Information Sciences 569 (2021) 287–301

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# *CDAE*: Color decomposition-based adversarial examples for screen devices

Huanyu Bian [a], Hao Cui [a], Kunlin Liu [a], Hang Zhou [a], Dongdong Chen [b,*], Wenbo Zhou [a], Weiming Zhang [a,*], Nenghai Yu [a]

[a] University of Science and Technology of China, China
[b] Microsoft Cloud AI, United States of America

ARTICLE INFO

*Article history:*
Received 15 July 2020
Received in revised form 28 February 2021
Accepted 1 April 2021
Available online 14 April 2021

*Keywords:*
Adversarial example
Color decomposition
Action recognition

ABSTRACT

Adversarial examples can easily fool existing powerful deep neural networks. However, we find that the attack ability of most existing adversarial attack methods is significantly degraded once the generated adversarial examples are shown on screen devices and are further captured. This is mainly attributed to two challenges: (1) Extra noises and variance during the capturing process, such as lens distortion and diverse capturing angle. (2) They get stuck in a self-contradictory problem between visual quality and attack ability. Inspired by the properties of the human visual system (HVS), this paper dedicatedly designs the first color decomposition-based adversarial example method ***CDAE*** for screen devices. Specifically, it decomposes one regular screen frame into two symmetric adversarial frames with maximum modifications while theoretically guaranteeing the visual quality perceived by human observers. Thanks to the powerful generalization ability of the proposed method, we can combine it with most adversarial example generation methods and achieve state-of-the-art attack ability. Additionally, it can also be used to protect important information from leakage and attack existing video action recognition networks.


## 1. Introduction

Screens are becoming some of the most common display devices in many application scenarios, such as personal computers, cellphones, cinemas, and traffic electronic signs. Essentially, screens show one image at a time but are refreshed at a high frequency. In this sense, screen content can be regarded as video streaming. Recently, adversarial example attacks have become a hot research topic, and many works [4,6,14,19,28,37] have been proposed for different image-based or video-based applications. There have been some successful physical-world attacks [12,13], including viewpoints, on road sign classification under various environmental conditions, which pose threats to the security of auto-driving systems. However, most of these works assume that their inputs are images or videos not captured from screen devices.

Capturing a screen image with a camera has become increasingly popular in real systems. Here we enumerate two very common scenarios: (1). **Auto-driving attack.** In auto-driving, electronic street signs are very common. It is possible to add adversarial perturbations on the original image displayed on an electronic screen to mislead the recognition results by the

* Corresponding authors.
*E-mail addresses:* hybian@mail.ustc.edu.cn (H. Bian), cvhc@mail.ustc.edu.cn (H. Cui), lkl6949@mail.ustc.edu.cn (K. Liu), zh2991@mail.ustc.edu.cn (H. Zhou), welbeckz@ustc.edu.cn (W. Zhou), zhangwm@ustc.edu.cn (W. Zhang), ynh@ustc.edu.cn (N. Yu).

https://doi.org/10.1016/j.ins.2021.04.005
0020-0255/

auto-driving systems. Adding adversarial perturbations, which maintain the original visual quality of an image when perceived by human observers, can have serious consequences for an auto-driving system that captures images on an electronic screen. (2). **Leak prevention.** In a confidential environment, hardware isolation is often used to prevent the leaking of confidential information. At this time it is difficult to steal confidential documents with traditional methods such as USB flash drives or sending emails. However, commercial spies (usually authorized employees) can steal a large amount of confidential information by simply opening many confidential documents on a screen and capturing images with a portable camera, such as the camera on a mobile phone, using optical character recognition (OCR) to obtain the documents from the pictures, and then directly transferring the documents to steal the confidential information. We believe that in the process of stealing, commercial spies can reduce the capacity of the transmitted information as much as possible. Compared with the direct transmission of high-definition pictures, when OCR is used on pictures to obtain text information and then to obtain the secret information, the chance of the thieves being caught is greatly reduced. This behavior has occurred many times in reality and is inevitable. To eliminate this security problem, we can add adversarial perturbations when displaying confidential documents on a screen, which blocks deep learning-based OCR systems from obtaining the right information from screen-captured pictures and thus prevents information leaks.

