Traitement du Signal
Vol. 41, No. 3, June, 2024, pp. 1451-1460
Journal homepage: http://iieta.org/journals/ts

Optimizing Image Recognition Algorithms with Differential Privacy Integration

Shaoyu Yang

College of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China

Corresponding Author Email: ysy@ncwu.edu.cn

https://doi.org/10.18280/ts.410333

Received: 2 January 2024
Revised: 15 May 2024
Accepted: 26 May 2024
Available online: 26 June 2024

Keywords:
differential privacy, image recognition, vision transformer, algorithm fusion, privacy protection

ABSTRACT

With the rapid advancement of artificial intelligence technology, image recognition has become a core task in the field of computer vision and is widely applied across various industries. Image recognition technology significantly improves work efficiency and decision accuracy through the automatic analysis and processing of image data. However, image data often contain a large amount of sensitive information, making privacy protection a crucial issue in the application of image recognition technology. Existing differential privacy techniques effectively prevent the leakage of sensitive information by introducing noise into data processing. However, when applied to image recognition, these techniques often lead to a decline in recognition performance. Additionally, current integration methods lack effective evaluation of prediction accuracy and stability when handling predictions from multiple models, affecting the reliability and accuracy of the final recognition results. This paper proposes a vision Transformer network model with differential privacy protection and designs an image recognition algorithm that integrates differential privacy. By incorporating differential privacy mechanisms, we aim to enhance image recognition performance while safeguarding privacy. Furthermore, we introduce an adaptive weighting method to fuse predictions from different models, further improving recognition accuracy and stability. Our research not only provides a novel solution for privacy protection in image recognition but also theoretically and practically verifies the feasibility and effectiveness of differential privacy techniques in real-world applications. This study holds significant academic and practical value.

1. INTRODUCTION

With the rapid development of artificial intelligence technology, image recognition has become one of the core tasks in the field of computer vision [1, 2]. Image recognition technology, widely applied in various industries, significantly improves work efficiency and decision accuracy through the automatic analysis and processing of image data [3-5]. However, image data often contains a large amount of sensitive information. How to improve the performance of image recognition while protecting privacy has become a hot and difficult issue in current research.

Privacy protection in image recognition has important research significance [6-9]. On the one hand, protecting personal privacy is a basic ethical requirement in data processing and application, and any privacy leakage may lead to serious legal and social consequences [10, 11]. On the other hand, differential privacy technology provides a solid theoretical foundation for privacy protection. By introducing noise in data processing, it can effectively prevent the leakage of sensitive information [12]. In this context, studying how to apply differential privacy technology to the field of image recognition, ensuring data privacy while improving recognition performance, has important practical significance and research value.

However, the current differential privacy protection methods often lead to a decline in recognition performance when applied to image recognition [13-16]. The prediction ability and accuracy of traditional image recognition models are often affected after the introduction of differential privacy mechanisms. In addition, the existing fusion methods lack effective evaluation of prediction accuracy and stability when handling the prediction results of multiple models, leading to insufficient reliability and accuracy of the final recognition results [17-20]. Therefore, an innovative algorithm is needed to optimize the overall performance of image recognition while protecting privacy.

This study mainly includes two parts: first, we propose a vision Transformer network model under differential privacy protection. By integrating the differential privacy mechanism, we aim to improve image recognition performance while protecting privacy. Second, we design an image recognition algorithm for integrating differential privacy. By adaptively weighting and fusing the prediction results of different models, we further improve the accuracy and stability of recognition. This study not only provides a new solution for privacy protection in image recognition but also verifies the feasibility and effectiveness of differential privacy technology in practical applications in theory and practice, which has important academic and practical value.

2. VISION TRANSFORMER NETWORK MODEL WITH DIFFERENTIAL PRIVACY PROTECTION

The application scenarios of differential privacy in image recognition algorithms are extremely extensive, especially in protecting user privacy and data security. In medical image analysis, differential privacy can ensure that the analysis results of medical images do not reveal any single patient's information, thereby protecting patient privacy. In social media and security monitoring, facial recognition technology is widely used, and differential privacy can protect users' biometric data, preventing unauthorized access and misuse. In the field of autonomous driving, differential privacy technology can protect the image data analyzed by autonomous vehicles regarding pedestrians, vehicles, and other objects on the road, preventing the disclosure of the identities of pedestrians and other road users. Smart home devices, such as smart cameras, can use differential privacy to protect the privacy of household members when capturing and processing image data, preventing external hackers or improper users from obtaining sensitive information. Social media platforms use image recognition technology to analyze and recommend content, and differential privacy can ensure that the image data uploaded by users will not be leaked or misused. On online education platforms, teachers use image recognition technology to analyze students' submitted assignments or exam answers, and differential privacy can protect students' data privacy, preventing their work results from being improperly used or disclosed. By applying differential privacy technology in these scenarios, the privacy protection capability of image recognition systems can be effectively enhanced while maintaining the efficiency and accuracy of the algorithms.

To achieve the optimization of image recognition with differential privacy integration, this paper proposes a vision Transformer network model with differential privacy protection and introduces the attention mechanism. In the proposed model, the introduced attention mechanism is not only used to capture and extract local and global features in image data but also needs to meet the requirements of privacy protection. Specifically, the differential privacy mechanism can introduce noise during the attention weight calculation phase to prevent the model from leaking sensitive information when processing image data. By adding noise to the weight calculations of each attention head, the contribution of a single image or its specific features will not significantly affect the overall result, thus protecting individual privacy. At the same time, the threshold attention mechanism adjusts between focusing on local features and recovering global features through learnable threshold parameters, which can also undergo differential privacy processing to ensure that the model does not leak specific information of sensitive data during the learning process.

