recovered secret images and the similarity with the original secret images are used to measure the accuracy of the revealed secret information. Average Pixel Discrepancy (APD), Peak Signal-to-noise Ratio (PSNR), and Structural Similarity (SSIM) are applied to evaluate performance of image hiding. APD indicates the average residual error between all pixels of two images. The lower APD represents better steganography performance. Peak Signal-to-noise Ratio (PSNR) and Structural Similarity (SSIM) are used to evaluate the image quality. Higher PSNR and SSIM denote better steganography performance.

### 4.2.1 Quantitative results

We compare our DAH-Net with existing image hiding methods, *i.e.*, Deep-Stego [20, 21], UDH [7], ISN [29] and HiNet [31]. The experiment results on hiding $\{1, 2, 3\}$ secret image in a cover image are reported in Table 1. It is noted the experimental results provided in this paper are the average of all test data. HiNet is proposed only for single secret image hiding. Meanwhile, due to the gradient explosion problem, HiNet is hard to train especially when hiding multiple images. Thus, the results of HiNet on multiple image hiding are not provided. In addition, there is no distortion in the processes of the transmission of stego images and the revealing of secret images. Meanwhile, our DAH-Net and all comparison methods are conducted under the same experimental conditions.

When hiding a single secret image in a cover image, the quality of stego and revealed secret images are effectively improved by our DAH-Net. Specifically, compared to the HiNet, the proposed DAH-Net improves the PSNRs of the stego and revealed images by 1.18 dB and 0.65 dB, respectively. Our DAH-Net also provides 0.034 and 0.032 improvement for SSIM of the stego and revealed secret images, compared to the ISN, respectively. Furthermore, the cover and secret APD is significantly decreased by the proposed DAH-Net. Specifically, compared to UDH, the proposed DAH-Net decreases the cover and secret APD by 0.33 and 0.88, respectively.

A larger steganography capacity will reduce the quality of the stego and revealed secret images in image hiding. Although most existing image hiding approaches only achieve single and dual image hiding, results of larger capacities image hiding are also provided in this paper for further exploring the potential of our DAH-Net. From the comparisons provided in Table 1, it is obvious that our DAH-Net achieves better image hiding performance on large capacities than other deep image hiding methods. Specifically, when hiding two secret images in a cover image, the proposed DAH-Net decreases the cover and secret APD by 0.37 and 1.10, compared to UDH, respectively. The proposed DAH-Net also achieves 0.035 and 0.039 improvement than the ISN for the SSIM of the stego and revealed secret images, respectively. At the task of hiding three images in a cover image, DAH-Net improves the PSNRs of the stego and revealed secret images by 1.1 dB and 2.71 dB, compared to the UDH, respectively.

The above quantitative results demonstrate that our DAH-Net can effectively improve the quality of stego and revealed secret images at different capacities. Since only the necessary cover and secret information are extracted and fused, our DAH-Net significantly promotes the quality of stego and revealed secret images. The improvement of cover image quality may also be attributed to multiple finer fusions used in different layers and frequency domains in the hiding stage.

### 4.2.2 Qualitative results

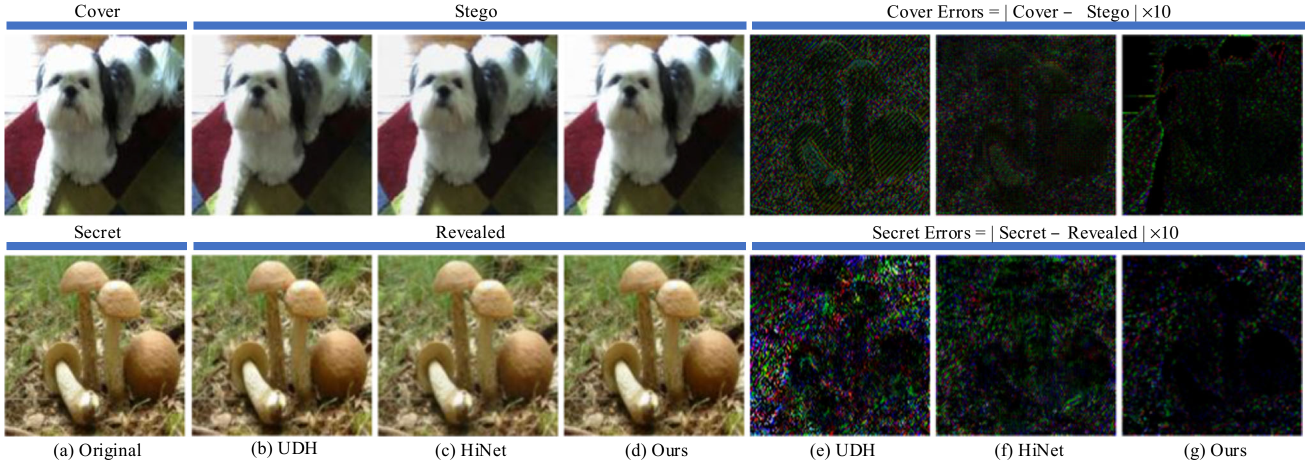
Figure 4 shows the comparison results of single image hiding and revealing. Cover errors indicate the residual information between cover and stego images. Correspondingly, secret errors denote the residual information between secret and revealed secret images. Obviously, there is almost no visual difference among the original

**Table 1** Comparisons of image hiding performances when hiding $\{1, 2, 3\}$ secret images in a cover image

| $n_s$ | $n_c$ | Method | APD-C↓ | PSNR-C↑ | SSIM-C↑ | APD-S↓ | PSNR-S ↑ | SSIM-S↑ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | Deep-stego [20 ,21] | 2.8 | 36.02 dB | 0.946 | 3.6 | 32.75 dB | 0.933 |
| | | UDH [7] | 2.35 | 39.13 dB | 0.985 | 3.56 | 35.00 dB | 0.976 |
| | | ISN [29] | – | 38.05 dB | 0.954 | – | 35.38 dB | 0.955 |
| | | HiNet [31] | 2.18★ | 38.75 dB★ | 0.987★ | 2.72★ | 36.66 dB★ | 0.985★ |
| | | **DAH-Net (Ours)** | **2.02** | **39.93 dB** | **0.988** | **2.68** | **37.31 dB** | **0.987** |
| 2 | 1 | Deep-stego [20, 21] | – | 30.18 dB | 0.919 | – | 29.17 dB | 0.898 |
| | | UDH [7] | 3.20★ | 36.58 dB ★ | 0.972★ | 5.74★ | 30.41 dB ★ | 0.940★ |
| | | ISN [29] | - | 36.86 dB | 0.945 | – | 32.21 dB | 0.920 |
| | | **DAH-Net (Ours)** | **2.83** | **36.88 dB** | **0.980** | **4.64** | **32.23 dB** | **0.959** |
| 3 | 1 | Deep-stego [20, 21] | – | – | – | – | – | – |
| | | UDH [7] | 3.42/3.57★ | 35.61 dB★ | 0.963★ | 6.74/6.67★ | 28.73 dB | 0.918★ |
| | | ISN [29] | – | – | – | – | – | – |
| | | **DAH-Net (Ours)** | **2.91** | **36.71 dB** | **0.977** | **5.04** | **31.44 dB** | **0.955** |