Since capturing screens with cameras has become increasingly popular in real systems, we attempt to address the adversarial attack problem for images captured from screen devices. As shown in Fig. 1, this is much more challenging from two aspects: (1) **More complex distortions.** The screen capturing process introduces extra noises and variances, such as lens distortion, moire pattern distortion, light source distortion and diverse capturing angle, which make most current adversarial attack methods fail. (2) **Easily perceived by HVS.** On the one hand, small modifications generated by current adversarial attack methods are still easily perceived by the human visual system (HVS), which makes them not applicable for real applications. On the other hand, to maintain the attack ability during the capturing process, current methods need to increase the amount of modification as much as possible. Hence, this is indeed a self-contradictory problem.

Inspired by the property of HVS, we propose to use the classic color mixing principle to solve the above self-contradictory problem. In other words, when the alternate flickering frequency of two colors is higher than the critical flicker frequency (CFF), human observers will perceive one fused color. Conversely, an original frame of a picture can be decomposed into two different frames, which are generated by adding and subtracting the same amount of modification from the original frame respectively. In this way, once the flickering frequency is high enough (over CFF), these two frames can theoretically be perceived by HVS to be the same as the original frame even when the modification amount is very large. In this paper, we refer to this property as the "color decomposition" property.

Based on this motivation, we successfully circumvent the above self-contrad-ictory problem in the existing methods and propose the first screen adversarial example generation framework: Color Decomposition based Adversarial Examples (**CDAE**). As shown in Fig. 2, we decompose one regular screen frame into two frames, as described above, while guaranteeing that they are both adversarial examples with the same amount of modification. However, different from previous adversarial methods, we choose to maximize but not minimize the modification amount so that the generated adversarial examples can still retain the attack ability after the screen capturing process. Thanks to the aforementioned color decomposition property, the adversarial examples generated by our method can perfectly maintain the original visual quality when perceived by human observers.

Our proposed algorithm can be used in scenarios with a single frame and a single photo. The essence of the picture displayed on the screen is to reproduce a picture continuously. The only difference between a picture and a video on a display screen is that a picture maintains the same display as the previous frame and the next frame, while a video displays different images in the previous frame and the next frame. On common display screens, the display of pictures is essentially the same


**Fig. 1.** Limitations of existing adversarial example generation methods after screen capture. The left picture is a clean image, and the middle picture is an image with added adversarial noise, and it was misclassified. However, the right picture shows that after screen capture, the attack ability of adversarial noise is destroyed.

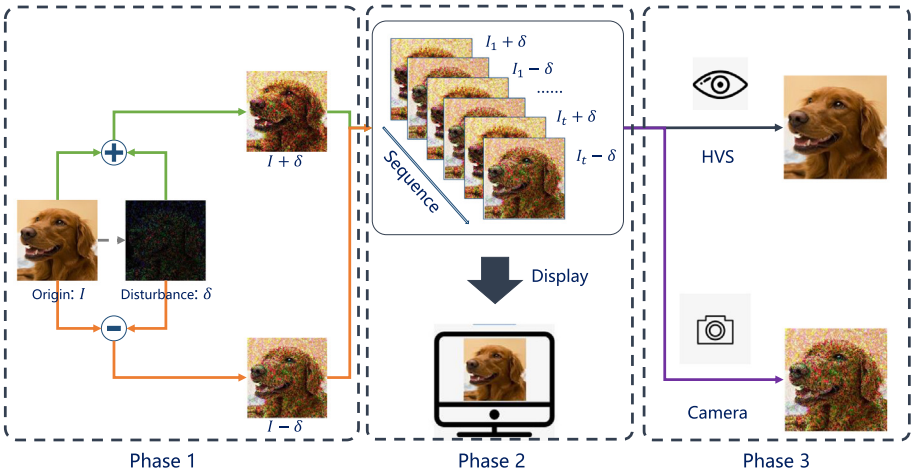


**Fig. 2.** The overall working principle of the proposed CDAE framework for screen devices. The CDAE method decomposes an original image $I$ to two symmetric adversarial examples by adding and subtracting perturbation $\delta$ respectively. Thanks to the color mixing principle, the human observer (HVS) can perceive the original image, while modern cameras can capture noisy adversarial examples.

as that of videos, that is, a single picture is played repeatedly in a loop with the display refresh frequency as the playback frequency. Due to the fusion characteristics of the HVS, when the display refresh frequency exceeds CFF, the human visual system only observes a single picture. The proposed CDAE can generate screen adversarial examples that are imperceptible to the human eye by making each picture into one adversarial example pair.