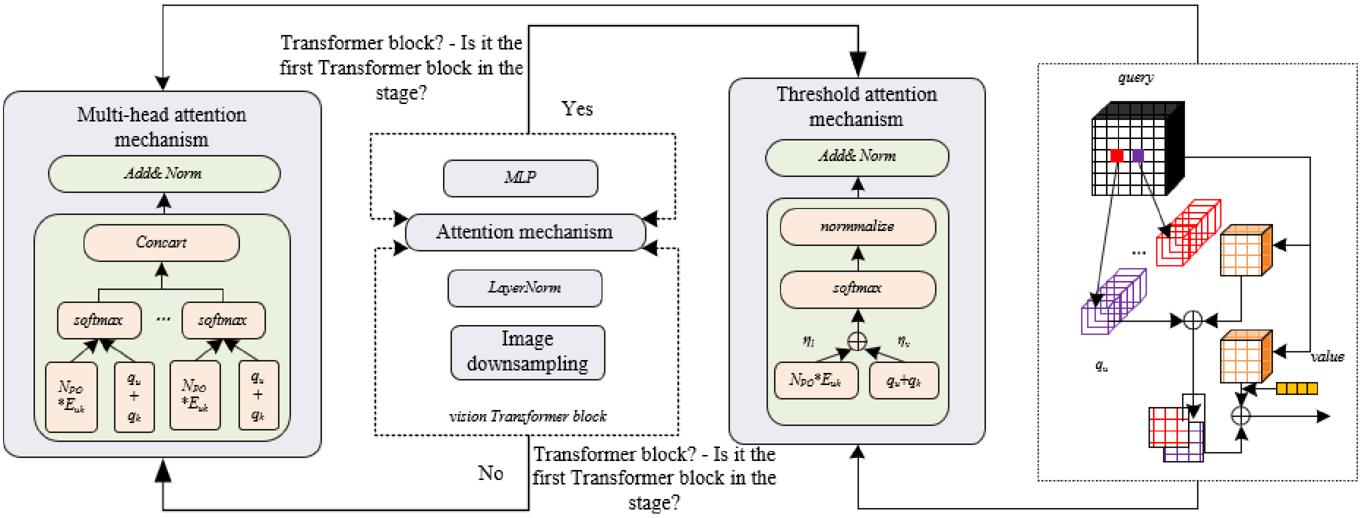


Figure 1. Hierarchical attention transformation network structure diagram

In the model, the hierarchical attention transformation mechanism and normalization play a crucial role in maintaining model performance while ensuring data privacy protection. Figure 1 shows the hierarchical attention transformation network structure diagram. The hierarchical attention transformation mechanism gradually refines feature extraction through multi-level attention calculations, thereby capturing local and global features of the image at different levels. Each layer's attention mechanism will combine differential privacy technology, introducing noise when calculating attention weights to ensure that no specific image information is leaked during feature extraction. Specifically, in each layer's attention head, the weight calculation formula will include a privacy-protecting noise term, which is appropriately adjusted according to the requirements of differential privacy to ensure the reasonable use of the privacy budget. Specifically, let the weight matrix be denoted by L, the trainable embedding vectors be denoted by N_{POS}, and the fixed relative position encoding be denoted by E_{uk}. The formulas of the hierarchical attention transformation mechanism network are as follows:

$$LGTX(A) = \text{concat}\left(\left(\text{softmax}\left(\frac{WJ^S}{\sqrt{F_g}}\right)N\right)_{1\ldots v}\right)AL \tag{1}$$

$$STX(A) = \text{normalize}\left[\begin{array}{l}\eta_v\left(\text{softmax}\left(W*J^S\right)\right)\\ +\eta_l\left(\text{softmax}\left(N_{POS}^S*E_{uk}\right)\right)\end{array}\right]AL \tag{2}$$

$$GXS(A) = STAGE_{BL1}(STX(A)) + STAGE_{BL(v-1)}(LGTX(A)) \tag{3}$$

Normalization in the vision Transformer is used to standardize the output of each layer to maintain training stability and accelerate convergence. Under differential privacy protection, the normalization operation needs to be specially designed to prevent potential privacy leaks during the training process. Specifically, when normalizing the output of each layer, differential privacy noise is introduced in the process of calculating the mean and standard deviation so that the normalized result does not directly reflect the characteristics of any single input data. This normalization process also follows the differential privacy mechanism to ensure that the model protects sensitive information of input data during training. Specifically, let each sample be denoted by $A=(a_1,a_2,...,a_G)$, the calculated mean and standard deviation be denoted by ω and δ, different attention mechanisms such as multi-head attention mechanisms or threshold attention mechanisms be denoted by X, and the features extracted by the u-th layer of the Transformer encoder be denoted by c_u and c'_u. Let the input image be denoted by $G\times Q\times Z$, where the height, width, and number of channels of the channel are denoted by G, Q, and Z, respectively. The bias and gain obtained from the normalized vector are denoted by y and h, respectively, then:

$$c_0 = \left\lfloor a_o^1 R; a_o^2 R; \ldots; a_o^v R \right\rfloor + R_{POS} \tag{4}$$

$$c_u = X(MV(c_{u-1})), MV = h\Phi\left(\frac{a-\omega}{\delta}\right) + y \tag{5}$$

$$c'_u = STAGE_G(X_{LGTX/STX}(c_{u-1})), u = 1...M \tag{6}$$

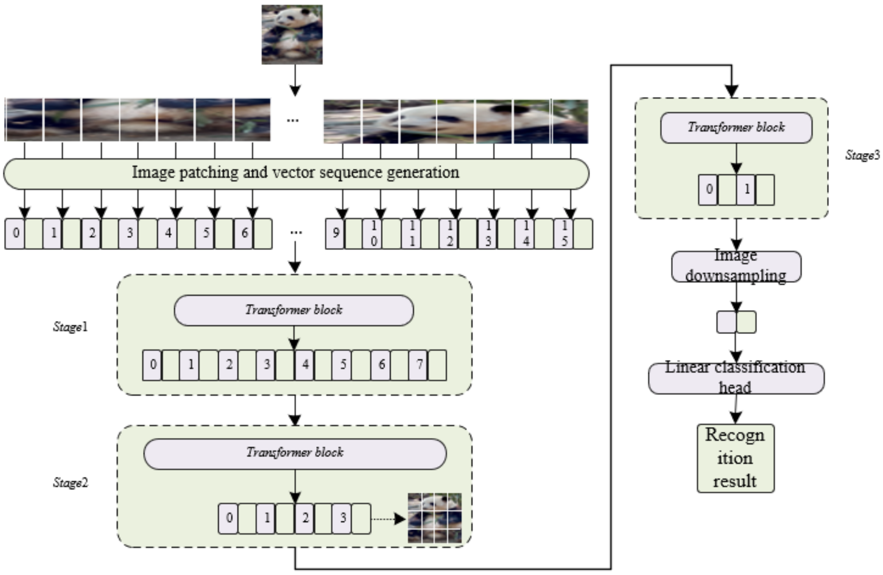


Figure 2. Hierarchical attention transformation mechanism-based vision transformer network structure diagram

