Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

HVS-inspired adversarial image generation with high perceptual quality

Yuan Xue[1], Jian Jin[2*], Wen Sun[2] and Weisi Lin[2]

Abstract

Adversarial images are able to fool the Deep Neural Network (DNN) based visual identity recognition systems, with the potential to be widely used in online social media for privacy-preserving purposes, especially in edge-cloud computing. However, most of the current techniques used for adversarial attacks focus on enhancing their ability to attack without making a deliberate, methodical, and well-researched effort to retain the perceptual quality of the resulting adversarial examples. This makes obvious distortion observed in the adversarial examples and affects users' photo-sharing experience. In this work, we propose a method for generating images inspired by the Human Visual System (HVS) in order to maintain a high level of perceptual quality. Firstly, a novel perceptual loss function is proposed based on Just Noticeable Difference (JND), which considered the loss beyond the JND thresholds. Then, a perturbation adjustment strategy is developed to assign more perturbation to the insensitive color channel according to the sensitivity of the HVS for different colors. Experimental results indicate that our algorithm surpasses the SOTA techniques in both subjective viewing and objective assessment on the VGGFace2 dataset.

Keywords Just noticeable difference, Privacy-preserving, Human visual system, Adversarial attack, Edge-cloud computing

Introduction

DNNs have achieved incomparable performance in various of traditional computer vision tasks, e.g., image classification [1], image processing [2–4],image quality assessment [5–8], etc.

Research has demonstrated that deep learning algorithms can surpass human performance in specific tasks, such as facial recognition [9] and image classification [10]. It has been shown that deep learning algorithms even outperform human beings for certain tasks, e.g., face recognition, image classification, and so on. These computing models has applied in edge computing which emphasizes distributing computation and data to enhance the system efficiency and performance. Edge computing [11–14]provides real-time data processing, minimizes network latency, and enhances system reliability and security. This technology has a broad range of applications, including industrial automation [15], healthcare [16], smart cities [17], environment [18], IoT [19] and so on.

However, the robustness of the DNN models are still weak, and DNN-based systems are highly vulnerable to adversarial examples [20]. For instance, the adversarial examples, which are generated by injecting elaborated perturbation into clean images, lead the classifier to misclassify it. If the weak models are applied on edge-cloud computing, then the security issue on identification will become severe. Although Kurakin et al. [21] stated that adversarial examples brought the hidden threat to the classifiers in the physical world scenarios, it had significant meanings in the privacy-preserving field and its relevant applications. Massive face photos are shared in social networking services (SNS) in our daily life. To

*Correspondence:
Jian Jin
jian.jin@ntu.edu.sg
[1] School of Software, Fudan University, Shanghai, China
[2] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

avoid their identity information being used maliciously, their private information is protected by injecting elaborated perturbation into photos. However, perturbation commonly leads to the degradation of the perceptual quality, which decreases the photo-sharing experience on social media. Hence, how to generate perturbation to avoid being perceived by the HVS and being able to fool the DNNs simultaneously is a significant problem to settle.

Most existing methods [20–29] focused on tricking DNN based recognition systems. Only a few of them [27–29] tried to preserve the quality of adversarial examples. However, the perceptual quality of adversarial images is still awful.

In this paper, an HVS-inspired adversarial example generation method is proposed for privacy-preserving with high perceptual quality. To achieve this, we first propose a novel perceptual loss based on JND characteristics, which assesses the adversarial examples along three quality-related factors between the adversarial example and its associated original image. The perturbation that is beyond the JND thresholds is perceived by the HVS, and therefore is counted in the perceptual loss in this work. Besides, we develop a perturbation adjustment strategy to integrate into JND-based perceptual loss, which is able to assign more perturbation to the insensitive color channel. All these designs make the perturbation unable to be perceived by the HVS as much as possible while successfully attacking the DNN-based recognition system and maintaining the high perceptual quality of adversarial examples. Experimental results indicate that the suggested method achieves state-of-the-art results in adversarial example generation with respect to perceptual quality.

Related work

Adversarial attack

Szegedy et al. [22] suggested the L-BFGS algorithm along with a box restraint to create the adversarial examples. Goodfellow et al. [23] created the Fast Gradient Sign Method (FGSM), which executes a single step on the clean image using a computed gradient with sign. After that, most of the SOTA works [20, 21, 24–26] are based on FGSM. e.g. Rozsa et al. [24] used the sign of the magnitude rather than gradient and proposed the Fast Gradient Value (FGV). Madry et al. [25] presented the Projected Gradient Method (PGD), which is a white-box attack with accessing the model gradients. Kurakin et al. [21] proposed the Iterative Fast Gradient Sign Method (I-FGSM), iteratively updating the image generation. Dong et al. [26] developed Momentum Iterative Fast Gradient Sign (MI-FGSM), which integrates the momentum term to stabilize update directions to escape poor local maxima during iteration. Xie et al. [20] presented the Diverse Inputs Iterative Fast Gradient Sign Method DI^2-FGSM and Momentum Diverse Inputs Iterative Fast Gradient Sign Method (M-DI^2-FGSM) [20], which considers the input diversity strategy to generate adversarial samples.