Another advantage of our method is its powerful generalization ability. Experiments show that **CDAE** method can be combined with most existing adversarial example methods and significantly improves their robustness and attack ability during the screen capturing process. As a side benefit, since the shutter frequency of modern cameras is generally much higher than CFF and does not satisfy the color decomposition property, our method can naturally protect important information from illegal screenshots or video recordings. This is because, if we apply our **CDAE** to the screen frames, the content captured by camera devices is not the fused content and will be very noisy. Moreover, it is also able to attack the following recognition systems if they exist.

For the leak prevention scenario, if commercial spies have sufficient time and tools, it is indeed possible to perform manual conversion regardless of the time and money required, however, the text adversarial examples we generated make distinguish detailed information from pictures difficult for the human visual system. Even if much manpower is invested, it is difficult to obtain complete and correct information. Therefore, our proposed method is effective in preventing information leaks.

We further demonstrate that the proposed **CDAE** method can be used to generate the first video action recognition adversarial examples against the Two-Stream Network [32]. By converting the original frames into adversarial frame pairs for the spatial stream network, we find that it can simultaneously be used to attack the spatial stream and flow stream networks, and then cause the final classification network to definitely fail. The experiments also show that the method achieves good performance.

In summary, our main contributions are threefold as follows:

- Based on the color decomposition property, we proposed the first screen adversarial example generation framework, called **CDAE**, which can theoretically guarantee the visual quality and attack ability of images captured from screen devices simultaneously.
- We demonstrated the generalization ability of the proposed **CDAE** method. It can be combined with most existing adversarial methods and achieves state-of-the-art attack abilities. Additionally, it can also naturally protect important information from leakages and attacks existing video action recognition networks as side benefits.
- We further proposed the first video action recognition adversarial examples against the Two-Stream Network [32] with **CDAE** and achieved very good attack performance in the experiment.

The rest of the paper is organized as follows. In Section 2, we describe the works most closely related to our proposed method. In Section 3, we propose our CDAE method, which can theoretically guarantee the visual quality and attack ability of images captured from screen devices. Section 4 shows the experimental setup and experimental results of our CDAE method compared with other state-of-the-art methods. In Section 5, we mainly discuss the influence of different parameters on our proposed CDAE method. Section 6 draws the conclusion.

## 2. Related work

### 2.1. Digital adversarial examples

L-BFGS attack [37], which is named by the optimization method they used, was the first optimization-based adversarial example generation method. Its main idea is to transform the adversarial example generation problem into an optimization problem and solve it with the L-BFGS optimization method. To further improve the attack ability, Carlini et al. [6] proposed C&W attack, which is named after initial letters of the authors' names, to follow the path of L-BFGS attack. They design a series of loss functions that have a smaller value for adversarial examples but a larger value for clean images so that adversarial examples can be searched by minimizing the loss function. However, unlike the L-BFGS method, this method uses the Adam algorithm instead of L-BFGS to solve this optimization problem and improve the speed.

Goodfellow et al. [14] pioneered the fast gradient sign function method (FGSM), which creates perturbations with symbols relative to the input loss gradient and uses back propagation to efficiently calculate the required gradient. It has the advantages of few calculations and fast calculation speeds and the disadvantage of poor attack ability. The Basic Iterative Method (BIM) [19], sometimes called iterative FGSM (I-FGSM), is an improvement of FGSM with a small step size and multiple iterations. It has a slightly higher computational cost but greatly improves the attack success rate. Madry et al. [27] further found that BIM could be significantly improved by starting with random points. This type of attack is often referred to as PGD.