Figure 2 presents the vision Transformer network structure diagram based on the hierarchical attention transformation mechanism. In the proposed vision Transformer network with differential privacy protection based on the hierarchical attention transformation mechanism, differential privacy technology is integrated to ensure data privacy protection in image recognition tasks. The input image size of the model is $G\times Q\times Z$. By dividing the image into non-overlapping patches of 14×14 pixels, the processed patch sequence is input into multiple structural blocks of the same dimension. The entire network is divided into L stages, with each stage containing multiple Transformer *blocks*. A local attention module is introduced at the beginning of each stage to capture local features, while the remaining blocks adopt a global attention mechanism to recover the perception of global features. This hierarchical attention transformation mechanism effectively balances the extraction of local and global features. To ensure differential privacy, the attention calculations and normalization operations in the model are specially designed. In the local and global attention modules, differential privacy noise is introduced into the calculation of attention weights. Specifically, when calculating the attention scores, a noise term is added after the *softmax* operation in the formula to protect the privacy of the input data. The normalization operation in each Transformer *block* also follows the differential privacy mechanism, introducing noise when calculating the mean and standard deviation to ensure that no sensitive information is leaked. Structurally, the model adopts a new positional encoding method to replace the traditional classification token. This method gradually shortens the sequence length through one-dimensional hierarchical max pooling, constructing hierarchical representations, reducing redundant information, and computational costs. The introduction of the max pooling layer not only improves the computational efficiency of the model but also enhances the robustness of feature extraction under the framework of differential privacy. The final classification of the model is computed through the *GELD* activation function and the cross-entropy loss function, and then it enters the *MLP* head for classification. Prediction and training are carried out without layer normalization, improving the model's accuracy. Through this design, the model not only performs excellently in image recognition tasks but also strictly adheres to the principles of differential privacy during training and inference, ensuring the security and privacy of user data.

Below, this paper explains the basic principles of image recognition using the vision Transformer network based on the hierarchical attention transformation mechanism under differential privacy protection. Specifically, the model divides the input image data into non-overlapping patches of 14×14 pixels and inputs these patch sequences into multiple Transformer blocks, achieving efficient extraction of local and global features. A local attention module is introduced at the beginning of each stage, while the remaining blocks adopt a global attention mechanism to balance the extraction of local and global features. To ensure privacy protection, a differential privacy mechanism is introduced into the model. Specifically, in calculating attention weights, Gaussian noise is added after the *softmax* operation to protect the privacy of the input data. Additionally, the normalization operations in the network also introduce differential privacy noise to ensure that no sensitive information is leaked when calculating the mean and standard deviation. These designs ensure that the model can strictly adhere to the principles of differential privacy when processing medical data. A convolution module *CONV(a)* is added to the model to further optimize feature extraction capabilities. Combining differential privacy optimization algorithms, data privacy is protected through gradient clipping and noise injection mechanisms. Specific steps include: first performing convolution operations on the network parameters *a* and sampling Gaussian noise *NOISE*. Setting the gradient clipping threshold *C* and noise level *δ*, then calculating the function of the aggregated gradient *h*. In each gradient update, the loss function *LOSS* is evaluated by the optimization algorithm *OPTIM*, and the parameter *P(a)* is iteratively updated, finally obtaining the classification detection result *T*.

$$LOSS = -\frac{1}{V}\sum_{u=1}^{V} \begin{matrix} b_u \cdot \log\left(o\left(b_u\right)\right) \\ +\left(1-b_u\right)\cdot\log\left(1-o\left(b_u\right)\right) \end{matrix} \tag{7}$$

In the following formula, o=[$o(b_0)$...$o(b_u)$] represents the probability vector of true labels, *LOSS* represents the loss function, $o(b_u)$ represents the probability of the true label, and 1-$o(b_u)$ represents the probability of the predicted label. Under this setting, the image data is processed and classified in the model, providing efficient and accurate classification detection results while ensuring data privacy.

$$T = P\left(h\begin{bmatrix} CONV\left(a\right)+NuS\left(a\right) \\ +NOISE\left(0,\delta^2 Z^2 U\right) \end{bmatrix} + OPTIM\left(LOSS\right)\right) \tag{8}$$

3. IMAGE RECOGNITION ALGORITHM WITH DIFFERENTIAL PRIVACY INTEGRATION

In image recognition under differential privacy protection, traditional methods typically achieve privacy protection by adding Gaussian noise to the model training parameters. However, these methods have several significant issues. First, the gradient clipping value *Z* needs to be manually input, making it difficult to accurately determine a suitable clipping value, thus reasonably clipping the gradient tensor. Second, as the training steps increase, the gradient norm gradually decreases, and a fixed *Z* value clipping strategy will lead to gradient information distortion, ultimately affecting the classification performance of the model. Additionally, as the privacy budget is consumed, the injected noise will gradually increase, severely impacting the training parameters and leading to a decline in model classification performance. To address these issues, this paper proposes a novel image recognition algorithm integrating differential privacy, aiming to optimize the training process of the vision Transformer model under differential privacy protection. Therefore, the goal of this study is to solve the problems of the difficulty in setting the gradient clipping value *Z*, gradient information distortion, and the excessive impact of noise on the model in traditional methods by improving the differential privacy mechanism. The specific implementation steps of the method will be described in sections below.

3.1 Hierarchical gradient clipping

In image recognition under the background of differential privacy, it is necessary to protect the privacy of sensitive data while ensuring the high accuracy and reliability of the model. The traditional global gradient clipping method uses a fixed clipping value *Z* to process the gradient, but this approach has significant shortcomings. Specifically, the *L*2 norm of the gradient decreases with the increase of training step *s* and gradually approaches zero. Additionally, the *L*2 norm of each layer of the global gradient tensor h_u is different. Therefore, choosing a fixed *Z* value for global gradient clipping is unreasonable because a fixed *Z* value cannot adaptively clip the gradients with gradually decreasing norms, nor can it provide appropriate clipping according to the different gradient norms of each layer. This fixed strategy will lead to gradient information distortion, severely affecting the performance of the model.