The methods mentioned above only pay attention to finding an efficient way to generate adversarial images to fool the DNNs tasks rather than preserving the adversarial image quality. To hide the unnecessary distortion, for example, Zhang et al. [27] first attempted to integrate the hand-crafted JND coefficients into FGSM while generating adversarial examples. Although the perceptual quality was improved to some degree, its low success attack ratio limited its application of privacy-preserving. To keep a high success attack ratio, Adil et al. [28] iteratively distorted a clean image until the classifier made wrong predictions. It reduced the injected perturbation and distortion of adversarial examples. After that, Sun et al. [29] utilized quality metric SSIM [30] to supervise the adversarial example generation. However, SSIM can't fully represent the perceptual quality of the HVS. Distortions can be observed in [29].

Hence, generating adversarial examples with high perceptual quality is still an open problem. As reviewed above, the perceptual quality of adversarial examples generated by the SOTA methods above is still unacceptable. The perturbation is not hidden in insensitive areas of images so that the distortion is still obvious.

Just notification difference

JND [31] reflects the minimum amount of change in visual signals that can be captured by the human visual system (HVS). This reflects the redundancy of perceptual information contained in visual signals. Generally, there are two categories of JND models: HVS-inspired [32, 33] and learning-based JND [34–38]. HVS-inspired JND is obtained by the characteristics of the HVS. For example, Chou et al. [31] put forward a JND model in the spatial domain by merging contrast masking (CM) and luminance adaptation (LA). Yang et al. [39] generated the JND model by introducing a nonlinear additivity model for masking effects (NAMM). Wu et al.[33] use the pattern complexity (PC) of visual content to further improve the accuracy of the HVS-inspired JND model. However, due to the HVS is not sufficiently ackownledged by human being, so these hand-crafed HVS feature based method cannot obtain the more accurate JND threshold as well.

As deep learning achieved incomparable success in dozens of visual problems, learning-based JND models were proposed [35–37, 40–47]. Due to inefficiency in generating labeled JND datasets, the unsupervised JND models are proposed recently. Jin et al. [37] proposed the

RGB-JND model, which takes the stimuli of the whole color space into account. They also proposed HVS-SD JND models using the prior information in image reconstruction task for better guiding the JND generation. It has already achieved best MOS performance among all SOTA models.

JND based privacy-preserving adversarial image generation

Problem formulation

In Sec. 1, the perturbation is required to fool the DNN based recognition systems without being perceived by the HVS. Thus, the identification information of adversarial samples is protected while maintaining their high perceptual quality. We have illustrated the overview of the adversarial attack process in Fig. 1. It is mentioned in figure that the attack process followed by two processes, which is perceptual quality perserving L_{per} and adversarial attack L_{adv}. The generated adversarial image I_{adv} is put through the DNN recognization system. If the recognization result is not matched the correct result then the iteration process will continue, else the attack will finish.

Here, the clean image and its associated adversarial example are marked by X and Z, respectively. The injected perturbation is marked by P. Then, we have $P = Z - X$. As a reference in our adversarial image generation method, we incorporate the HVS-SD JND [38] and denote the JND of the clean image as J. The HVS-SD JND is a recent learning-based model that outperforms both handcrafted JND methods [32, 33] and other learning-based JND methods [34–38], achieving state-of-the-art performance. Then, the generation of adversarial examples for privacy-preserving can be explicated as follows

$$\arg\min_{Z} \mathcal{L} = \begin{cases} \alpha \cdot \mathcal{L}_{adv}(Z, X) + \mathcal{L}_{per}(Z, X, J), \text{if non-targeted.} \\ \alpha \cdot \mathcal{L}_{adv}(Z, C) + \mathcal{L}_{per}(Z, X, J), \text{otherwise.} \end{cases} \quad (1)$$

Both non-targeted attacks and targeted attacks are included in this work. For non-targeted attack, we have $\mathcal{L}_{adv}(Z, X) = -\mathcal{E}(\theta(Z), \theta(X))$, where $\mathcal{E}(\cdot, \cdot)$ is a variant of cross-entropy loss [29]. It makes sure that the recognition result of adversarial example Z (denoted by $\theta(Z)$, where $\theta(\cdot)$ denotes the DNN-based recognition system) has deviated from that of the clean image X (denoted by $\theta(X)$). For targeted attack, we have $\mathcal{L}_{adv}(Z, X) = \mathcal{E}(\theta(Z), C)$. It makes sure that the recognition result of Z is close to a targeted label C. $\mathcal{L}_{per}(\cdot, \cdot, \cdot)$ is a JND-based perceptual loss, which is used to maintain the high perceptual quality of Z and is to be introduced in the next subsection. J is the JND threshold generated through [38], which is the latest learning-based JND model that overperforms the SOTA JND models. The hyper-parameter α has been set to 1 in order to balance between these two factors.