Although the above two types of methods have very good attack abilities, they require a large amount of calculations due to their inherent iterative procedures. To solve this problem, Baluja et al. [4] directly trained a generation model to transform input images to adversarial examples. This kind of method can be viewed as an accelerated version of optimization-based and gradient-based methods. By training on a large number of images with the pre-defined attacking objective, these models do not need to access the target model again and can generate adversarial examples with only a fast forward pass during the runtime, so they are very fast. GAP [30] is one of the classic generation-based methods. Other representative algorithms include [48,10,15,44,47].

Although all the above methods can attack recognition models for regular images, their attack ability is significantly degraded for the images captured from screen devices because of the extra noises and variances added during the capturing process. Another problem is that the visual quality of the input images is sacrificed.

### 2.2. Temporal image modulating

Temporary image Modulating is a broad and classical research theory that has been studied for more than 60 years. [31] is a representative work proposed in 1952, which summarized its theoretical background and potential applications. One of the key observations is that when two isoluminant colors alternate at sufficiently high frequencies, the human observer typically perceives only the fused color in an additive way. Based on this theory, many great works [42,46,1] have been proposed in the past. VRCodes [42] was a novel bar code design based on color decomposition, and can be read by cameras. However, it is still unobtrusive to the HVS. Kaleido [46] proposed a color decomposition-based video encoding and display system that does not affect HVS but prevents opponents from using the camera or smart device to record HD video. Abe et al. [1] designed a system based on color decomposition to embed five different data values using imperceptible color vibrations.

Our work is also motivated and based on this great theory. However, the direct application of this theory to our task is non-trivial. First, it can only guarantee the attacking ability of only one frame rather than two adjacent symmetric frames. Second, most previous methods [46,1] consider Lab and XYZ color spaces for color decomposition, which have to be pre-transformed from the original RGB color space. By contrast, we consider RGB color space decomposition instead. This is advantageous because it avoids complex color difference formulas such as the CIEDE2000 [26] color difference, which may cause less chromatic aberration and higher computational efficiency. Third, to make our method general to different types of adversarial attacking methods, we have dedicatedly designed different types of algorithms, which are shown in the following parts, for different methods.

### 2.3. Physical-world adversarial examples

Kurakin et al. [19] first demonstrated that adversarial attacks can be implemented in the physical world. They printed adversarial examples on papers and then used cameras to capture pictures on the papers. However, subsequent research [3] found that due to different illumination and angles, camera noise, and other transformations, it is difficult to generate successful adversarial examples in the physical world. Therefore, to enhance the adaptability of the methods to natural transformations such as illumination, rotation and camera noise, powerful physical-world adversarial examples require larger perturbations. Adversarial Patch [5] uses large perturbations to enhance the robustness and generalization ability of the adversarial examples. By generating a visible fixed-size adversarial patch, adversarial examples can be generated by adding this patch at any position on the original image. Cameras were used to capture pictures and these special pictures were fed to deep learning classifiers. This shows that even if captured photos contain the noise from the physical world, a large portion of adversarial examples will still be misclassified. A similar idea was also used by Evtimov et al. [12] to attack traffic signs. By

pasting stickers onto a sign to create disturbances, a classifier can be attacked from a different distance at different angles. Other representative algorithms include [12,23,24].

Though these methods can generate physical adversarial examples, (1) **Application scenario.** they cannot be directly applied to screen devices, and (2) **Easily perceived.** to increase the attack ability, all of these physical-world adversarial examples generate large perturbations, which inevitably cause these existing adversarial examples to have distortions that can be easily perceived by HVS and then removed by human observers.

### 2.4. Video adversarial examples

In addition to the adversarial examples designed for image classifiers, there are some attack methods designed for video classification systems. For example, Hosseini et al. [17] first proposed an attack method for video classifiers and exploited the vulnerability in the Google video classification API. Because the underlying API algorithm only processes the first frame of a video every second, by inserting an attack image into the video each second, the API would only output the tags associated with the inserted attack image. However, this method is not universal. Li et al. [21] then proposed the first universal adversarial example generation method for video classification. They proposed two different types of perturbations and achieved an 80% attack success rate on the C3D classification network [39] with the UCF-101 dataset.

However, to the best of our knowledge, there is no attack method designed for the Two-Stream Network [32] yet.