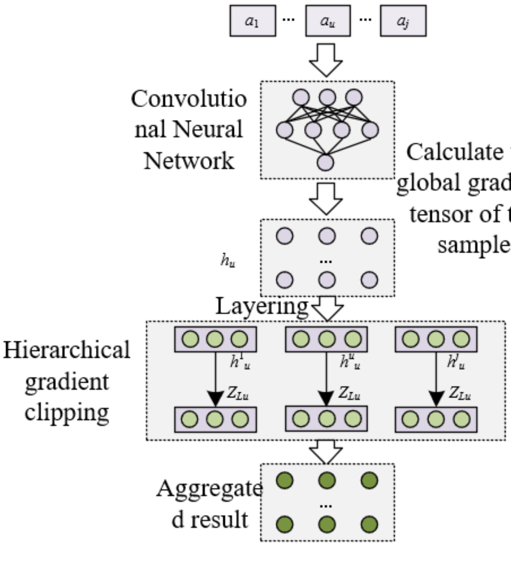

a_1
···
a_u
···
a_j
Convolutional Neural Network
Calculate the global gradient tensor of the sample
h_u
...
Layering
Hierarchical gradient clipping
h^1_u
h^u_u
h^j_u
Z_{Lu}
Z_{Lu}
Z_{Lu}
Aggregated result
...

Figure 3. Schematic diagram of hierarchical gradient clipping scheme

Under differential privacy protection, network models will accumulate privacy loss during the backpropagation process, further leading to a decline in classification performance. To address these challenges, this paper proposes an adaptive hierarchical gradient clipping scheme, as shown in Figure 3. This scheme can adaptively adjust the clipping value based on the L2 norm of the gradients at each layer, ensuring that the gradient clipping process is more reasonable and efficient. Specifically, through hierarchical clipping, the clipping value *Z* for each layer's gradient can be adjusted according to its actual $L2$ norm, avoiding the inapplicability and information distortion problems caused by a fixed clipping value.

First, the global gradient tensor h_u of each sample a_u is divided into j layers according to the model structure, forming a gradient set $\{h^1{}_u,\ldots,h^j{}_u\}$. Next, the $L2$ norm of these layered gradient tensors is calculated, and these $L2$ norms are arranged in ascending order to obtain the $L2$ norm set $T=\{\|h^1{}_u\|_2,\ldots,\|h^j{}_u\|_2\}$, where $\|h^j{}_u\|_2\geq\|h^{j-1}{}_u\|_2\geq\ldots\geq\|h^1{}_u\|_2$. In this way, the size of each layer's gradient can be intuitively compared and analyzed, providing a basis for the subsequent clipping operation. Specifically, the median of the $L2$ norm set T is taken as the clipping value, denoted by Z_{Lu}, and its calculation formula is:

$$Z_{Lu}=\frac{1}{2}\left(\left\|h_u^{\frac{j}{2}}\right\|_2+\left\|h_u^{\frac{j}{2}+1}\right\|_2\right) \tag{9}$$

Furthermore, the Z_{Lu} is used to clip each layered gradient tensor $h^j{}_u$. Through the clipping operation, it can be ensured that the $L2$ norm of each layer's gradient tensor does not exceed Z_{Lu}, effectively controlling the magnitude of the gradient and preventing gradient explosion. The calculation formula is:

$$\bar{h}_s^j(a_u)=h_s^j(a_u)\Big/ MAX\left(1,\frac{\left\|h_s^j(a_u)\right\|_2}{Z_{Lu}}\right) \tag{10}$$

After clipping, the layered gradients $h^{-j}{}_u(a_u)$ are re-aggregated into the global gradient $h^{-}{}_s(a_u)$, ensuring that the model can still maintain effective gradient information transmission while protecting privacy, which helps to improve the model's classification performance.

3.2 Noise addition

To further protect the privacy of image data, we introduced a noise addition scheme integrating differential privacy protection. This scheme, based on the proposed algorithm, adaptively adjusts the sensitivity of the noise to reduce its impact on the model's classification performance. In this scheme, privacy is typically protected by adding Gaussian noise to the aggregated gradient tensor, preventing sensitive information in the model training parameters from being easily reverse-engineered. In standard differential privacy stochastic gradient descent methods, a fixed clipping threshold Z is used as the sensitivity for noise addition. Assuming that the batch of samples is denoted by M, the noise multiplier by δ, the unit matrix by U, and the average noise added to the aggregated gradient tensor by V, the noise addition formula is:

$$\bar{h}_{s\_NO}=\frac{1}{M}\sum\nolimits_u\left(\bar{h}_s(a_u)+\bar{V}\left(0,\delta^2Z^2U\right)\right) \tag{11}$$

However, as the training steps increase, the gradient norm tends to gradually decrease, leading to an increasing consumption of the privacy budget during the model training process, thus requiring larger and larger noise additions. This increase in noise can adversely affect the model's classification performance. To address this issue, the noise addition strategy can be dynamically adjusted using the characteristics of the gradient norm. Specifically, the dynamic gradient norm Z_{Lu} in each training step can replace the fixed clipping threshold Z as the sensitivity. The advantage of this method is that Z_{Lu} will gradually decrease with the increase of training steps, thereby reducing the negative impact of noise on the model's classification performance. However, to prevent Z_{Lu} from converging to 0 and resulting in no noise being added to the gradient, thereby losing the effect of differential privacy protection, a minimal boundary parameter α needs to be set on Z_{Lu}. This parameter ensures that even when the gradient norm is very small, noise will still be added to the gradient, maintaining the continuous effectiveness of privacy protection.

$$\bar{Z}_u=Z_{Lu}+\alpha \tag{12}$$

In specific operations, the dynamic gradient norm Z_{Lu} is first calculated in each training step, and then a boundary parameter α is added to form a new sensitivity value. Next, this dynamically adjusted sensitivity value is used to calculate the Gaussian noise to be added and then added to the aggregated gradient tensor. Through this method, the effect of differential privacy protection is maintained while reducing the negative impact of noise on model performance, thereby improving the classification accuracy of the image recognition model. Assuming that the average noise added to the aggregated gradient after changing the sensitivity is denoted by V^{-}, the noise addition formula with $Z^{-}{}_u$ as the sensitivity is as follows:

$$\bar{h}_{s\_NO}=\frac{1}{M}\sum\nolimits_u\left(\bar{h}_s(a_u)+\bar{V}\left(0,\delta^2\bar{Z}_u^2U\right)\right) \tag{13}$$

3.3 Adaptive weighted fusion module

Under the differential privacy protection mechanism, noise needs to be added in each update of the model to protect data privacy. This noise addition inevitably affects the model's accuracy, especially in multi-layer network structures where the cumulative effect of noise becomes more significant, leading to a decline in model performance. Secondly, image recognition tasks typically require high accuracy, and directly increasing the number of network layers cannot effectively improve the model's performance under differential privacy protection. Instead, it can reduce the model's classification capability due to excessive noise interference. Therefore, to solve this problem and further improve the model's classification capability under differential privacy protection, we designed an adaptive weighted fusion module. Figure 4 shows the architecture of the adaptive weighted fusion module.