★Denotes the results we re-implemented. $n_s$ and $n_c$ represent the numbers of secret and cover images, respectively. -C and -S indicate the performance of the stego and revealed secret images, respectively

The best result is bold


Cover
Stego
Cover Errors = | Cover – Stego | ×10
Secret
Revealed
Secret Errors = | Secret – Revealed | ×10

(a) Original (b) UDH (c) HiNet (d) Ours (e) UDH (f) HiNet (g) Ours

**Fig. 4** Visual comparisons for hiding a secret image in a cover image

cover image and the stego images generated by UDH, HiNet, and our DAH-Net. Thus, cover errors are introduced to visually measure the difference between cover and stego images. It is clear that the cover errors magnified 10 times of UDH and HiNet are mainly the edge information of the secret images. However, the cover errors of our DAH-Net contain little secret information. That indicates the stego images generated by UDH and HiNet are not safe enough when the cover images are available for the third parties, while our DAH-Net can avoid this dangerous risk to some extent. In addition, the cover errors generated by our DAH-Net are almost negligible, compared to UDH and HiNet. This show the higher quality of the stego images generated by the proposed DAH-Net.

Due to the high visual similarity among secret images, revealed secret images reconstructed by UDH, HiNet, and our DAH-Net, secret errors are also introduced to visually compare the quality of revealed secret images. It is obvious that the secret errors magnified 10 times of our DAH-Net are the least visually obvious in all the compared methods. This indicates that the revealed secret images generated by our DAH-Net have more similarities with the original secret images. This proves our DAH-Net can reveal the higher quality of the revealed secret images. To further verify the steganography performance of our DAH-Net on

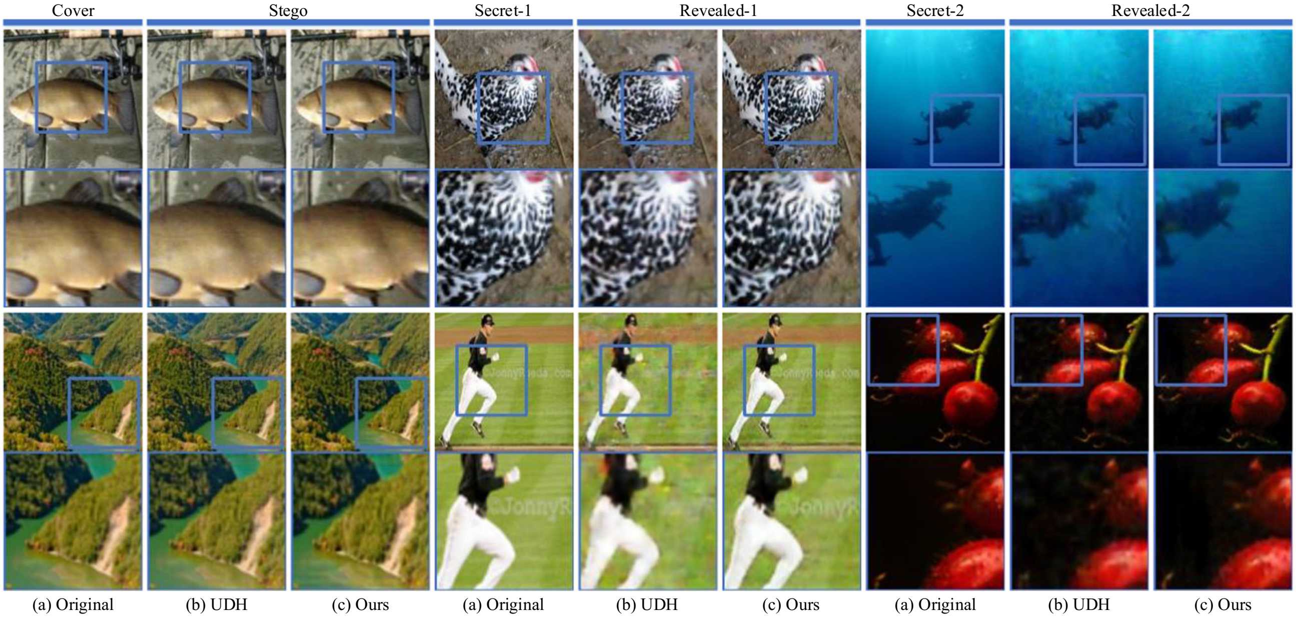


**Fig. 5** Visual comparisons for hiding two secret images in a cover image

large capacity, the visualizations of hiding two secret images are shown in Fig. 5. We enlarge the details of the stego and revealed secret images generated by UDH and our DAH-Net for further comparison. From these figures, our method can reconstruct the details and smooth the areas of secret images more effectively than the UDH, producing the revealed secret images closer to the original secret images. The visualization results of hiding three secret images are shown in Fig. 6. It is clear that our DAH-Net can still reveal the secret images with high quality even when hiding three secret images. According to all of the above results, the proposed DAH-Net can generate the stego and revealed secret images with high quality at single and multiple image hiding.

## 4.3 Universality

### 4.3.1 Watermarking by hiding barcode

The proposed DAH-Net is universal for the task of watermarking. In order to evaluate the robustness of our DAH-Net and other state-of-the-art methods for watermarking, different distortions, *i.e.*, Gaussian blurring, pixel-wise dropout, and JPEG compression are applied on the stego images in the process of the transmission of stego images. The dropout probability $p$ is set to 0.3, and the Gaussian kernel is set to 3 with variance $\sigma = 2$. Meanwhile, bits accuracy under different distortions is utilized to evaluate the robustness of watermarking methods. UDH