## 3. Method

### 3.1. Motivation

Since extra noise is often introduced during the image/video capturing process, the attack ability of adversarial examples generated by most existing methods degrades considerably when images are captured from screen devices. This is because these methods are often designed to minimize adversarial perturbations to ensure visual quality. However, small perturbations are be easily overwhelmed by noise and variance captured by screens. That is, most existing methods face a serious self-contradictory problem between visual quality and attack ability.

Inspired by the color mixing principle of the human vision system, we adopt a totally different idea and circumvent the above contradiction successfully. This principle [9] can be summarized as **"When the alternating flicker frequency of two colors is above the CFF, the human observers can only perceive their fused color.**" In the RGB color space, it can be formally formulated as:

$$\begin{cases} R = (R_1 + R_2)/2 \\ G = (G_1 + G_2)/2 \\ B = (B_1 + B_2)/2 \end{cases} \tag{1}$$

where $(R_1, G_1, B_1)$ and $(R_2, G_2, B_2)$ are the original two colors in the RGB color space respectively, and $R, G, B$ is the fused color perceived by human observers. **CFF** represents the fluctuation frequency above which human eyes cannot perceive flicker [2]. Notably, by real human visual tests, we find that this color mixing principle is perfectly valid when the illuminance change is under some threshold; otherwise, some flickering effects will appear.

For an adversarial example, we consider the dual problem of the color mixing principle, i.e., **"color decomposition"**. In detail, given an original frame image $I$, we can decompose it into two frames $\{I - \delta, I + \delta\}$, where $\delta$ represents the symmetric modification value. Regardless of how large $\delta$ is, as long as these two frames are alternatively shown at a frequency above the CFF threshold (satisfied by most screen devices), humans can only perceive the original frame $I$. In this paper, we refer to this property as the "color decomposition property". Recalling that all current adversarial attack methods are based on adding/-subtracting some adversarial perturbations onto/from the original input image, a natural idea would be to incorporate the "color decomposition" idea into the adversarial example generation process. In fact, if we can guarantee that $I - \delta$and $I + \delta$ are both adversarial examples, even if $\delta$ is large, humans can still perceive the original frame $I$, thus theoretically preserving the original visual quality. Another advantage of a large $\delta$ is that it can help to maintain the attack ability during the screen capturing process.

### 3.2. CDAE: Color decomposition-based adversarial examples

Formally, given an original frame $I$, previous methods search for an optimal and minimal $\delta$ to maintain $I + \delta$'s attack ability. By contrast, as shown in Fig. 2, we aim to generate two symmetric adversarial examples $I - \delta, I + \delta$. Thanks to the color decomposition property, we do not need to minimize $\delta$ to ensure the visual quality but maximize it to retain the attack ability after screen capture. Although our method can be incorporated with different types of adversarial attack methods, we first formulate it as an optimization problem with the two basic constraints below:

- *Adversarial constraints:* $I - \delta$ and $I + \delta$ are both adversarial examples that can attack a given recognition model.
- *Boundary constraints:* The pixel values of $I - \delta$ and $I + \delta$ must be in the range of 0 to 255.

The proposed optimization problem can be formulated as:

$$\begin{aligned} \max_{\delta} \quad & D(I, I+\delta) + D(I, I-\delta) \\ \text{s.t.} \quad & f(I+\delta) \neq f(I) \\ & f(I-\delta) \neq f(I) \\ & I+\delta \in [0,255]^n \\ & I-\delta \in [0,255]^n \end{aligned} \tag{2}$$

where $D$ is the distance metric that can be $L_0, L_2$ or $L_\infty$; $f$ is the target recognition model; and $I$ is fixed. The objective is to find the value of $\delta$ that maximizes the distances $D(I, I+\delta)$ and $D(I, I-\delta)$ while ensuring that $f(I-\delta)$ and $f(I+\delta)$ have classification labels that are different from the labels of origin $f(I)$.

