Supplementary material: Hiding Visual Information via Obfuscating Adversarial Perturbations

Zhigang Su[1,*] Dawei Zhou[1,*] Nannan Wang[1,†] Decheng Liu[1,†]
Zhen Wang[2] Xinbo Gao[3]
[1]Xidian University, [2]Zhejiang Lab
[3]Chongqing University of Posts and Telecommunications
{zgsu, dwzhou}.xidian@gmail.com, {nnwang, dchliu}@xidian.edu.cn
wangzhen@zhejianglab.com, gaoxb@cqupt.edu.cn

A. Time Spent on Protection

The GPU model used for all our experiments is NVIDIA TITAN Xp. Since our method is an iterative method, it takes a certain amount of time to protect images one by one. However, we can increase the protection efficiency by increasing the batch size. As shown in Figure. 1, we tested the average time it takes to protect an image under different batch sizes, from which we can conclude that adjusting the batch size can greatly improve the efficiency of protection. Moreover, the number of iterations also greatly affects the time required for protection. We tested the effect of different iterations on the time required for protection and the quality of protected images as shown in Table 1 and Figure. 2. We can conclude that, within a certain range, we can speed up the protection time with little impact on protection quality by reducing the number of iterations. We can also spend more time increasing the protection quality by increasing the number of iterations.

B. Security Analysis

Key Model Space Analysis. In our proposed method, we choose the generative model as the key, which has a large key space. Take the key model used in our experiments as an example, it has a size of 11.383M. Such a large key space can be disastrous for those who use exhaustive attacks.

Histogram Analysis and Correlation Analysis. Plain images have a strong correlation between two adjacent pixels in the horizontal and vertical directions[1], and protection methods with good properties often need to break this correlation[5]. Therefore, we performed a correlation analysis of our proposed method, and its results are shown in Figure. 3. Compared to the strong correlation of the original image, the protected image we obtained greatly reduces the correlation between adjacent pixels. We also performed histogram analysis on the original and protected images, and the results are shown in Figure. 4. From it, it can be concluded that the histogram statistical properties of the protected image and the original image are completely different, and the protected image more closely resembles a gaussian distribution. So the protected images obtained by the AVIH method have well statistical characteristics.

Identifiability of Protected Images. We use one face recognition model as the target model to generate protected images, and then use another face recognition model to identify the results obtained in Table 2. It can be concluded that the protected image obtained by a target model cannot be used normally by other face recognition models. Storing such protected images in a cloud environment can greatly improve security.

C. Privacy Protection for Classification Tasks

Detailed Experimental Setup. When implementing the AVIH method on the classification task, we adjust the batch size to 10 and initialize the protected image as the original image. We set the number of iterations to 600.

The Quality of Recovered Images. We show the protection results of LIE [4] and ITP [2] in Figure. 6. We tested the average SSIM values of the AVIH method for the protected and recovery images of the test set in *CIFAR-10* [3]. The results

*Equal contributions. † Corresponding author.

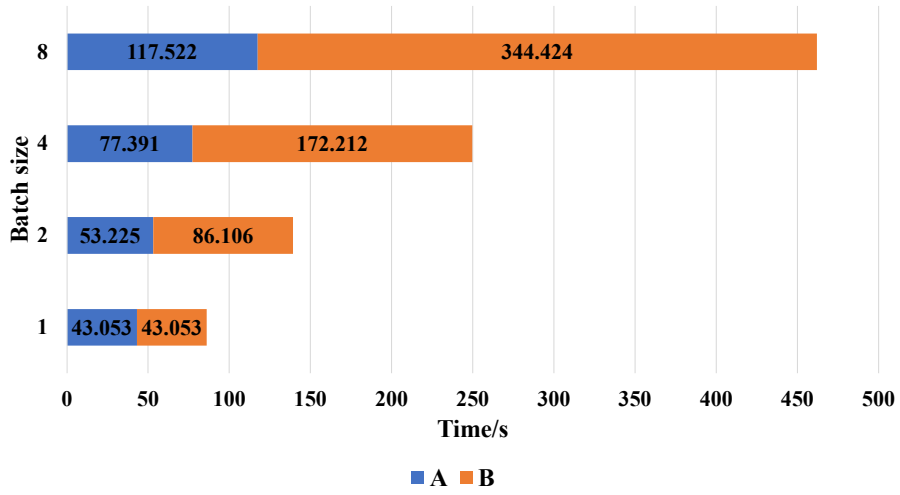

Figure 1: The time required to protect an image by the AVIH method and the effect of batchsize on protection time. A represents the average time required to protect a batch, and B represents the average time required to protect the corresponding number of samples when the batch size is 1. We protected ten batches and averaged the time spent.