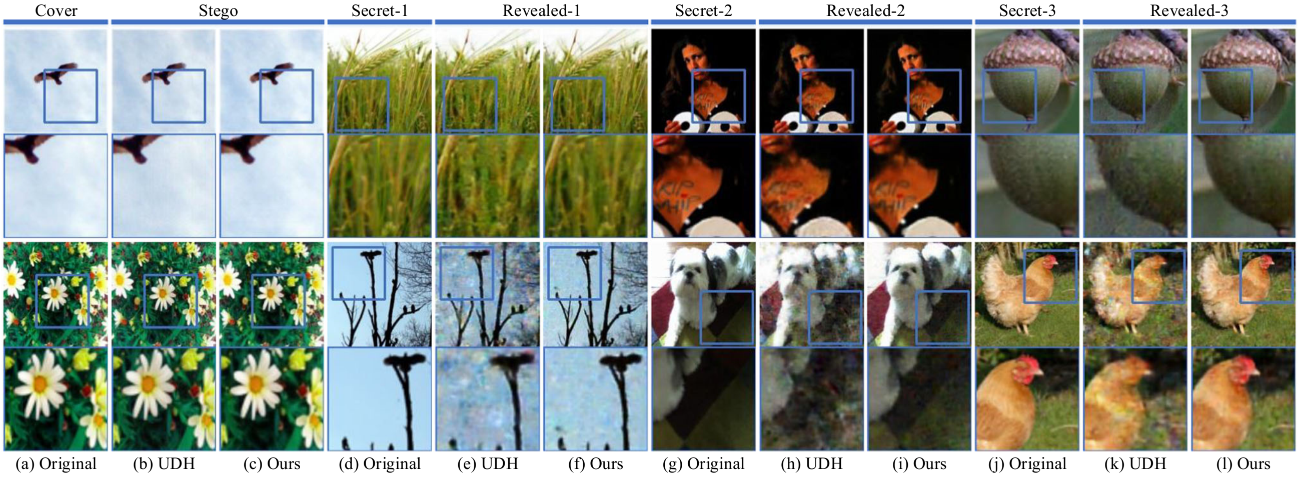


**Fig. 6** Visual comparisons for hiding three images in a cover image

proposes transforming the byte information to bit information by setting the pixel intensity lower than 128 as bit 0 and that higher than 128 as bit 1. The information in barcodes is pseudo-binary information. Following UDH, we divide the secret images into small pieces, each of which has a value of 0 or 255 to represent the binary 0 or 1. When the patch size is set to 8×8×3, a cover image with the size of 128×128×3 can hide 256 bits.

The performances of our DAH-Net compared to UDH and HiNet are demonstrated in Table 2. We re-implement UDH and HiNet under the same experiment settings as DAH-Net. From Table 2, our DAH-Net can achieve better performances in most cases compared to the UDH and HiNet methods.

Especially, the proposed DAH-Net improves the bits accuracy by 23.2% and 32.2% using patch size 2×2×3 compared to the UDH method under the JEPG-50 and JPEG-85, respectively. Compared to HiNet, our DAH-Net also promotes the identity bits accuracy by 6.4% when hiding 1024 bits of information. In order to further verify the results of hiding barcodes by DAH-Net, the visualization of watermarking by hiding barcodes with the patch size of 8×8×3 on DAH-Net is seen in Fig. 7. It is obvious that our DAH-Net could successfully hide barcodes in an image under different distortions. From all of the above comparisons, our DAH-Net is also more effective in watermarking compared to the other methods.

### 4.3.2 Photographic steganography

The proposed DAH-Net is also universal for the task of photographic steganography. Photographic steganography requires the secret images should be revealed from the captured photos of the stego images displayed on the screen. Owing to the color differences between different displaying devices, photographic steganography is also named Light Field Messaging (LFM). In the task of photographic steganography, the final obtained stego images are usually cropped. Thus, it is more difficult to reveal the secret information of photographic steganography. Different from traditional LFM methods [36], we use random homography matrices and uniform noise to simulate the noise of photographic steganography in the process of the transmission of stego images. Moreover, we still hide a secret image instead of a barcode. We re-implement UDH and HiNet under the same experiment settings as the DAH-Net. The photographic steganography performances of DAH-Net compared to the UDH and HiNet are presented in Table 3. The results demonstrate that our DAH-Net can reveal the secret images from the captured stego images with higher quality. Especially, the proposed DAH-Net improves the PSNR and SSIM of the revealed secret images by 0.98 dB and 0.029, compared to the HiNet, respectively. Meanwhile, compared to the UDH, the proposed DAH-Net decreases the APD of the stego images by 0.53 and produces competitive results on other evaluations. To intuitively illustrate the advantages of our DAH-Net, the visualization comparisons among the UDH, HiNet, and our DAH-Net for photographic steganography are shown in Fig. 8. It is apparent that our DAH-Net reveals the secret images with higher quality than the UDH and HiNet on photographic steganography. All of the above results show the satisfactory effect of the revealing secret images produced by the proposed DAH-Net on photographic steganography.

## 4.4 Ablation study

### 4.4.1 Effect of GDE and AFE

To verify the effectiveness of the proposed GDE and AFE methods, we gradually equip the GDE and AFE in the DAH-Net and evaluate these different networks. The comparison results are reported in Table 4. The SSIM of

**Table 2** Bits accuracy for hiding a barcode in an image under different distortions

| Method | Total bits | Patch size | Identity (%) | Dropout (%) | Gaussian (%) | JPEG-50 (%) | JPEG-85 (%) |
|---|---|---|---|---|---|---|---|
| | 256 | 8×8×3 | 100 | 98.4 | 81.3 | 93.6 | 89.6 |
| UDH [7] | 1024 | 4×4×3 | 99.9 | 88.0 | 60.0 | 74.7 | 71.5 |
| | 4096 | 2×2×3 | 98.2 | 76.0 | 52.2 | 61.0 | 61.0 |
| | 256 | 8×8×3 | 100 | 99.7 | **98.21** | 99.52 | 99.80 |
| HiNet [31] | 1024 | 4×4×3 | 93.6 | 83.98 | 85.68 | 90.83 | 93.79 |
| | 4096 | 2×2×3 | 71.5 | 65.08 | 64.38 | 65.78 | 68.5 |
| | 256 | 8×8×3 | 100 | **99.8** | 93.8 | **100** | **100** |
| **DAH-Net (Ours)** | 1024 | 4×4×3 | **100** | **94.1** | **91.0** | **99.4** | **100** |
| | 4096 | 2×2×3 | **100** | **82.1** | **83.5** | **84.2** | **93.2** |

Identity indicates the results of testing the steganography performance without any distortion