**Illuminance Constraint.** As emphasized above, when the change in illuminance between $I-\delta$ and $I+\delta$ is above some threshold (+2, −3), humans can perceive some flickering effects. This phenomenon is also highlighted in [29]. Therefore, to guarantee that there is no impact on the visual quality, we further incorporate an extra illuminance constraint; *i.e.*:

$$\begin{aligned} & Y = 0.222485R + 0.716905G + 0.060610B \\ & Y(I-\delta) - Y(I+\delta) <= \Delta Y \end{aligned} \tag{3}$$

To demonstrate the generalization ability of our method, we combine our color decomposition-based adversarial example generation idea with different types of adversarial attack methods. For gradient-based methods, we select I-FGSM [19] as the basic algorithm. For generation-based methods, we design our generation-based CDAE. For the optimization-based method [6], because it is too time-consuming, we do not have enough computational resources to perform massive experiments, so we ignore it in this part. However, we have to emphasize that our method can support it in principle.

**Gradient-based CDAE.** Directly solving the above objective function defined in Eq. (2) together with the illuminance constraint in Eq. (3) is a non-trivial problem, because $f$ is often a very complex recognition model without explicit formulation and Eq. (3) is a hard constraint. Therefore, we resort to an alternative optimization algorithm to approximate its solution instead.

I-FGSM [19] is an adversarial example generation method that increases the attack capability by accumulating modifications in the direction of the gradient during the iteration process. I-FGSM seeks an adversarial example by solving the constrained optimization problem [11]:

$$\arg\max_{\delta} \quad J(I+\delta, x), \quad \text{s.t.} \quad \|\delta\|_\infty \leqslant \epsilon \tag{4}$$

where $x$ is the ground-truth label, $J(I+\delta, x)$ is the loss function, and $\epsilon$ is the size of the adversarial perturbation. We use multiple iterations to help achieve a stronger attack capability and robustness with the same modification amount. We apply the idea of iteration to generate our gradient-based CDAE method and obtain great benefits.

To generate a pair of adversarial examples $I+\delta$ and $I-\delta$ from a real example $I$, which satisfies our CDAE bounds, gradient-based CDAE seeks the adversarial example pair by changing Eq. (4) to the following optimization problem:

$$\begin{aligned} \arg\max_{I+\delta, I-\delta} \quad & J(I+\delta, x) + J(I-\delta, x) \\ \text{s.t.} \quad & I+\delta \in [0,255]^n \\ & I-\delta \in [0,255]^n \\ & \|\delta\|_\infty \leqslant \epsilon \\ & Y(I- \quad \delta) - Y(I+\delta) <= \Delta Y \end{aligned} \tag{5}$$

where we set $\epsilon$ as large as possible to ensure the maximum modification.

**Algorithm 1:** Color decomposition-based adversarial example (***CDAE***) generation method

**Symbols:** original frame **I**, iteration number **n**, illuminance constraint $\Delta$**Y**, pixel position $(i,j)$, illuminance **Y**, specific adversarial attack method **GenAdv**

**Output**: $\delta$

Initialize $\delta^0 = 0$

**for** $k \leftarrow 1$ to **n do**:

$I_-^k := I - \delta^{k-1}$ → get frame one

$\delta^k := \textbf{GenAdv}(I_-^k)$ → adversarial perturbation of $I_-^k$

$I_+^k := I + \delta^k$ → get frame two

$\delta^k := \textbf{GenAdv}(I_+^k)$ → adversarial perturbation of $I_+^k$

$I_-^k := I - \delta^k$ → update frame one

$I_+^k := I + \delta^k$ → update frame two
**for** pixels $(i,j)$ that satisfy $\mathbf{Y}(I_+^k) - \mathbf{Y}(I_-^k) \notin [-\Delta\mathbf{Y}, \Delta\mathbf{Y}]$ **do**
$\delta^k[i,j] := \delta^{k-1}[i,j]$ → rollback to last $\delta$
**return** $\delta^n$ as the final $\delta$

As shown in Algorithm 1, we first initialize $\delta$ with zeros, then for each iteration, we alternatively obtain the adversarial perturbation of the first decomposed frame $I - \delta$ and the second decomposed frame $I + \delta$ respectively. To ensure that the change in illuminance between $(I - \delta, I + \delta)$ is always below threshold $\Delta Y$, we roll back to the last $\delta$ values for positions that do not satisfy this illuminance constraint.