Figure 2: Protected and recovery images generated by different iterations. Each pair of images includes the protected image and recovery image, respectively. The number of iterations they correspond to is marked below the image.

are shown in Table. 3. We compare our results with LIE [4] and ITP [2], where LIE recovers the original image, but the protection quality is weaker and has a significant impact on the model accuracy. ITP mitigates the impact of the protected image on the model accuracy, but the protected image becomes unrecoverable. And our method ensures a strong protection strength while the impact on the model accuracy is very slight.

Test on IamgeNet. We set the target model as ResNet50 and selected 50 images in ImageNet for testing as in Table 4.

D. Details of Variance Consistency Loss

Motivation. We replaced variance consistency loss with different losses and test the impact of these losses on visual information hiding separately. The results are shown in Figure. 7, where VC Loss represents our proposed variance consistency loss, MSE Loss represents the mean square error loss of the protected image and the original image, T Loss represents the mean square error of the protected image and the full gray image, and TV Loss represents the total variation loss of the protected image. From it, it can be concluded that the conventional loss focus on the difference between pixels. Images obtained by maximizing MSE between the protected image and the original image still retain some spatial features of the original image (*e.g.*, the contours of a human face) despite large pixel changes (*e.g.*, color textures), as shown in Figure. 7. To solve this issue, we consider making the pixel distribution of the protected image as consistent as possible at each location, so that the protected image cannot exhibit obvious spatial features in the pixel space. However, it is difficult (and unnecessary) to make every pixel converge to the same value. We thus block the image so that the pixel distribution between each block is similar while giving more possibilities for variation of pixels within the block.

References

[1] Hossein Movafegh Ghadirli, Ali Nodehi, and Rasul Enayatifar. An overview of encryption algorithms in color images. *Signal Processing*, 164:163–185, 2019. 1

Table 1: The effect of different iterations on protection time and protected image quality. Time represents the average time required to protect a batch, SSIM represents the average SSIM between the recovery image and the original image, and COS represents the average cosine similarity between the original image feature and the protected image features. We tested 10 samples and averaged these metrics.

Iterations	100	200	400	600	800
Time(s)	5.651	11.251	22.703	32.703	43.053
SSIM	0.721	0.829	0.886	0.895	0.912
COS	0.973	0.983	0.995	0.997	0.997

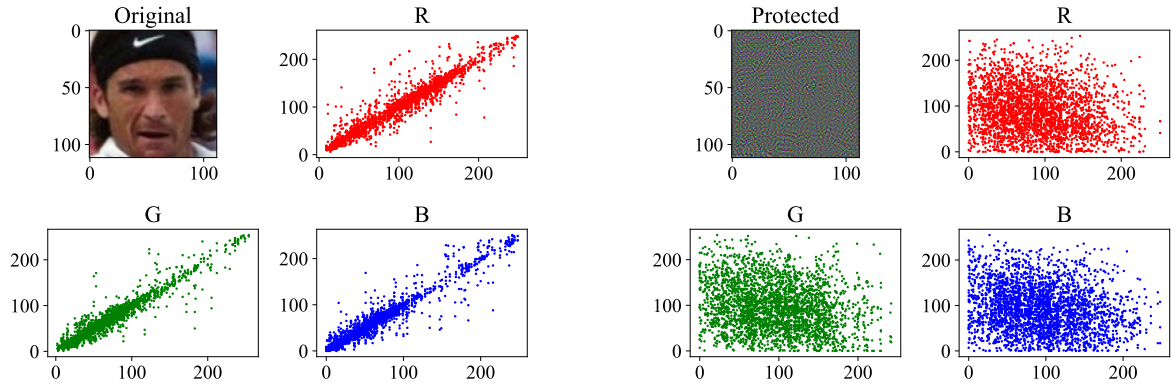

Figure 3: Results of correlation analysis. We randomly select 1000 pairs of adjacent pixel points for the original and protected images, respectively, and calculate their correlation coefficients in horizontal, vertical and diagonal directions.

Table 2: Accuracy (percentage) of predicting protected images using a model different from the target face recognition model. The same name represents the same model.