The best result is bold

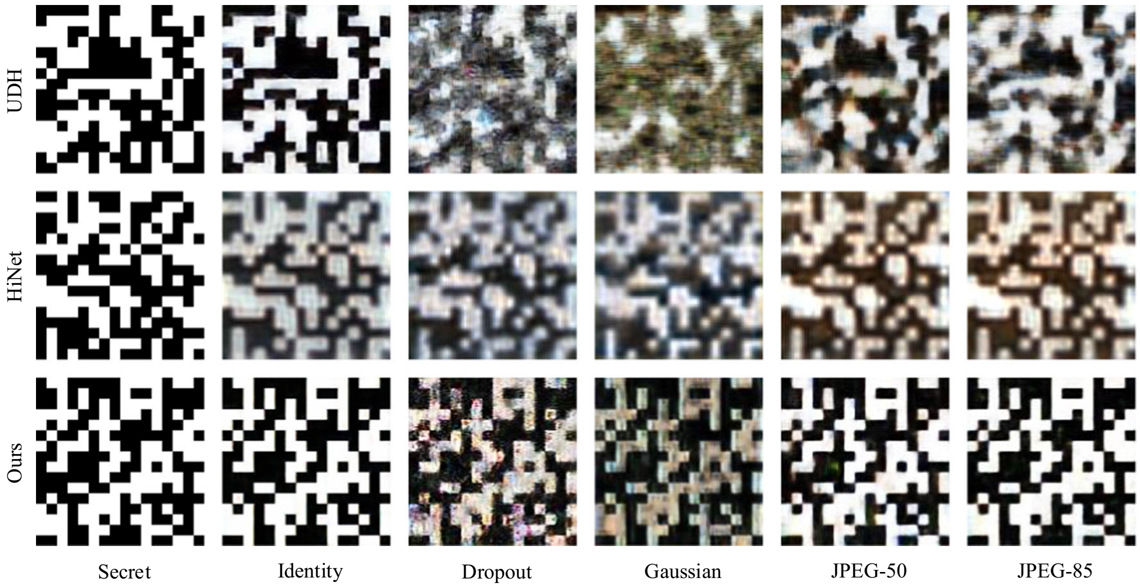


**Fig. 7** Visual comparisons for watermarking by hiding barcodes under different image distortions

**Table 3** Experiment results of photographic steganography

| | Method | APD↓ | PSNR↑ | SSIM↑ |
|---|---|---|---|---|
| C | UDH [7] | 6.93 | **30.48 dB** | 0.908 |
| | HiNet [31] | 8.50 | 27.95 dB | 0.895 |
| | **DAH-Net (Ours)** | **6.40** | 30.06 dB | **0.922** |
| S | UDH [7] | 12.71 | 23.00 dB | 0.807 |
| | HiNet [31] | 11.26 | 24.53 dB | 0.847 |
| | **DAH-Net (Ours)** | **9.38** | **25.41 dB** | **0.876** |

C, S indicate the performances of the stego and revealed secret images, respectively

The best result is bold

the stego and revealed secret images produced using GDE are 0.984 and 0.980, respectively. GDE+AFE$^S$ denotes that AFE is only used to extract the essential part of secret information. Applying the GDE+AFE$^S$ improves the APDs of cover and revealed secret images by 0.54 and 0.22, respectively. Applying the GDE+AFE$^S$ improves the cover and secret APD by 0.54 and 0.22, respectively. Furthermore, the image quality of the stego and revealed secret images are also promoted in the GDE+AFE$^S$. When the AFE is applied both on the secret and cover images, the image quality of the stego and revealed secret images are improved the best in the DAH-Net (GDE+AFE$^{C,S}$). These comparison results show the effectiveness of the proposed GDE and AFE methods in our DAH-Net.

#### 4.4.2 Adaptiveness of attentive frequency weights

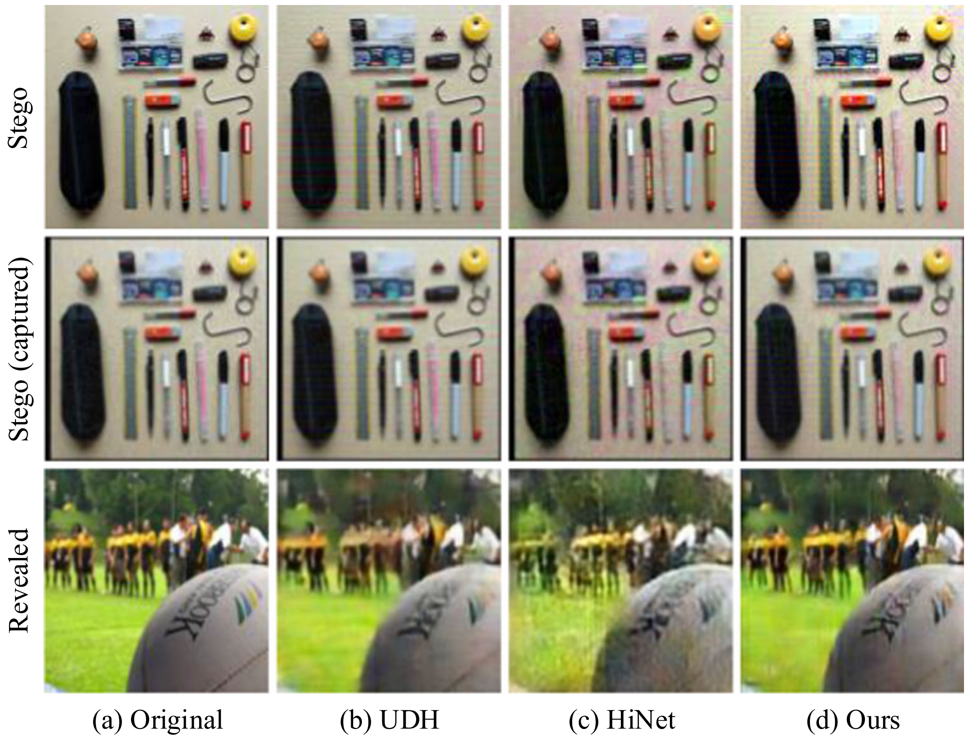
In order to further analyze the success of image hiding, we visualize the importances of attentive frequency weights in AFE. Frequency importances are calculated from every transposed convolution layer output of the cover and secret images. Figure 9a shows the different frequency importances of the cover and secret images. It is clear that frequency $f \in \{0, 1, \cdots, n-1\}$ has different importances in different depth layers $l \in \{0, 1, \cdots, L-1\}$ of cover and secret images. In addition, frequency $f$ at the same depth layers $l$ of the cover and secret images have different importances. Combined with the steganography performance of DAH-Net, it strongly verifies the adaptiveness and rationality of the attentive frequency weights produced by the proposed AFE.

Furthermore, the average frequency importances of the cover and secret images in every transposed convolution layer are also calculated and illustrated in Fig. 9b. It is obvious that frequency $f = 0$ and $f = 8$ are more important than other frequencies on the cover and secret images for image hiding. This phenomenon guides to achieving image hiding by fusing different frequencies in future. From all of the above comparisons, the attentive frequency weights in AFE adaptively and effectively extract the essential frequency information of the cover and secret images.