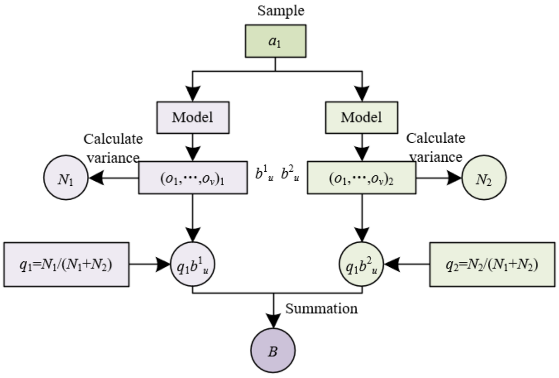


Figure 4. Adaptive weighted fusion module

Specifically, the sample image is input into the adaptive weighted fusion module. The sample image is respectively passed into two identical models with differential privacy protection for processing. These two models will generate prediction tensors, which consist of multiple prediction probabilities, denoted as $b^{1}{}_{u}$ and $b^{2}{}_{u}$. Since each prediction tensor contains the probabilities o_y for multiple prediction categories, the variance N_1 and N_2 of these prediction probabilities can be calculated. The variance reflects the degree of dispersion of the prediction probabilities, that is, the accuracy of the prediction results. The greater the degree of dispersion, the greater the variance, indicating that the model's prediction for that category is more certain and accurate. Assuming the module input sample is a_u, and v represents the number of prediction categories, the calculation formulas are:

$$\bar{o}=\frac{1}{v}\sum\nolimits_{y=1}^{v}o_y \tag{14}$$

$$N=\frac{1}{v}\sum\nolimits_{y=1}^{v}\left(o_y-\bar{o}\right)^2 \tag{15}$$

Since the more accurate the prediction result, the greater the dispersion of the prediction probabilities, thus the greater the variance. Therefore, by comparing the variances of the prediction tensors, we can determine which model's prediction is more accurate for that sample. When fusing the prediction results, the more accurate prediction tensor will be assigned a greater weight, thereby increasing the accuracy of the final fusion result.

Furthermore, weights are assigned to each prediction tensor according to the size of the prediction tensor variance. Specifically, the prediction tensor with greater variance will receive a higher weight. This is because a prediction tensor with greater variance usually represents a more certain prediction result by the model. Therefore, by giving these tensors higher weights, the overall accuracy of the fused predictions will be improved. Finally, by summing the weighted prediction tensors, a comprehensive prediction result is obtained. Assuming the assigned weight is q_u:

$$q_1=\frac{N_1}{N_1+N_2},q_2=\frac{N_2}{N_1+N_2} \tag{16}$$

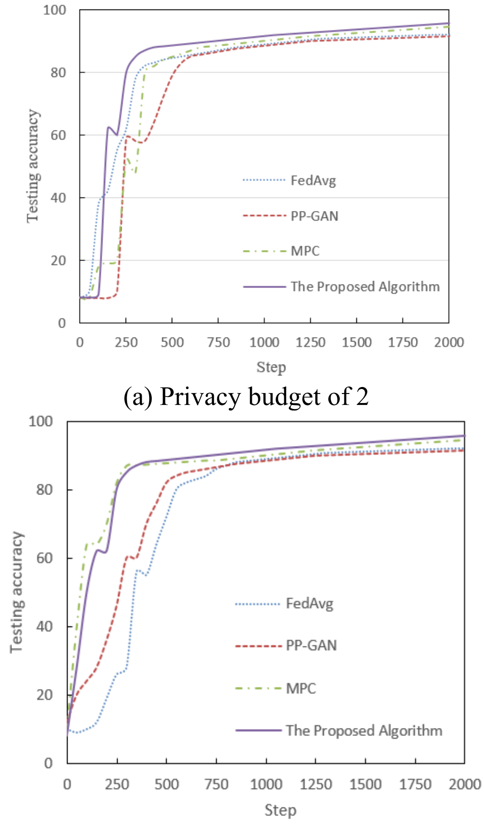
Further, the calculation formula for obtaining the fused prediction tensor B by weighted summation of the prediction tensors is:

$$B=q_1b_u^1+q_2b_u^2 \tag{17}$$

Finally, through the adaptive weighted fusion module, the image recognition model not only effectively protects image privacy but also optimizes the prediction results without affecting the model's classification performance.

4. EXPERIMENTAL RESULTS AND ANALYSIS

Figure 5 shows the comparison data of the accuracy of different image recognition models with differential privacy fusion under different privacy budgets. In the case of a privacy budget of (b), the proposed algorithm performs relatively stable in the early stages of training (step 0-250), but quickly improves after step 500, increasing significantly from 62% at step 500 to 95.7% at step 2000. In contrast, the performances of Federated Averaging (FedAvg), Privacy-Preserving Generative Adversarial Network (PP-GAN*)*, and Multi-Party Computation (MPC) models are relatively inferior under the same privacy budget. FedAvg improves slowly after step 500, finally reaching 92% at step 2000; PP-GAN's growth rate is even slower after step 500, with a peak accuracy of 91.4%; although MPC improves rapidly after step 500, its final accuracy is still lower than that of the algorithm in this paper, reaching 94.6%. It can be seen that the proposed algorithm can quickly improve accuracy in the early stages and maintain high performance stably in the later stages under a privacy budget of (b). Under the condition of a privacy budget of 8, the algorithm in this paper still performs excellently. At step 500, the algorithm in this paper has already reached 62% and continues to improve, finally reaching a high accuracy of 95.7% at step 2000. This is significantly better than FedAvg and PP-GAN. FedAvg grows slowly after step 500, with a maximum of 92%; PP-GAN grows quickly in the initial stage but grows steadily in the later stage, with a maximum of 91.4%; the MPC model improves quickly after step 500, but the final accuracy is 94.6%. Combining the experimental results under two privacy budgets, the vision Transformer network model with differential privacy protection and the image recognition algorithm integrating differential privacy proposed in this paper significantly improve the accuracy and stability of image recognition, verifying its effectiveness in improving image recognition performance while protecting privacy.