Target model \ Test model	AdaFace	ArcFace	CosFace	SphereFce
AdaFace	98.6	0	-	-
ArcFace	0	96.5	-	-
CosFace	-	-	89.4	0
SphereFce	-	-	0	80.3

[2] Hiroki Ito, Yuma Kinoshita, Maungmaung Aprilpyone, and Hitoshi Kiya. Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks. *IEEE Access*, 9:64629–64638, 2021. 1, 2, 4

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[4] Masayuki Tanaka. Learnable image encryption. In *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2, 2018. 1, 2, 4

[5] Aqeel ur Rehman, Xiaofeng Liao, Rehan Ashraf, Saleem Ullah, and Hueiwei Wang. A color image encryption technique using exclusive-or with dna complementary rules based on chaos theory and sha-2. *Optik*, 159:348–367, 2018. 1

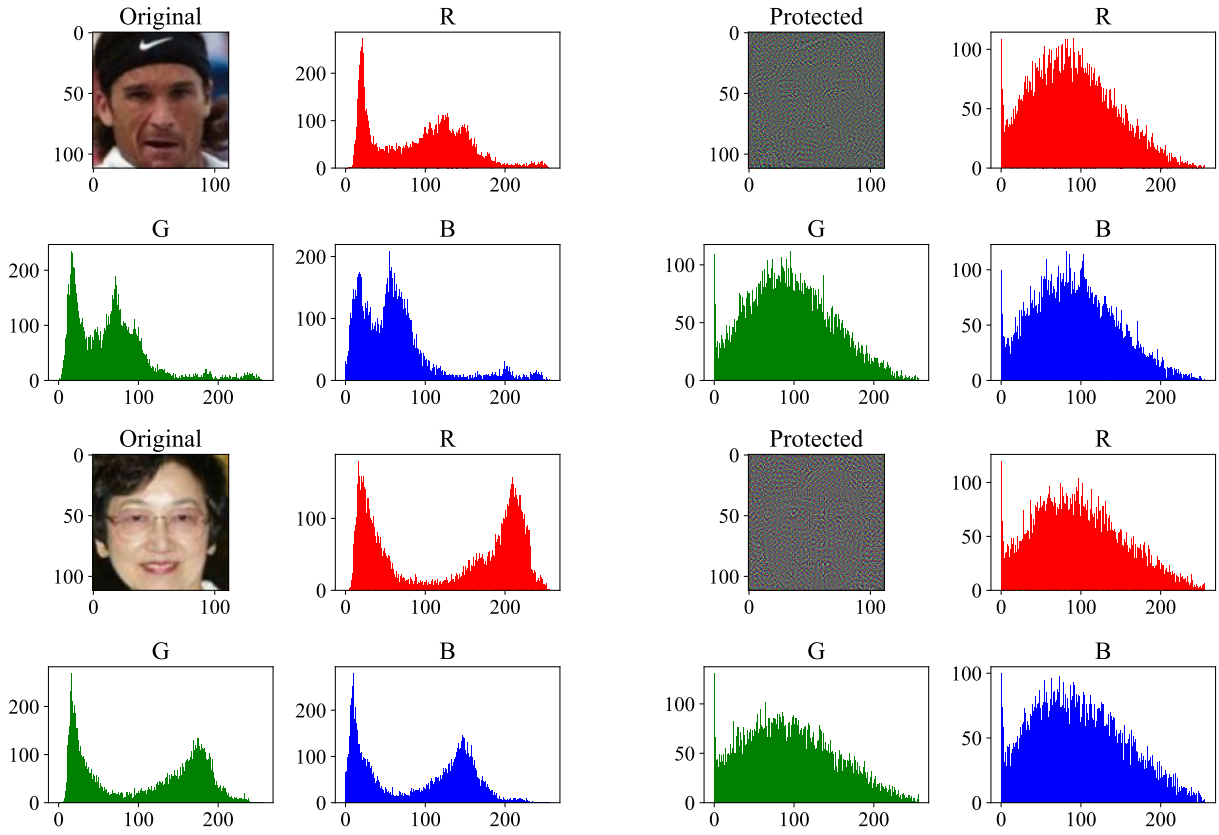

Figure 4: Results of histogram analysis of some original and protected images. The original image is on the left and its corresponding protected image is on the right.

Table 3: Image visual quality metrics of visual information hiding methods of classification models on *CIFAR-10*.

Method	Model	$SSIM_e$	$SSIM_d$
LIE [4]	VGG19	0.178	1.000
	Resnet50	0.178	1.000
ITP [2]	VGG19	0.068	-
	Resnet50	0.073	-
AVIH(our)	VGG19	0.171	0.900
	Resnet50	0.198	0.923

Table 4: Evaluation results of 50 samples in ImageNet. $SSIM_d$ represents the SSIM between the original and the recovered image.