#### 4.4.3 Robustness of DAH-net

To analyze the robustness of our DAH-Net, different distortions, *i.e.*, Gaussian blurring, pixel-wise dropout, and JPEG compression are applied on the stego images to simulate image distortions. The Gaussian layer blurs the stego images by convolution layer with Gaussian kernel. The kernel size of Gaussian blurring is set to 3, and the

**Fig. 8** Visual comparisons for photographic steganography



(a) Original (b) UDH (c) HiNet (d) Ours

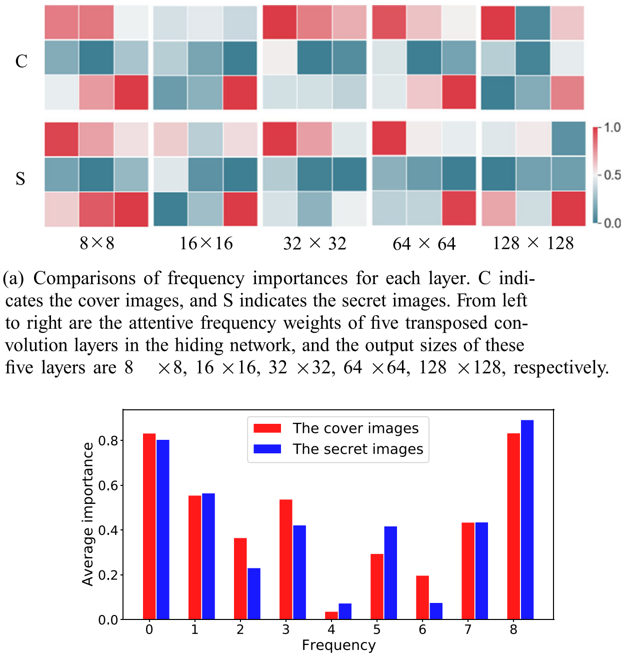
**Table 4** Effect of GDE and AFE

| | Method | APD↓ | PSNR↑ | SSIM↑ |
|---|---|---|---|---|
| C | GDE | 2.45 | 38.21 dB | 0.984 |
| | GDE+AFE$^{S}$ | 2.23 | 39.04 dB | 0.987 |
| | **DAH-Net (GDE+AFE$^{C,S}$)** | **2.05** | **39.84 dB** | **0.988** |
| S | GDE | 3.69 | 34.76 dB | 0.980 |
| | GDE+AFE$^{S}$ | 3.15 | 36.11 dB | 0.986 |
| | **DAH-Net (GDE+AFE$^{C,S}$)** | **2.71** | **37.19 dB** | **0.987** |

AFE$^{S}$Denotes that AFE is only used to extract the essential part of secret information. AFE$^{C,S}$ indicates that AFE is applied to extract the necessary cover and secret information

The best result is bold

variance is set to 2. The dropout layer randomly drops the pixels of stego images and substitutes them with the pixels of the cover images. It drops the pixels of stego images with the percentage $p$, and $p$ is set to 0.3 in our work. The JPEG layer applies JPEG compression to the stego images with quality factor $Q \in (0, 100)$. We train DAH-Net with $Q = 50$, and test the robustness with $Q = 50$ and $Q = 85$. The robustness results with different distortions are shown in Table 5. It is clear that our DAH-Net achieves much better performances under different image distortions compared to UDH and HiNet. Specifically, compared to UDH and HiNet, the proposed DAH-Net improves PSNRs of the revealed secret images by 8.41 dB and 6.12 dB



(a) Comparisons of frequency importances for each layer. C indicates the cover images, and S indicates the secret images. From left to right are the attentive frequency weights of five transposed convolution layers in the hiding network, and the output sizes of these five layers are 8 ×8, 16 ×16, 32 ×32, 64 ×64, 128 ×128, respectively.

(b) Comparisons of average frequency importances of all the layers.

**Fig. 9** Frequency importance comparisons of the cover and secret images. **a** Comparisons of frequency importances for each layer. C indicates the cover images, and S indicates the secret images. From left to right are the attentive frequency weights of five transposed convolution layers in the hiding network, and the output sizes of these five layers are 8×8, 16×16, 32×32, 64×64, 128×128, respectively. **b** Comparisons of average frequency importances of all the layers.

**Table 5** Performances of image hiding under different image distortions

| Noisy | Method | APD-S↓ | PSNR-S↑ | SSIM-S↑ |
|---|---|---|---|---|
| Dropout | UDH [7] | 27.73 | 17.66 dB | 0.552 |
| | HiNet [31] | 29.93 | 17.21 dB | **0.634** |
| | **DAH-Net (Ours)** | **24.15** | **18.37 dB** | 0.597 |
| Gaussian | UDH [7] | 37.41 | 14.73 dB | 0.385 |
| | HiNet [31] | 39.95 | 13.68 dB | **0.484** |
| | **DAH-Net (Ours)** | **36.64** | **14.92 dB** | 0.444 |
| JPEG-50 | UDH [7] | 21.10 | 19.41 dB | 0.615 |
| | HiNet [31] | 31.29 | 16.46 dB | 0.521 |
| | **DAH-Net (Ours)** | **14.35** | **22.64 dB** | **0.730** |
| JPEG-85 | UDH [7] | 33.04 | 15.34 dB | 0.447 |
| | HiNet [31] | 27.36 | 17.63 dB | 0.611 |
| | **DAH-Net (Ours)** | **12.54** | **23.75 dB** | **0.776** |

The best result is bold

under JPEG-85, respectively. To intuitively illustrate the robustness of our DAH-Net, the visualization for steganography under different image distortions are shown in in Fig. 10. It is obvious that our DAH-Net can generate the revealed secret images with much better image quality compared to the UDH and HiNet under different distortions. Consequently, these experiment results sufficiently prove that our DAH-Net is more robust than the other methods.

### 4.4.4 Passive attack analysis

In this part, we conduct StegExpose [37], SRNet [38], and ManTra-Net [39] to analyze the security of stego images generated by UDH, HiNet, and the proposed DAH-Net. These three methods are common security detection methods for image hiding. StegExpose consists of five steganalysis methods, *i.e.*, Chi-square attack [40], Primary sets [41], RS analysis [42], Sample pair analysis [43], and Fusion steganalysis. Because the attack detection ability of the Chi-square attacks and Primary sets is limited, we use the other three methods to analyze the security of the stego images. The detection results with StegExpose of UDH, HiNet, and our method are shown in Fig. 11. Area Under Curve (AUC) indicates the area under the Receiver Operating Characteristic (ROC) curves. The smaller AUC denotes the lower steganalysis performance and the better steganography performance. It is apparent that compared to UDH and HiNet, our DAH-Net can generate stego images with significant security promotion using AFE and GDE.