JND based perceptual loss

In this subsection, we introduced a new perceptual loss that is based on JND, $\mathcal{L}_{per}(\cdot, \cdot, \cdot)$, where three quality-related factors (including the deviation, fidelity, and gradient of the adversarial example) are formulated with three sub-losses by taking the JND into account. Besides, we also design a perturbation adjustment

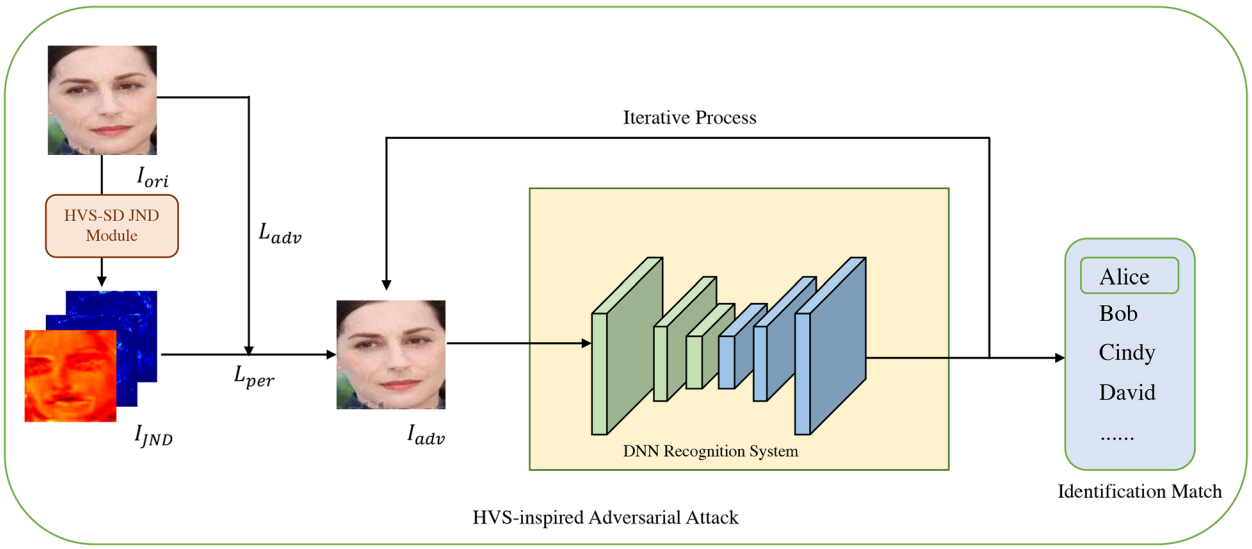

Fig. 1 The process of HVS-inspired adversarial attack: I_{ori} is the original clear image, I_{jnd} is the three-channel JND map through the HVS-SD JND method, I_{adv} is the generated adversarial image. L_{adv} and I_{per} is the process of adversarial attack and quality preserving process respectively

matrix $\mathcal{M}(\cdot)$ to assign more perturbation to the color channels that are insensitive. The details are as follows.

$\mathcal{L}_{per}(\cdot,\cdot,\cdot)$ mainly contains three JND based sub-losses: 1) deviation loss $\mathcal{L}_1(\cdot,\cdot,\cdot)$, 2) fidelity loss $\mathcal{L}_2(\cdot,\cdot,\cdot)$, and 3) gradient loss $\mathcal{G}(\cdot,\cdot,\cdot)$. The effectiveness of the three sub-losses will be proved in ablation experiments in Sec. 4.2. We have

$$\mathcal{L}_{per} = \beta_1 \cdot \mathcal{L}_1(Z,X,J) + \beta_2 \cdot \mathcal{L}_2(Z,X,J) + \beta_3 \cdot \mathcal{G}(Z,X,J) \tag{2}$$

$\mathcal{L}_1(Z,X,J)$ is used to control the magnitude of the perturbation. $\mathcal{L}_2(Z,X,J)$ is used to constrain the fidelity distortion between X and Z. $\mathcal{G}(Z,X,J)$ ensures that Z and X keep a similar gradient. Hyper-parameters β_1 ,β_2 and β_3 are used to balance the three items. Notice that the three items above are designed based on the JND.

Deviation loss $\mathcal{L}_1(Z,X,J)$ describes the actual deviation from adversarial example Z to its associated clean image X. Here, only the deviation beyond the JND threshold is considered as the actual deviation that HVS can obviously be perceived. In other words, the deviation under the JND is not perceived by the HVS, which is not taken into account while calculating the deviation loss. Therefore, we have

$$\begin{aligned}\mathcal{L}_1(Z,X,J) = \sum_n \sum_h \sum_w (|Z(n,h,w) - X(n,h,w)| \\ - J(n,h,w)) \cdot \lambda(n,h,w)\end{aligned} \tag{3}$$

and

$$\lambda(h,w,n) = \begin{cases} 1, \text{if } |Z(n,h,w) - X(n,h,w)| > J(n,h,w), \\ 0, \text{otherwise.} \end{cases} \tag{4}$$

where $Z(n,h,w)$, $X(n,h,w)$, and $J(n,h,w)$ denote the pixels' value located at (n,h,w) of Z, X, and J, respectively. n denotes the index of color channel. We have $n \in \{r,g,b\}$.