Model	Original Acc.(%)	Protect Acc.(%)	$SSIM_d$
ResNet50	0.86	0.86	0.90

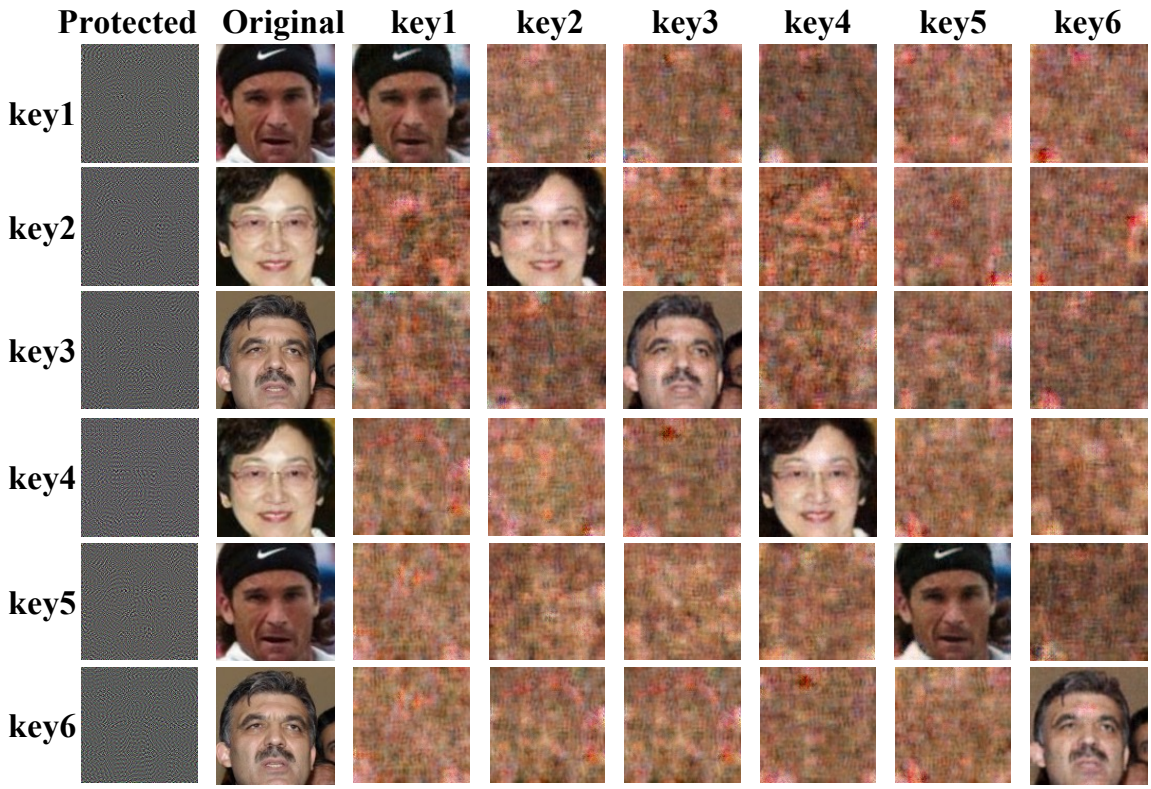

Figure 5: Results of the key model randomness analysis. Due to space limitations, we show the results for 6 key models.

Figure 6: Results of the LIE method and the ITP method. The top group is the result of the LIE method, and the bottom group is the result of the ITP method. Each column represents a different sample.

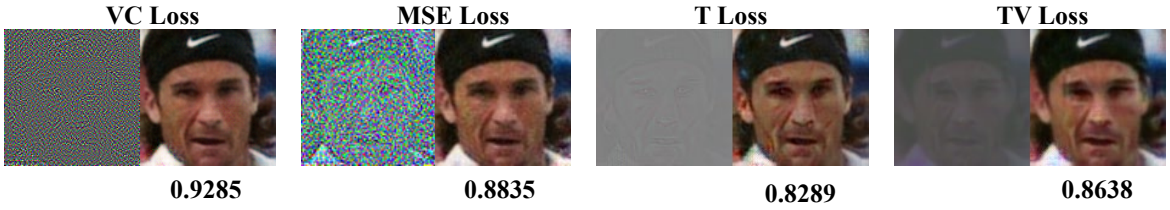

Figure 7: Impact of different losses on protected quality and recovery quality. For each pair of images, the former is the protected image and the latter is the recovery image. Below the recovery image is marked the SSIM value between it and the original image.