Furthermore, the SOTA steganalysis method SRNet is also adopted to measure the anti-steganalysis ability of our DAH-Net and the comparison methods. The detection results using SRNet are reported in Table 6. Note that the accuracy closer to 50% (random guess) indicates a higher security level. It is clear that the detection rate of the stego images generated by our DAH-Net is the smallest and the closest to 50%. In addition, following HiNet, we also conduct experiments to investigate the anti-steganalysis

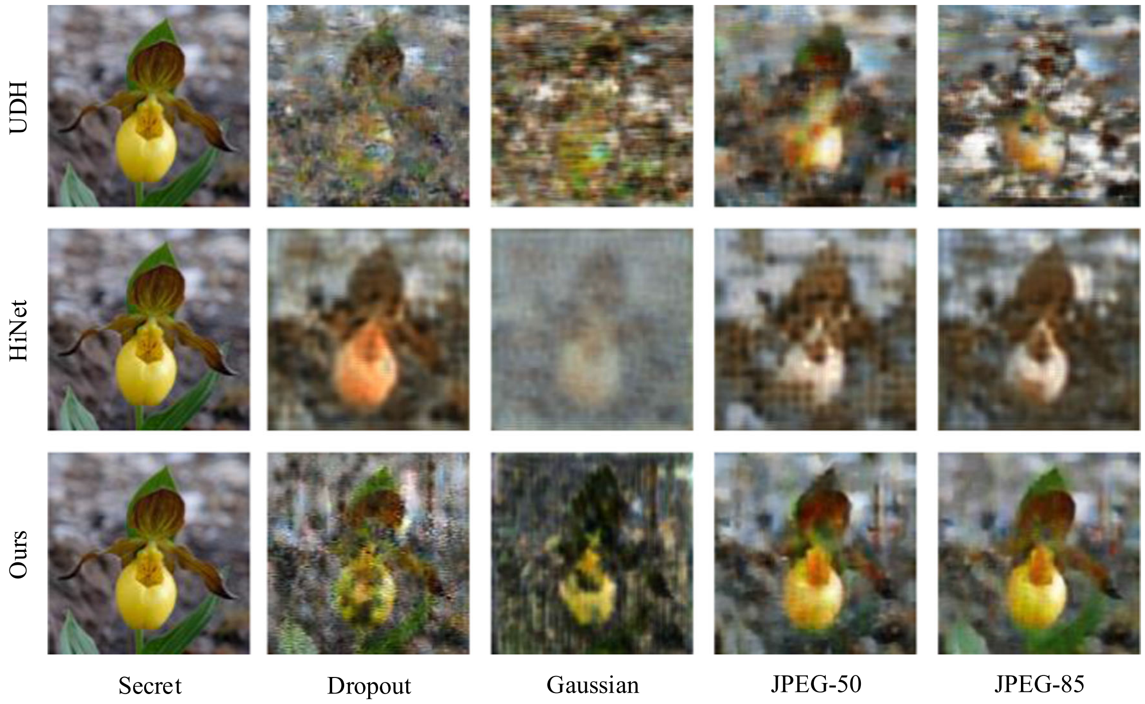


**Fig. 10** Visual comparisons for image hiding under different image distortions

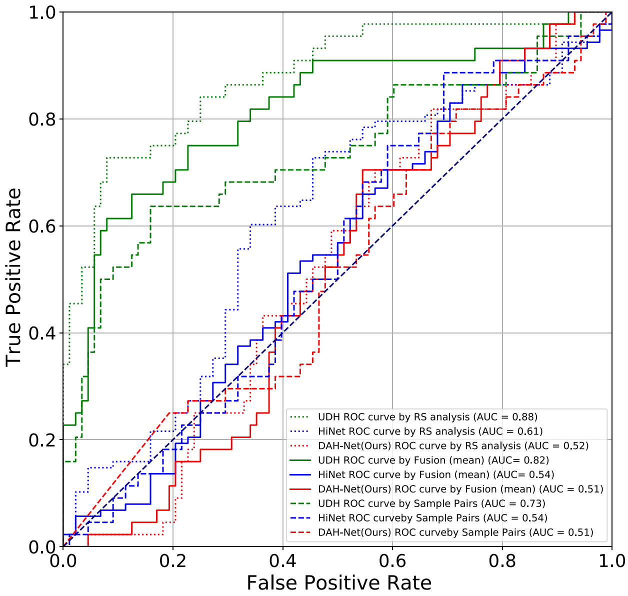


**Fig. 11** Visual comparisons of forgery detection for the stego images generated by UDH, HiNet, and our DAH-Net

**Table 6** The detection accuracy using SRNet [38]

| Methods | Accuracy (%) ±std |
|---|---|
| UDH [7] | 99.45 ± 0.34 |
| HiNet [31] | 89.25 ± 0.75 |
| **DAH-Net (Ours)** | **73.41 ± 0.19** |

The accuracy closer to 50% (random guess) indicates higher security level, and the best result is bold

performances of our DAH-Net under different numbers of training data. The results are shown in Fig. 12. Clearly, compared to UDH and HiNet, our DAH-Net can generate higher security stego images under different numbers of training data. Meanwhile, when the training data is limited, SRNet hardly distinguishes the stego images generated by our DAH-Net from cover images.