Similarly, for fidelity loss $\mathcal{L}_2$, the perturbation beyond (or under) the JND threshold is (or not) counted in, as the fidelity distortion beyond (or under) the JND threshold is (or not) perceived by the HVS. We have

$$\begin{aligned}\mathcal{L}_2(Z,X,J) = \sum_n \sum_h \sum_w (|Z(n,h,w) - X(n,h,w)| \\ - J(n,h,w))^2 \cdot \lambda(n,h,w).\end{aligned} \tag{5}$$

Gradient similarity, as a major perceptual metric for the HVS, is also taken into account based on the JND so that we can better constrain the adversarial example generation. We use the ℓ_1-norm variation to describe the gradient loss. The gradient loss $\mathcal{G}(Z,X,J)$ can be formulated as:

$$\mathcal{G}(Z,X,J) = \mathcal{L}_1(g(Z),g(X),g(J)) \tag{6}$$

where $g(\cdot)$ is the Sobel operator [48]. We use $g(\cdot)$ to calculate the gradient in both horizontal and vertical directions of Z, X, and J. Also, only the gradient beyond JND is counted in the gradient loss.

Considering that the HVS has a different sensitivity to different colors. For instance, the HVS has high, medium, and low sensitivity in green, red, and blue, respectively. Hence, there are small, medium and large JNDs in the green, red, and blue channels demonstrated in [37] as well.

In view of this, we design a perturbation adjustment matrix $\mathcal{M}(n)$ to adjust the distribution of perturbation among different color channels. Hence, Eq. (4) can be reformulated as:

$$\lambda(n,h,w) = \begin{cases} \mathcal{M}(n), \text{if } |Z(n,h,w) - X(n,h,w)| > J(n,h,w), \\ 0, \text{otherwise.} \end{cases} \tag{7}$$

$\mathcal{M}(n)$ is able to assign more perturbation to insensitivity color channels, like blue and red channels, while less perturbation is assigned to the green channel. The specific value of $\mathcal{M}(n)$ is adjusted according to the regular pattern of different color channels perceived in [38].

Optimization

During our optimization of the function in Eq. (1), we still use the gradient descent algorithm. Besides, as the proposed algorithm is a balance between the perceptual quality and classification deviation, unsuccess attacks are inevitable when perceptual loss trade off too much against classification deviation loss. To make sure that all the generated examples successfully trick the recognition system and are used for privacy-preserving, the hyper-parameter β_i in Eq. (2) will be adjusted when unsuccess attacks occur. The adjustment of β_i is as follows

$$\beta_i = \beta_i + \delta \tag{8}$$

where δ is an adjuster. The adjustment of β_i will be activated when unsuccess attacks occur. Then, a new adversarial example will be generated, This will attack the identification system again until it succeeds. With such optimization, we can achieve the 100% success attack ratio on VGGFace2.

Experiments

Experimental settings

Datasets and Anchors. Experiments are conducted on VGGFace2 [9] dataset. The DNN based face recognition system is trained on VGGFace2, which contains 8,631 identities. To determine whether the provided face image applys to corresponding identities, 100 images are randomly selected from VGGFace2 for non-attacked attack and targeted attack evaluation among 6 SOTA anchor methods and the proposed method. The anchor methods include BIM [21], PGD [25], MIFGSM [26], DI^{2}FGSM

[20], JNDMGA [28], MND [29]. In this paper, it's important to note that all the outcomes from the suggested approach were achieved.

Evaluation metric. The objective evaluation metrics used to assess the quality of adversarial examples generated with anchor and proposed methods are the Peak Structural Similarity Index (SSIM) and Signal-to-Noise Ratio (PSNR). 60 subjects are invited to conduct the subjective viewing test based on ITU-R BT.500-11 criterion [49]. The participants in the subjective viewing test encompass individuals from diverse backgrounds, including students, teachers, doctors, artists, researchers, and others, all of whom hold a bachelor's degree or higher. The age range of the participants falls between 18 and 45 years, and they do not have any vision impairments. Out of the total participants, 38 are male, while the remaining participants are female.

Ablation Setting. To verify the reasonability of $\mathcal{L}_{per}$, ablation experiments are conducted with different settings. The details are listed below:

- $\mathcal{L}_1$ *evaluation:* Only deviation loss $\mathcal{L}_1(\cdot,\cdot,\cdot)$ in formula (1) is optimized. That is, $\beta_1 \neq 0, \beta_2 = \beta_3 = 0$.
- $L1_{ori}$ *evaluation:* Only original $L1$ loss $L1_{ori}(\cdot,\cdot)$ in formula (1) is optimized. That is, $\beta_1 \neq 0, \beta_2 = \beta_3 = 0$.
- $\mathcal{L}_2$ *evaluation:* Only fidelity loss $\mathcal{L}_2(\cdot,\cdot,\cdot)$ in formula (1) is optimized. We have $\beta_2 \neq 0, \beta_1 = \beta_3 = 0$.
- $L2_{ori}$ *evaluation:* Only original $L2$ loss $L2_{ori}(\cdot,\cdot)$ in formula (1) is optimized. We have $\beta_2 \neq 0$, $\beta_1 = \beta_3 = 0$.
- $\mathcal{L}_1 + \mathcal{L}_2$ *evaluation:* Deviation loss $\mathcal{L}_1(\cdot,\cdot,\cdot)$ and fidelity loss $\mathcal{L}_2(\cdot,\cdot,\cdot)$ are optimized for generating Z. That is, $\beta_1, \beta_2 \neq 0, \beta_3 = 0$.
- $\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{G}$ *evaluation:* All the three items in formula (1) are used for constrain Z. $\beta_1, \beta_2, \beta_3 \neq 0$