(a) Privacy budget of 2

(b) Privacy budget of 8

Figure 5. Comparison of accuracy of different image recognition models with differential privacy fusion under different privacy budgets

In Table 1, the performance of different image recognition models with differential privacy fusion is compared in terms of computational cost (*FLOPs*), parameter amount, and *Top*-1 accuracy at different training epochs. The *Top*-1 accuracy of the proposed Model 1 (4 stages) at 10, 20, 30, 40, and 50 epochs is 16.32%, 26.39%, 33.26%, 37.54%, and 41.26%, respectively, which is significantly better than the performance of FedAvg and PP-GAN at the same epochs. Although the MPC model's accuracy is slightly higher than Model 1 at some epochs, overall, Model 1 demonstrates higher efficiency in terms of parameters and computational resource consumption. For example, Model 1 has only 1.58 *FLOPs* and 22.14 parameter amount at 10 epochs, whereas the corresponding *FLOPs* and parameter amount for the MPC model are 1.36 and 21.56, respectively, indicating that Model 1 maintains high image recognition capability while keeping computational complexity low. Combining these experimental results, it can be concluded that the proposed vision Transformer network model with differential privacy protection significantly improves the accuracy and stability of image recognition while protecting privacy. Especially in the early training stages, the performance of Model 1 is particularly outstanding, demonstrating its effectiveness across different training epochs.

Table 1. Performance comparison of different image recognition models with differential privacy fusion

Model	***FLOPs***	**Parameter Amount**	***Top*-1 Accuracy /*Epoch* (%)**				
			10	**20**	**30**	**40**	**50**
FedAvg	15.21	85.69	11.03	19.25	24.21	30.12	35.21
PP-GAN	4.23	23.14	12.89	23.16	30.26	34.56	39.87
MPC	1.36	21.56	17.85	27.54	33.15	36.59	40.26
The proposed model 1(4 *stages*)	1.58	22.14	16.32	26.39	33.26	37.54	41.26

Table 2. Image recognition accuracy of the model with different attention mechanisms

Model	***Top*-1 Accuracy (%)**	***Top*-5 Accuracy (%)**
+Threshold Attention	54.32	77.24
+Multi-head Self-Attention	56.98	79.65
+Hierarchical Attention Module	62.35	82.31

Table 2 shows the image recognition accuracy of the model with different attention mechanisms. Specifically, the model with threshold attention mechanism achieved a *Top*-1 accuracy of 54.32% and a *Top*-5 accuracy of 77.24%; the model with multi-head self-attention mechanism showed a slight improvement, reaching a *Top*-1 accuracy of 56.98% and a *Top*-5 accuracy of 79.65%; while the model with hierarchical attention module performed the best, achieving a *Top*-1 accuracy of 62.35% and a *Top*-5 accuracy of 82.31%. These data indicate that different attention mechanisms significantly affect model performance, with the hierarchical attention module providing the best results, significantly improving image recognition accuracy. Combining these experimental results, it can be concluded that the proposed vision Transformer network model with differential privacy protection significantly enhances image recognition performance when integrating different attention mechanisms. The introduction of the hierarchical attention module not only improves *Top*-1 and *Top*-5 accuracy but also further demonstrates the effectiveness of the proposed model in complex tasks.

In Table 3, the performance of the model in terms of image recognition accuracy, computational cost (*FLOPs*), and the parameter amount is shown for different numbers of attention heads. When the number of attention heads is 4, the model's *Top*-1 accuracy is 60.54% and *Top*-5 accuracy is 82.31%, with *FLOPs* of 0.712 and parameter amount of 9.87. When the number of attention heads increases to 8, the *Top*-1 accuracy improves to 62.31% and the *Top*-5 accuracy is 82.41%, but the corresponding *FLOPs* and parameter amount also increase significantly to 2.68 and 38.26, respectively. When the number of attention heads further increases to 12, the *Top*-1 accuracy slightly improves to 62.65%, but the *Top*-5 accuracy slightly decreases to 81.23%, while the *FLOPs* and parameter amount significantly increase to 6.12 and 86.23. These data indicate that although increasing the number of attention heads can improve the *Top*-1 accuracy of the model to some extent, the computational resource consumption and the number of parameters also increase significantly, and the *Top*-5 accuracy does not improve significantly, even slightly decreasing when the number of heads is 12. Combining these experimental results, it can be concluded that the proposed vision Transformer network model with differential privacy protection exhibits different trade-offs under different numbers of attention heads. Although increasing the number of attention heads can improve the *Top*-1 accuracy of the model to some extent, the associated computational resource and parameter consumption also increase significantly, and beyond a certain number of heads, the improvement in *Top*-5 accuracy is not significant.

Table 3. *Top*-1 and *Top*-5 accuracy of the model with different Numbers of attention heads

Heads	***FLOPs***	**Parameter Amount**	***Top*-1 Accuracy (%)**	***Top*-5 Accuracy (%)**
4	0.712	9.87	60.54	82.31
8	2.68	38.26	62.31	82.41
12	6.12	86.23	62.65	81.23

In the comparison of loss function values before and after using the vision Transformer network shown in Figures 6 and 7, significant differences and improvements can be observed. Before using the vision Transformer network, the model's loss function values fluctuate greatly across epochs. For example, at epoch 0, the loss values range from 0.72 to 0.7, and at epoch 15, the loss values still fluctuate between 0.73 and 0.77, without a significant downward trend, indicating large overall fluctuations and a lack of stable convergence. In contrast, after using the vision Transformer network, the model's loss function values show more stable changes and a gradual downward trend across epochs. At epoch 0, the loss value is 0.768, and at epoch 15, the loss value gradually decreases to 0.704, demonstrating significant convergence and stability.