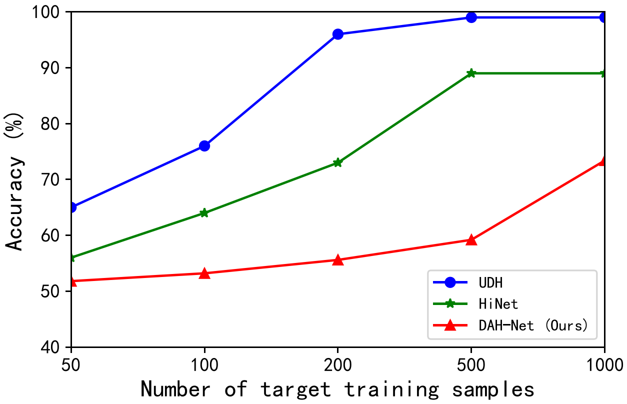


**Fig. 12** The steganalysis results using SRNet with different training samples

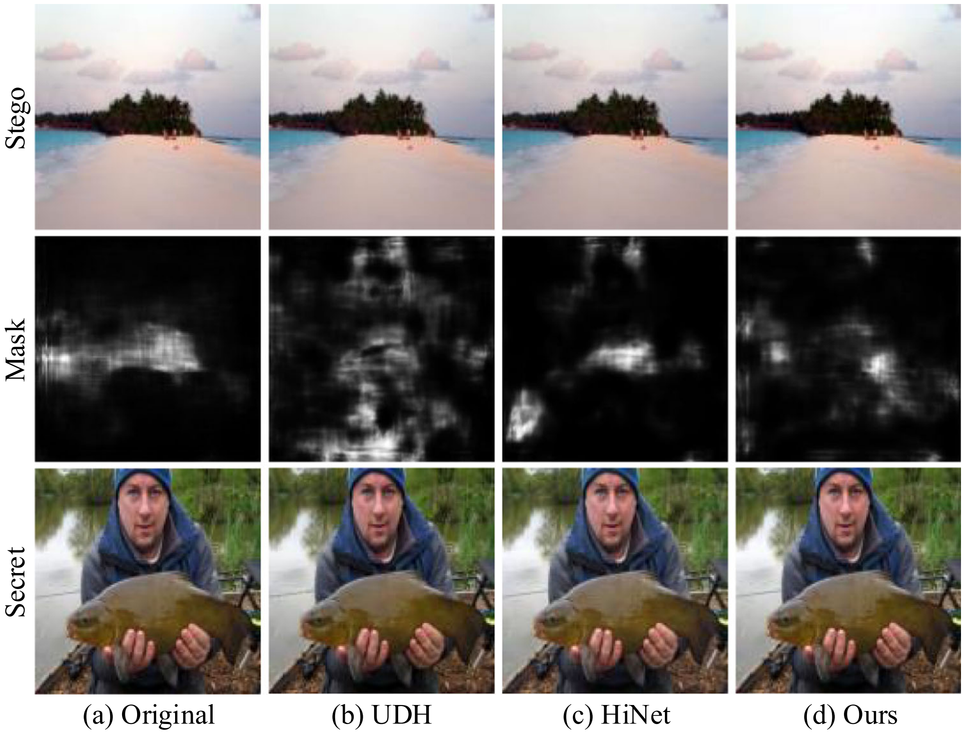
StegExpose and SRNet can only detect whether the target images contain secret information, and fail to detect the specific content of the secret image information. Therefore, we utilize ManTra-Net to further detect the manipulation region of stego images. The comparisons of detection results are shown in Fig. 13. The detection results of the cover image are also provided for comparison. The white region in the mask images represents the manipulation region detected by ManTra-Net. Apparently, the detected manipulation region of our DAH-Net has the smallest similarity with the secret image. More clear contour information of the head can be detected from the stego image obtained by UDH and HiNet, and the contour information is consistent with the edge information of the object in the secret image. This indicates the stego images generated by our DAH-Net are undetectable for steganalysis methods. These visualization results prove that the stego images generated by our DAH-Net are safer than other state-of-the-art methods.

### 4.4.5 Effect of kernel size of convolution layer

To explore the impact of kernel size of convolution layer on steganography performance, we conduct experiments with kernel sizes of 3 × 3, 4 × 4, and 5 × 5, respectively. The results are shown in Table 7. Obviously, when the kernel size of convolution layer is set to 4 × 4, the performance of steganography is the best. Specially, compared to $\kappa = 3$ and $\kappa = 5$, the PSNRs of the stego and revealed secret images have been improved by $\{2.15, 1.16\}$ dB and $\{1.42, 1.08\}$ dB using $\kappa = 4$. Considering that the kernel size of transpose convolution layer is set to 4 × 4, this may be attributed to the fact that better steganography can be achieved when the encoding and decoding parts of the hiding network have the same receptive field.

### 4.4.6 Discussion of limitations

In this paper, we mainly introduce the Attentive Frequency Extraction and Gradual Depth Extraction of the proposed DAH-Net for image hiding. Future research should explore more steganography tasks of different modals, such as hiding images within audio information. Meanwhile, although our DAH-Net significantly improves quality and security of stego images, how to further promote the performance of larger capacity image hiding with a lighter network is still a challenge.



(a) Original (b) UDH (c) HiNet (d) Ours

**Fig. 13** Passive attack analysis for the stego images generated by UDH, HiNet, and our DAH-Net

**Table 7** Effect of kernel size of convolution layer in the hiding network

| Kernel size ($\kappa \times \kappa$) | s | p | APD-C↓ | PSNR-C↑ | SSIM-C↑ | APD-S↓ | PSNR-S↑ | SSIM-S↑ |
|---|---|---|---|---|---|---|---|---|
| 3 × 3 | 2 | 1 | 2.56 | 37.78 dB | 0.983 | 3.05 | 36.15 dB | 0.983 |
| 4 × 4 | 2 | 1 | **2.02** | **39.93 dB** | **0.988** | **2.68** | **37.31 dB** | **0.987** |
| 5 × 5 | 2 | 2 | 2.39 | 38.51 dB | 0.986 | 3.05 | 36.23 dB | 0.984 |

The best result is bold

## 5 Conclusion

This paper proposes the Deep Adaptive Hiding Network (DAH-Net) for image hiding, watermarking, and photographic steganography. This is the first attempt to extract and fuse only the necessary information of the secret and cover images for image hiding. The proposed DAH-Net is mainly constructed by two proposed extractions, including Attentive Frequency and Gradual Depth Extractions. Attentive Frequency Extraction (AFE) is applied to adaptively extract the necessary information of the secret and cover images at the frequency extent. Gradual Depth Extraction (GDE) is proposed to gradually extract and fuse secret and cover information at the depth extent in the hiding network. Experiment results prove that the proposed DAH-Net can generate significantly high-quality stego and revealed secret images with satisfactory safety promotion.