In targeted attack, if $\beta_i \neq 0$, we set $\beta_i = 30$ and adjuster $\delta = -1$. In non-targeted attack, for the $\beta_i \neq 0$, we set $\beta_i = 1500$ and adjuster $\delta = -50$. α is set to 1 in both targeted attacks and non-targeted attacks. For metric $\mathcal{M}(n)$, we set $\mathcal{M}(r) = 3, \mathcal{M}(g) = 5, \mathcal{M}(b) = 1$, respectively.

Comparison and objective evaluation

Figure 2 shows a comparison of the adversarial images generated with anchors and our proposed method under targeted attack and non-targeted attack. As the perturbation in our adversarial examples under the JND are as much as possible, our adversarial examples are closest to clean images. The algorithm of our adversarial process is as following Algorithm 1. In the algorithm, We generated the adversarial image through JND, if the adversarial image cannot successfull attack the DNN recognization system, then we iterate the process and update the pertubation following the JND threshold.

Generate JND through HVS-SD JND J
$t = 0$
$Z_0 = X$
Calculate the pertubation according to the equation(2)
while *adversarial attack is not success* **do**
  update Z_{t+1} according to equition(8)
  $t = t + 1$
end

Algorithm 1 Optimization for generating HVS-inspired Adversarial image

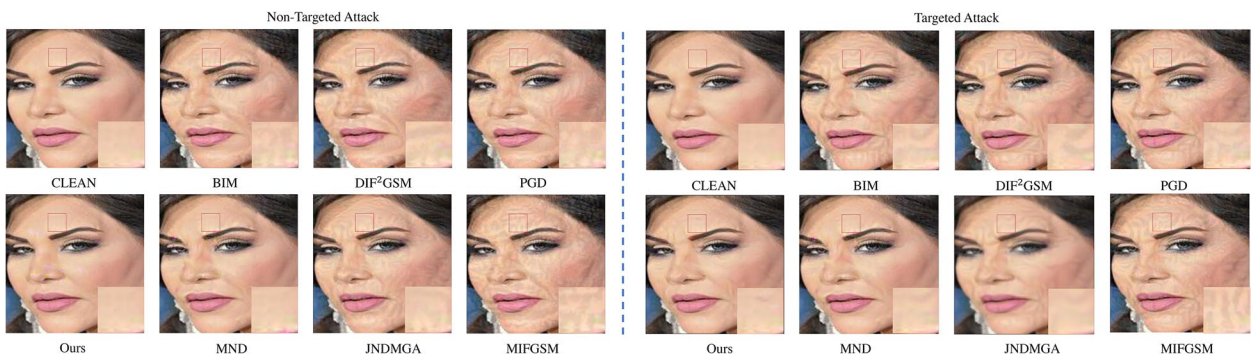

Fig. 2 Comparison among five anchor methods and the proposed method under targeted and non-targeted attack on face recognition. Anchor methods include BIM [21], DI^{2}FGSM [20], MIFGSM [26], MND [29], JNDMGA [28] and PGD [25]. A specific part in each image is enlarged in the red box

Table 1 Objective Viewing Test Of The Proposed Method And Anchor Methods Conducted Under Targeted/Non-targeted Attacks

Image Index	Non-Targeted Attack		Non-Targeted Attack	
	PSNR	SSIM	PSNR	SSIM
MIFGSM [26]	31.78±0.10	0.8421±0.0169	31.96±0.12	0.8451±0.0165
PGD [25]	32.63±0.26	0.8613±0.0132	32.94±0.22	0.8671±0.0133
DI^{2}FGSM [20]	32.58±0.01	0.8709±0.0126	32.94±0.21	0.8764±0.0127
BIM [21]	33.38±0.48	0.8872±0.0094	34.31±0.48	0.9054±0.0080
JNDMGA [28]	34.91±1.99	0.9124±0.0218	35.69±2.39	0.9422±0.0163
MND [29]	36.84±1.48	0.9176±0.0330	38.39±1.58	0.9651±0.0047
$L2_{ori}$	34.34±1.43	0.9025±0.0113	32.27±1.56	0.8521±0.0262
$\mathcal{L}_2$	33.98±1.95	0.9294±0.0175	31.57±2.66	0.8743±0.0364
$L1_{ori}$	36.72±0.32	0.9324±0.0233	35.03±1.23	0.9323±0.0212
$\mathcal{L}_1$	36.32±1.59	0.9496±0.0127	34.33±1.74	0.9545±0.0164
$\mathcal{L}_1 + \mathcal{L}_2$	36.08±1.53	0.9524±0.0128	35.38±1.87	0.9589±0.0160
$\mathcal{L}_1 + \mathcal{L}_2 + \mathcal{G}$	**38.02±1.60**	**0.9628±0.0095**	**38.12±2.10**	**0.9782±0.0092**