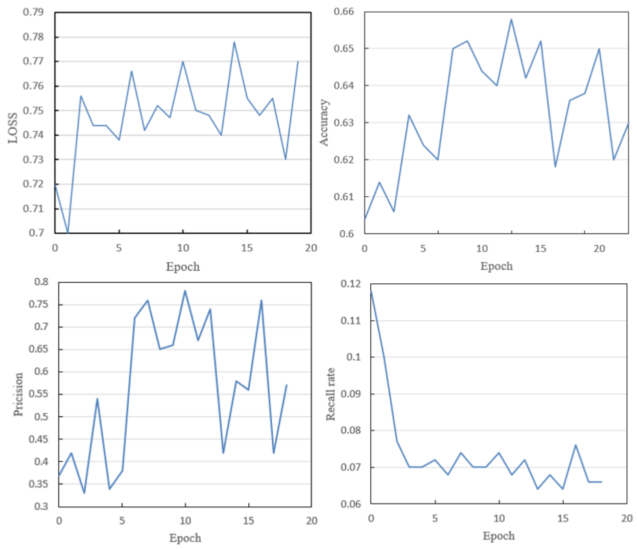


Figure 6. Performance comparison before using vision transformer network (four graphs: LOSS, accuracy, precision, and recall)

Before using the vision Transformer network, the model's accuracy fluctuates greatly across epochs. For example, at epoch 0, the model's accuracy is 0.604, and it slightly increases to 0.632 at epoch 5, but then the accuracy does not significantly improve and even decreases in some epochs, such as at epoch 15 where the accuracy drops to 0.618. At epoch 20, the accuracy still fluctuates between 0.62 and 0.65, without showing a significant convergence trend. In contrast, after using the vision Transformer network, the model's accuracy shows significant improvement and stability across epochs. At epoch 0, the accuracy already reaches 1.405, and it continues to rise during subsequent training, reaching 1.55 at epoch 15, showing good convergence and significant performance improvement.

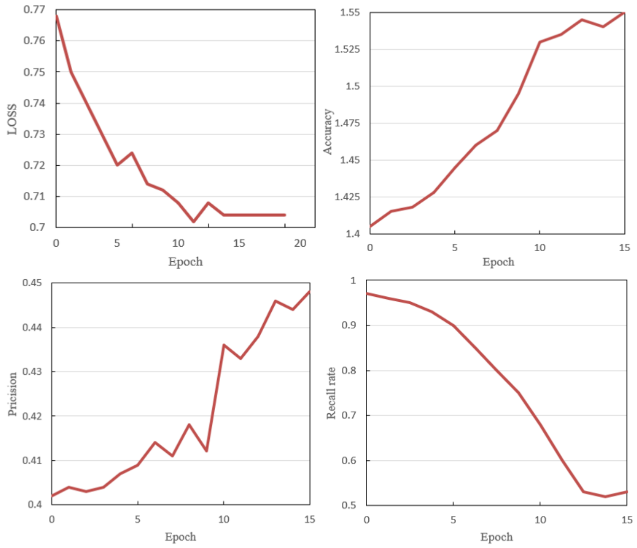


Figure 7. Performance comparison after using vision transformer network (four graphs: LOSS, accuracy, precision, and recall)

Before using the vision Transformer network, the precision of the model fluctuates greatly and is unstable across epochs. For example, at epoch 0, the model's precision is 0.37, and at epoch 5, it improves to 0.54, but then decreases again. Especially at epoch 10, the precision fluctuates significantly, ranging from 0.65 to 0.78, and at epoch 15, it fluctuates between 0.42 to 0.76, indicating that the model's precision is unstable and lacks a consistent upward trend. In contrast, after using the vision Transformer network, the model's precision shows significant improvement and stability across epochs. At epoch 0, the precision is 0.402, and it continues to rise during subsequent training, reaching 0.448 at epoch 15, with relatively stable changes between epochs, demonstrating good convergence and consistent performance improvement.

Before using the vision Transformer network, the model's recall rate is not only low but also fluctuates greatly and is unstable across epochs. For example, at epoch 0, the model's recall rate is 0.118, and it drops to 0.1 at epoch 5, further fluctuating in subsequent training, with a minimum of 0.064 and a maximum of only 0.076, showing overall poor performance and a lack of a consistent upward trend. In contrast, after using the vision Transformer network, the model's recall rate significantly increases to 0.97 at epoch 0, and although it decreases in subsequent training, it remains at a relatively high level, reaching 0.52 at epoch 15. This indicates that although the recall rate decreases, the model's recall rate significantly improves in the early epochs and maintains a certain level of stability during training.

Combining these experimental results, it can be concluded that the proposed vision Transformer network model with differential privacy protection shows significant advantages in reducing loss function values, improving model accuracy, improving model precision, and improving model recall rate. By introducing the vision Transformer network, the model not only shows better convergence and stability in loss function values but also effectively reduces the loss during training, indicating a significant improvement in the model's feature extraction and learning ability. The model also shows better improvement and stability in accuracy, with accuracy continuously improving stably during training, indicating a significant enhancement in the model's feature extraction and learning ability. The model shows better improvement and stability in precision, with precision continuously improving stably during training, indicating a significant enhancement in the model's feature extraction and learning ability. The model's recall rate significantly improves in the early epochs, demonstrating the advantages of the vision Transformer in feature extraction and classification ability. Even though the recall rate decreases in later training, the overall level is still much higher than the model before using the vision Transformer. The above experimental results further verify the effectiveness of this study, showing that by integrating the differential privacy mechanism and adaptively weighting the fusion of different models' prediction results, it is possible to improve image recognition performance while ensuring data privacy protection. This method significantly improves the model training effect, proving its potential and innovation in practical applications.

5. CONCLUSION

This study mainly includes two aspects: first, a vision Transformer network model with differential privacy protection is proposed, which enhances image recognition performance while protecting privacy by integrating the differential privacy mechanism; second, an image recognition algorithm for integrating differential privacy is designed, which further improves recognition accuracy and stability by adaptively weighting the fusion of different models' prediction results. The experimental results demonstrate the effectiveness and superior performance of the model in multiple aspects, including accuracy comparison under different privacy budgets, performance comparison of different image recognition models with differential privacy fusion, recognition accuracy with different attention mechanisms, *Top*-1 and *Top*-5 accuracy with different numbers of attention heads, and comparison of loss, accuracy, precision, and recall rate before and after using the vision Transformer network. Specifically, the experimental results show that under different privacy budgets, the proposed model significantly improves image recognition accuracy while protecting data privacy. The performance of different image recognition models with differential privacy fusion also shows significant differences, and the adaptive weighted fusion technique further enhances the model's stability and accuracy. After adding different attention mechanisms, the model's image recognition accuracy improves, indicating the important role of attention mechanisms in enhancing Transformer model performance. At the same time, under different numbers of attention heads, the model's *Top*-1 and *Top*-5 accuracy also show different trends, further verifying the effectiveness of the model design. Especially after using the vision Transformer network, the model shows significant improvements in loss, accuracy, precision, and recall rate, proving the advantages of the vision Transformer in image recognition tasks.