**Funding** This work was supported in part by Guangdong Shenzhen joint Youth Fund under Grant 2021A151511074, in part by Shenzhen Key Technical Project under Grant 2022N063, in part by National Key Research and Development Program of China under Project Number 2018AAA0100100, in part by the NSFC funds 62206073 and 62176077, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019Bl515120055, in part by the Shenzhen Key Technical Project under Grant 2022N001, 2020N046, in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20210324132210025, and in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

**Data availability** The data generated and analyzed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

## References

1. You W, Zhang H, Zhao X (2020) A siamese CNN for image steganalysis. IEEE Trans Inf Forensics Secur 16:291–306

2. Zhu J, Kaplan R, Johnson J, Fei-Fei L (2018) Hidden: Hiding data with deep networks. In: Proceedings of the European conference on computer vision, pp 657–672
3. Yin Z, Peng Y, Xiang Y (2020) Reversible data hiding in encrypted images based on pixel prediction and bit-plane compression. IEEE Trans Depend Secure Computing 19(2):992–1002
4. Chen F, Yuan Y, He H, Tian M, Tai H-M (2020) Multi-msb compression based reversible data hiding scheme in encrypted images. IEEE Trans Circuits Syst Video Technol 31(3):905–916
5. Thodi DM, Rodríguez JJ (2007) Expansion embedding techniques for reversible watermarking. IEEE Trans Image Process 16(3):721–730
6. Chen B, Lu W, Huang J, Weng J, Zhou Y (2020) Secret sharing based reversible data hiding in encrypted images with multiple data-hiders. IEEE Trans Depend Secure Computing 19(2):978–991
7. Zhang C, Benz P, Karjauv A, Sun G, Kweon IS (2020) Udh: universal deep hiding for steganography, watermarking, and light field messaging. Adv Neural Inf Process Syst 33:10223–10234
8. Lu Y, Lu G, Li J, Zhang Z, Xu Y (2021) Fully shared convolutional neural networks. Neural Comput Appl 33(14):8635–8648
9. Li Y, Zhang Z, Chen B, Lu G, Zhang D (2022) Deep margin-sensitive representation learning for cross-domain facial expression recognition. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2022.3141604
10. Lu Y, Lu G, Xu Y, Zhang B (2018) Aar-cnns: auto adaptive regularized convolutional neural networks. In: International joint conference on artificial intelligence, pp 2511–2517
11. Lu Y, Lu G, Li J, Xu Y, Zhang D (2020) High-parameter-efficiency convolutional neural networks. Neural Comput Appl 32(14):10633–10644
12. Xie Q, Zhang P, Yu B, Choi J (2021) Semisupervised training of deep generative models for high-dimensional anomaly detection. IEEE Trans Neural Networks Learn System 33(6):2444–2453
13. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK (2020) Improved inception-residual convolutional neural network for object recognition. Neural Comput Appl 32(1):279–293
14. Ignatov A, Byeoung-su K, Timofte R, Pouget A (2021) Fast camera image denoising on mobile gpus with deep learning, mobile ai 2021 challenge: Report. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2515–2524
15. Feng X, Pei W, Jia Z, Chen F, Zhang D, Lu G (2021) Deep-masking generative network: a unified framework for background restoration from superimposed images. IEEE Trans Image Process 30:4867–4882
16. Zhang C, Hu W, Jin T, Mei Z (2018) Nonlocal image denoising via adaptive tensor nuclear norm minimization. Neural Comput Appl 29(1):3–19
17. Das PK, Meher S, Panda R, Abraham A (2021) An efficient blood-cell segmentation for the detection of hematological disorders. IEEE Trans Cybern
18. Abdel-Basset M, Chang V, Mohamed R (2021) A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems. Neural Comput Appl 33(17):10685–10718
19. Brempong EA, Kornblith S, Chen T, Parmar N, Minderer M, Norouzi M (2022) Denoising pretraining for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4175–4186
20. Baluja S (2017) Hiding images in plain sight: deep steganography. Adv Neural Inf Process Syst 30:2069–2079
21. Baluja S (2019) Hiding images within images. IEEE Trans Pattern Anal Mach Intell 42(7):1685–1697
22. Thai TH, Cogranne R, Retraint F (2014) Statistical model of quantized dct coefficients: application in the steganalysis of jsteg algorithm. IEEE Trans Image Process 23(5):1980–1993
23. Denemark T, Fridrich J (2015) Side-informed steganography with additive distortion. In: 2015 IEEE international workshop on information forensics and security (WIFS), pp 1–6. IEEE
24. Li B, Tan S, Wang M, Huang J (2014) Investigation on cost assignment in spatial image steganography. IEEE Trans Inf Forensics Secur 9(8):1264–1277
25. Bandyopadhyay SK, Bhattacharyya D, Ganguly D, Mukherjee S, Das P (2008) A tutorial review on steganography. In: International conference on contemporary computing, vol 101, pp 105–114
26. Wu H-T, Cheung Y-M, Zhuang Z, Xu L, Hu J (2022) Lossless data hiding in encrypted images compatible with homomorphic processing. IEEE Trans Cybernetics. https://doi.org/10.1109/TCYB.2022.3163245
27. Pevnỳ T, Filler T, Bas P (2010) Using high-dimensional image models to perform highly undetectable steganography. International workshop on information hiding. Springer, Berlin, pp 161–177
28. Hayes J, Danezis G (2017) Generating steganographic images via adversarial training. arXiv preprint arXiv:1703.00371
29. Lu S-P, Wang R, Zhong T, Rosin PL (2021) Large-capacity image steganography based on invertible neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10816–10825
30. Xiao M, Zheng S, Liu C, Wang Y, He D, Ke G, Bian J, Lin Z, Liu T-Y (2020) Invertible image rescaling. In: Proceedings of the European conference on computer vision. Springer, Berlin. pp 126–144.
31. Jing J, Deng X, Xu M, Wang J, Guan Z (2021) Hinet: deep image hiding by invertible network. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4733–4742
32. Gonzalez RC, Woods RE et al (2002) Digit Image Process. Prentice hall, Upper Saddle River
33. Ulicny M, Krylov VA, Dahyot R (2020) Harmonic convolutional networks based on discrete cosine transform. arXiv preprint arXiv:2001.06570
34. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232
35. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. IEEE
36. Wengrowski E, Dana K (2019) Light field messaging with deep photographic steganography. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1515–1524
37. Boehm B (2014) Stegexpose-a tool for detecting LSB steganography. arXiv preprint arXiv:1410.6656
38. Boroumand M, Chen M, Fridrich J (2018) Deep residual network for steganalysis of digital images. IEEE Trans Inf Forensics Secur 14(5):1181–1193
39. Wu Y, AbdAlmageed W, Natarajan P (2019) Mantra-net: manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9543–9552
40. Westfeld A, Pfitzmann A (1999) Attacks on steganographic systems. International workshop on information hiding. Springer, Berlin, pp 61–76
41. Dumitrescu S, Wu X, Memon N (2002) On steganalysis of random lsb embedding in continuous-tone images. In: Proceedings

of the international conference on image processing, vol 3, pp 641–644. IEEE

42. Fridrich J, Goljan M, Du R (2001) Reliable detection of lsb steganography in color and grayscale images. In: Proceedings of the 2001 workshop on multimedia and security: new challenges, pp 27–30
43. Dumitrescu S, Wu X, Wang Z (2002) Detection of lsb steganography via sample pair analysis. International workshop on information hiding. Springer, Berlin, pp 355–372

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.