Table 1 displays the PSNR and SSIM values when comparing a clean image with its corresponding adversarial example through various targeted and non-targeted attack methods. Our method achieves the SOTA performance in both PSNR and SSIM. Besides, the Ablation experiments shown in Table 1 demonstrate the effectiveness of our loss setting for perceptual quality. For instance, by gradually adding $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{G}$ into the perceptual loss, the PSNR and SSIM of our method have been improved. Our method achieves the SOTA when all $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{G}$ are combined together. To better evaluate the disparity between the $\mathcal{L}_1$ loss and the original L_1 loss which denoted as $L1_{ori}$, we have the comparasion included an ablation experiment. Similarly, we also added the comparison between the original $\mathcal{L}_2$ loss and the original L_2 loss which denoted as $L2_{ori}$.

Subjective viewing test

Eight images are randomly selected from the VGGFace2 dataset for the subjective viewing test. The results of non-targeted and targeted attacks are exhibited in Tables 2 and 3. A positive (or negative) value of 'mean' represents that the adversarial example generated with our method has higher (or lower) perceptual quality than that of the anchors. Besides, the larger value of "mean" represents better perceptual quality. All the positive average values of 'mean' demonstrate that the adversarial examples generated by our method are of higher perceptual quality than those generated by the anchor methods. To prove the generalizability of our proposed method, we also selected eight images from the ImageNet dataset [10] for the subjective viewing test. Similar results are shown in Tables 4 and 5.

Conclusion

In this work, an HVS-inspired adversarial example generation method with high perceptual quality is proposed. Specifically, a JND-based perceptual loss has been proposed by taking three quality-related factors to account for the constraint of the JND thresholds. Besides, we designed a perturbation adjustment strategy to adjust the distribution among different color channels. All these designs above made the perturbation can be tolerated by the HVS as much as possible and demonstrated high perceptual quality. Ablation experiments have demonstrated the reasonability of the proposed JND-based perceptual loss. After wide experiment comparations, the proposed method has achieved the SOTA performance in subjective and objective evaluation.

It should be mentioned that our adversarial examples are iteratively generated under the constraint of the proposed JND-based perceptual loss and adversarial

Table 2 Comparison Of Subjective Viewing Tests Conducted Between The Proposed Method And The Anchor method Using The VGGFace2 Dataset Under Non-targeted Attacks

Image Index	Non-Targeted Attack					
	BIM [21]	PGD [25]	MIFGSM [26]	DI^{2}FGSM [20]	JNDMGA [28]	MND [29]
P9	1.84	1.76	2.02	2.23	2.05	1.60
P10	1.94	1.96	1.74	1.58	1.97	1.55
P11	1.12	1.47	1.64	1.38	0.69	1.17
P12	0.92	0.88	1.42	1.58	1.51	1.11
P13	1.08	1.10	1.42	1.30	1.44	1.25
P14	1.59	1.78	1.68	1.40	1.15	1.36
P15	1.22	1.10	1.32	1.38	0.51	0.96
P16	0.92	0.84	1.51	1.13	1.36	0.87
Average	**1.33**	**1.36**	**1.59**	**1.50**	**1.34**	**1.23**

Table 3 Comparison Of Subjective Viewing Tests Conducted Between The Proposed Method And The Anchor method Using The VGGFace2 Dataset Under Targeted Attacks

Image Index	Targeted Attack					
	BIM [21]	PGD [25]	MIFGSM [26]	DI^{2}FGSM [20]	JNDMGA [28]	MND [29]
P9	1.98	2.27	2.14	2.03	2.13	2.10
P10	1.18	1.93	1.86	1.95	0.44	1.53
P11	1.93	1.82	1.41	1.55	0.51	1.27
P12	1.43	1.45	1.37	1.28	1.00	1.06
P13	1.42	1.38	1.14	1.18	0.56	1.18
P14	1.67	1.62	1.71	1.83	-0.23	1.31
P15	1.10	1.28	1.12	1.05	0.41	0.76
P16	0.98	1.30	1.25	1.23	0.41	0.67
Average	**1.46**	**1.63**	**1.50**	**1.51**	**0.65**	**1.24**

Table 4 Comparison Of Subjective Viewing Tests Conducted Between The Proposed Method And The Anchor method Using The ImageNet Dataset Under Non-targeted Attacks

Image Index	Non-Targeted Attack					
	BIM [21]	PGD [25]	MIFGSM [26]	DI^{2}FGSM [20]	JNDMGA [28]	MND [29]
P12	1.87	2.20	1.90	1.97	1.35	1.54
P13	2.14	1.43	1.73	1.94	2.23	1.31
P14	1.31	1.21	1.51	1.09	1.13	0.87
P15	1.53	1.66	1.62	1.09	1.42	0.94
P16	1.88	0.98	1.33	1.06	0.87	0.76
P17	0.76	1.14	1.10	1.69	0.98	0.65
P18	1.13	1.21	0.74	1.34	0.99	1.02
P19	1.42	0.87	1.25	0.83	1.08	1.43
Average	**1.51**	**1.34**	**1.40**	**1.38**	**1.26**	**1.01**