This study has important theoretical and practical value. By introducing the differential privacy protection mechanism, the proposed model significantly improves image recognition performance while ensuring data privacy, providing important reference value for research combining privacy protection and machine learning. Additionally, by adaptively weighting the fusion of different models' prediction results, this method further improves recognition accuracy and stability, demonstrating its potential in practical applications.

However, this study also has certain limitations. First, although the differential privacy protection mechanism is effective, the model performance is still limited under high privacy budgets, and the balance between privacy and performance needs further exploration. Second, this study mainly focuses on image recognition tasks, and its application and effects on other types of tasks (such as natural language processing or time series analysis) have not been verified. Future research can further explore the following directions: first, exploring more efficient differential privacy protection mechanisms to further enhance privacy protection while reducing the impact on model performance; second, applying this method to more types of tasks to verify its generality and effectiveness in different application scenarios; third, optimizing the design of attention mechanisms to further improve the model's recognition accuracy and performance stability. Additionally, combining other advanced deep learning techniques, such as GAN and self-supervised learning, may bring more improvements and innovations.

REFERENCES

[1] Thammastitkul, A. (2023). Assessing the effectiveness of image recognition tools in metadata identification through semantic and label-based analysis. International Journal of Metadata, Semantics and Ontologies, 16(3): 227-237. https://doi.org/10.1504/IJMSO.2023.137174

[2] Kwak, D., Choi, J., Lee, S. (2023). Rethinking breast cancer diagnosis through deep learning based image recognition. Sensors, 23(4): 2307. https://doi.org/10.3390/s23042307

[3] Cai, L. (2023). Investigation of the theory and applications of deep learning-based image recognition. In International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023), 12707: 19-25. https://doi.org/10.1117/12.2681279

[4] Mansor, Z., Maharum, S.M.M., Ghani, S.B.N.A., Anwar, R., Ahmad, I., Nurmantris, D.A. (2023). Utilizing image recognition to identify water quality in polluted river environments. In 2023 IEEE 9th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Kuala Lumpur, Malaysia, pp. 230-234. https://doi.org/10.1109/ICSIMA59853.2023.10373521

[5] Huang, J., Jiang, Y. (2023). Gnosis system of early gastric cancer based on artificial intelligence algorithm. In Proceedings - 2023 3rd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS), Shenyang, China, pp. 77-81.

[6] Pratik, R., Sendhil, R. (2023). Privacy protection against reverse image search. In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, pp. 1207-1214. https://doi.org/10.1109/ICAIS56108.2023.10073803

[7] Ghani, M.A.N.U., She, K., Rauf, M.A., Khan, S., Khan, J.A., Aldakheel, E.A., Khafaga, D.S. (2024). Enhancing security and privacy in distributed face recognition systems through blockchain and GAN technologies. Computers, Materials & Continua, 79(2): 2610-2623. https://doi.org/10.32604/cmc.2024.049611

[8] Nakamura, K., Nitta, N., Babaguchi, N. (2018). Encryption-free framework of privacy-preserving image recognition for photo-based information services. IEEE Transactions on Information Forensics and Security, 14(5): 1264-1279. https://doi.org/10.1109/TIFS.2018.2876752

[9] Tanwar, V.K., Raman, B., Rajput, A.S., Bhargava, R. (2022). SecureDL: A privacy preserving deep learning model for image recognition over cloud. Journal of Visual Communication and Image Representation, 86: 103503. https://doi.org/10.1016/j.jvcir.2022.103503

[10] Ishikawa, Y., Kondo, M., Kataoka, H. (2024). Learnable cube-based video encryption for privacy-preserving action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 7003-7013.

[11] Osorio-Roig, D., Rathgeb, C., Drozdowski, P., Terhörst, P., Štruc, V., Busch, C. (2022). An attack on facial soft-biometric privacy enhancement. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(2): 263-275. https://doi.org/10.1109/TBIOM.2022.3172724

[12] Morris, J., Newman, S., Palaniappan, K., Fan, J., Lin, D. (2021). "Do you know you are tracked by photos that you didn't take": Large-scale location-aware multi-party image privacy protection. IEEE Transactions on Dependable and Secure Computing, 20(1): 301-312. https://doi.org/10.1109/TDSC.2021.3132230

[13] Le, M.H., Carlsson, N. (2023). IdDecoder: A face embedding inversion tool and its privacy and security implications on facial recognition systems. In

Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy, NC, Charlotte, USA, pp. 15-26. https://doi.org/10.1145/3577923.3583645

[14] Guo, E., Li, P., Yu, S., Wang, H. (2022). Efficient video privacy protection against malicious face recognition models. IEEE Open Journal of the Computer Society, 3: 271-280. https://doi.org/10.1109/OJCS.2022.3218559

[15] Ouchi, Y., Uchida, H., Abe, N. (2023). Privacy-preserving image transformation method for person detection and Re-ID. In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, pp. 1798-1803. https://doi.org/10.1109/APSIPAASC58517.2023.10317180

[16] Tian, H., Zhu, T., Zhou, W. (2022). Fairness and privacy preservation for facial images: GAN-based methods. Computers & Security, 122: 102902. https://doi.org/10.1016/j.cose.2022.102902

[17] Kagan, D., Alpert, G.F., Fire, M. (2023). Zooming into video conferencing privacy. IEEE Transactions on Computational Social Systems, 11(1): 933-944. https://doi.org/10.1109/TCSS.2022.3231987

[18] Zhao, Y., Yu, Z., Li, X., Cai, M. (2019). Chinese license plate image database building methodology for license plate recognition. Journal of Electronic Imaging, 28(1): 013001-013001. https://doi.org/10.1117/1.JEI.28.1.013001

[19] Prakash, P., Ding, J., Li, H., Errapotu, S.M., Pei, Q., Pan, M. (2020). Privacy preserving facial recognition against model inversion attacks. In GLOBECOM 2020-2020 IEEE Global Communications Conference, Taipei, Taiwan, pp. 1-6. https://doi.org/10.1109/GLOBECOM42002.2020.9322508

[20] Tekli, J., Al Bouna, B., Tekli, G., Couturier, R. (2023). A framework for evaluating image obfuscation under deep learning-assisted privacy attacks. Multimedia Tools and Applications, 82(27): 42173-42205. https://doi.org/10.1007/s11042-023-14664-y