Table 5 Comparison Of Subjective Viewing Tests Conducted Between The Proposed Method And The Anchor method Using The ImageNet Dataset Under Targeted Attacks

Image Index	Targeted Attack					
	BIM [21]	PGD [25]	MIFGSM [26]	DI^{2}FGSM [20]	JNDMGA [28]	MND [29]
P12	2.04	1.82	1.88	1.76	2.27	1.54
P13	1.92	1.86	1.83	1.96	1.42	1.31
P14	1.88	1.05	1.21	1.47	1.15	0.87
P15	1.27	0.91	1.54	0.88	1.38	0.94
P16	1.58	1.05	1.33	1.10	1.35	0.76
P17	1.73	1.64	1.13	1.78	1.46	0.65
P18	1.15	1.23	0.83	1.10	1.35	1.02
P19	1.42	0.86	1.33	0.84	1.08	1.43
Average	**1.62**	**1.30**	**1.39**	**1.36**	**1.43**	**1.07**

loss. On the one hand, the iterative generation allows us to achieve a high attack rate. On the other hand, the iterative method reduces the efficiency of adversarial examples generation compared with methods without iterations. Hence, we will focus on improving efficiency while keeping high perceptual quality in our future work.

Authors' contributions
Yuan Xue is the first author, who took the idea of this paper, did related experiments, and wrote this paper, Jian Jin is the instructor and corresponding author of this paper and gave the main suggestion on this idea. Wen Sun did part of the coding work, Weisi Lin is the professor and gave the overall instruction for this paper.

Funding
This work was supported by the National Natural Science Foundation of China (No. 62120106009).

Availability of data and materials
The corresponding author may provide the supporting data on request.

Declarations

Ethics approval and consent to participate
The authors declare that the study is applicable for both human and/or animal studies.

Competing interests
The authors declare no competing interests.

Received: 2 May 2023 Accepted: 1 June 2023
Published online: 13 June 2023

References

1. Ning X, Tian W, Yu Z, Li W, Bai X, Wang Y (2022) Hcfnn: high-order coverage function neural network for image classification. Pattern Recognit 131:108873
2. Bai X, Zhou J, Ning X, et al (2022) 3D data computation and visualization. Displays: 102169
3. Zhang P, Zhou L, Bai X, Wang C, Zhou J, Zhang L, Zheng J (2022) Learning multi-view visual correspondences with self-supervision. Displays 72(102):160
4. Tang L, Hui Y, Yang H, Zhao Y, Tian C (2023) Medical image fusion quality assessment based on conditional generative adversarial network. Multimodal Brain Image Fusion: Methods Eval Appl 16648714:54
5. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment, vol 2. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, IEEE, p 1398–1402
6. Zhang L, Zhang L, Mou X, Zhang D (2011) Fsim: A feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386
7. Zhang W, Ma K, Yan J, Deng D, Wang Z (2018) Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Trans Circ Syst Video Technol 30(1):36–47
8. Ding K, Ma K, Wang S, Simoncelli EP (2020) Image quality assessment: Unifying structure and texture similarity. arXiv preprint arXiv:2004.07728
9. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognizing faces across pose and age. In: 13th IEEE international conference on automatic face & gesture recognition (FG 2018), IEEE, p 67–74.
10. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, p 248–255
11. Yu W, Liang F, He X, Hatcher WG, Lu C, Lin J, Yang X (2017) A survey on the edge computing for the internet of things. IEEE Access 6:6900–6919
12. Chaopeng G, Zhengqing L, Jie S (2023) A privacy protection approach in edge-computing based on maximized DNN partition strategy with energy saving. J Cloud Comput 12(1):1–16
13. Peng K, Liu P, Tao P, Huang Q (2021) Security-aware computation offloading for mobile edge computing-enabled smart city. J Cloud Comput 10(1):47
14. Liu Y, Wu H, Rezaee K, Khosravi MR, Khalaf OI, Khan AA, Ramesh D, Qi L (2022) Interaction-enhanced and time-aware graph convolutional network for successive point-of-interest recommendation in traveling enterprises. IEEE Trans Ind Inform 19(1):635–643
15. Tange K, De Donno M, Fafoutis X, Dragoni N (2020) A systematic survey of industrial internet of things security: Requirements and fog computing opportunities. IEEE Commun Surv Tutor 22(4):2489–2520
16. Dash S, Biswas S, Banerjee D, Rahman AU (2019) Edge and fog computing in healthcare-a review. Scalable Comput: Pract Experience 20(2):191–206
17. Perera C, Qin Y, Estrella JC, Reiff-Marganiec S, Vasilakos AV (2017) Fog computing for sustainable smart cities: A survey. ACM Comput Surv (CSUR) 50(3):1–43
18. Liu Y, Li D, Wan S, Wang F, Dou W, Xu X, Li S, Ma R, Qi L (2022) A long short-term memory-based model for greenhouse climate prediction. Int J Intell Syst 37(1):135–151
19. Qi L, Liu Y, Zhang Y, Xu X, Bilal M, Song H (2022) Privacy-aware point-of-interest category recommendation in internet of things. IEEE Internet Things J 9(21):21398–21408
20. Xie C, Zhang Z, Zhou Y, Bai S, Wang J, Ren Z, Yuille AL (2019) Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, p 2730–2739
21. Kurakin A, Goodfellow IJ, Bengio S (2018) Adversarial examples in the physical world. In: Artificial intelligence safety and security, Chapman and Hall/CRC, pp 99–112
22. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199
23. Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
24. Rozsa A, Rudd EM, Boult TE (2016) Adversarial diversity and hard positive generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, IEEE, p 25–32
25. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083
26. Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, Li J (2018) Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, p 9185–9193
27. Zhang Z, Qiao K, Jiang L, Wang L, Chen J, Yan B (2020) Advjnd: Generating adversarial examples with just noticeable difference. In: International Conference on Machine Learning for Cyber Security, Springer International Publishing, p 463–478
28. Akan AK, Genc MA, Vural FTY (2020) Just noticeable difference for machines to generate adversarial images. In: 2020 IEEE International Conference on Image Processing (ICIP), IEEE, p 1901–1905
29. Sun W, Jin J, Lin W (2022) Minimum noticeable difference based adversarial privacy preserving image generation. arXiv preprint arXiv:2206.08638
30. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
31. Chou CH, Li YC (1995) A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. IEEE Trans Circ Syst Video Technol 5(6):467–476
32. Wu J, Shi G, Lin W, Liu A, Qi F (2013) Just noticeable difference estimation for images with free-energy principle. IEEE Trans Multimed 15(7):1705–1710
33. Wu J, Li L, Dong W, Shi G, Lin W, Kuo CCJ (2017) Enhanced just noticeable difference model for images with pattern complexity. IEEE Trans Image Process 26(6):2682–2693
34. Shen X, Ni Z, Yang W, Zhang X, Wang S, Kwong S (2020) Just noticeable distortion profile inference: A patch-level structural visibility learning approach. IEEE Trans Image Process 30:26–38
35. Zhang Y, Liu H, Yang Y, Fan X, Kwong S, Kuo CJ (2021) Deep learning based just noticeable difference and perceptual quality prediction models for compressed video. IEEE Trans Circ Syst Video Technol 32(3):1197–1212
36. Wu Y, Ji W, Wu J (2020) Unsupervised deep learning for just noticeable difference estimation. In: IEEE International Conference on Multimedia & Expo Workshops, IEEE, p 1–6
37. Jin J, Yu D, Lin W, Meng L, Wang H, Zhang H (2022a) Full RGB just noticeable difference (JND) modelling. arXiv preprint arXiv:2203.00629

38. Jin J, Xue Y, Zhang X, Meng L, Zhao Y, Lin W (2022b) Hvs-inspired signal degradation network for just noticeable difference estimation. arXiv preprint arXiv:2208.07583
39. Yang X, Lin W, Lu Z, Ong E, Yao S (2005) Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile. IEEE Trans Circ Syst Video Technol 15(6):742–752
40. Liu H, Zhang Y, Zhang H, Fan C, Kwong S, Kuo CCJ, Fan X (2019) Deep learning-based picture-wise just noticeable distortion prediction model for image compression. IEEE Trans Image Process 29:641–656
41. Tian T, Wang H, Kwong S, Kuo CCJ (2021) Perceptual image compression with block-level just noticeable difference prediction. ACM Trans Multimed Comput Commun Appl (TOMM) 16(4):1–15
42. Jin L, Lin JY, Hu S, Wang H, Wang P, Katsavounidis I, Aaron A (2016) Kuo CCJ (2016) Statistical study on perceived jpeg image quality via mcl-jci dataset construction and analysis. Electron Imaging 13:1–9
43. Wang H, Gan W, Hu S, Lin JY, Jin L, Song L, Wang P, Katsavounidis I, Aaron A, Kuo CCJ (2016) Mcl-jcv: a JND-based h. 264/avc video quality assessment dataset. In: IEEE International Conference on Image Processing, IEEE, p 1509–1513
44. Wang H, Katsavounidis I, Zhou J, Park J, Lei S, Zhou X, Pun MO, Jin X, Wang R, Wang X et al (2017) Videoset: A large-scale compressed video quality dataset based on JND measurement. J Vis Commun Image Represent 46:292–302
45. Liu X, Chen Z, Wang X, Jiang J, Kowng S (2018) JND-pano: Database for just noticeable difference of jpeg compressed panoramic images. In: Pacific Rim Conference on Multimedia, IEEE, p 458–468
46. Lin H, Chen G, Jenadeleh M, Hosu V, Reips UD, Hamzaoui R, Saupe D (2022) Large-scale crowdsourced subjective assessment of picture-wise just noticeable difference. IEEE Trans Circuits Syst Video Technol 32(9):5859–5873
47. Jin J, Zhang X, Fu X, Zhang H, Lin W, Lou J, Zhao Y (2021) Just noticeable difference for deep machine vision. IEEE Trans Circ Syst Video Technol
48. Kanopoulos N, Vasanthavada N, Baker RL (1988) Design of an image edge detection filter using the Sobel operator. IEEE J Solid-State Circ 23(2):358–367
49. BT RIR (2002) Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union

Publisher's Note