The experiments show that the adversarial perturbation $\delta$ first increased during the beginning stage and then converges to stable values after enough iterations. By alternatively obtaining the adversarial perturbations for the two decomposed frames, this algorithm can converge well and guarantee $I - \delta$ and $I + \delta$ both adversarial examples.

### 3.3. Generation-based CDAE

Since the above gradient-based CDAE requires an alternative optimization procedure, such as the underlying gradient-based adversarial attacking methods, its computational cost is relatively high. To speed it up and make it applicable in real-time applications, we further propose the generation-based CDAE method.

Similar to many existing end-to-end image generation models based on deep learning, the proposed generation-based CDAE uses the basic architecture of a convolutional encoder-decoder network. Existing end-to-end image generation networks can be generally divided into two kinds: directly generate the target image and generate the residual image between the input image and the target image. For our task, unlike most existing image generation methods, which only generate one output image, we need to generate a pair of adversarial examples. If we use the encoder-decoder network to directly generate one target image, we need to train two encoder-decoder networks, which increases the training difficulty and computation time. Hence, we design a generation network that directly outputs one perturbation and then generates a pair of adversarial examples by adding and subtracting the perturbation image from the input image respectively.

Therefore, we require a mapping $G : I \to \delta$, which generates a pair of perturbed images $I + G(I), I - G(I)$ for each clean image $I$. A desirable function $G$ must result in a high fooling ratio and a large modification. Our approach is to approximate the difference in the clean and adversarial image pairs with a generation network $G$. We require that for each clean image $I : f(I + G(I)) \neq f(I)$ and $f(I - G(I)) \neq f(I)$, the $L_\infty$ norm of the additive perturbation needs to be as large as possible while maintaining the illuminance constraint. Like GAP, clean image $I$ is passed through the generator $G$ to generate one perturbation $G(I)$. The result is the perturbation $G(I)$, which is added and subtracted to the clean image. We feed the modified images $f(I + G(I))$ and $f(I - G(I))$ to the network to obtain two output probabilities. At the inference stage, we can discard the classification network and only use the generation network $G$ to generate adversarial example pairs. This avoids the need for iterative gradient computations.

As mentioned above, our goal is to develop an end-to-end generation model that directly generates adversarial example pairs with large perturbations and fits CDAE's optimization conditions. We use a generative model to directly generate adversarial example perturbations to achieve the goal of generating adversarial sample pairs. Compared with GAP, the generated adversarial example pair no longer needs to maintain appearance similarity with the original image but needs to ensure that their luminance difference is as small as possible. Therefore, we use three intuitive loss functions to optimize the generative adversarial network. The first is the classification loss $L_{cls}$, which guarantees the attack ability of the generated adversarial example pair; the second is the perturbation loss $L_{per}$, which controls the amplitude of the perturbation; and the last is the illuminance loss $L_{lum}$, which guarantees the illuminance difference between $I + G(I)$ and $I - G(I)$. To calculate the classification loss, the generated images are fed into the target classification network to obtain the predicted labels.

We call this final loss the **CDAE** loss:

$$\mathscr{L}_{CDAE} = \mathscr{L}_{cls} + \mathscr{L}_{per} + \omega\mathscr{L}_{lum} \tag{6}$$

**Classification loss.** The classification loss $\mathscr{L}_{cls}$ encourages the classification results between the generated image and the original image to be different. The detailed function formulation is defined as cross-entropy loss:

$$\mathscr{L}_{cls} = -log(p_y(I + G(I))) - log(p_y(I - G(I))) \tag{7}$$

where $p_y()$ is the probability output of the attacked model $f$ with respect to classification class $y$.

**Perturbation loss.** The above classification loss can make the generated adversarial image pair have attack capabilities; however, it cannot control the magnitude of the adversarial perturbation. The magnitude of the adversarial perturbation can be controlled by the perturbation loss $\mathscr{L}_{per}$:

$$\mathscr{L}_{per} = -\|G(I)\|_\infty \tag{8}$$

**Illuminance loss.** The illuminance loss $\mathscr{L}_{lum}$ is the relaxed version of Eq. (3) and measures the difference in illuminance between the symmetric adversarial example pair, and we directly use the simple MSE loss here:

$$\mathscr{L}_{lum} = \frac{1}{M}(Y_{I+G(I)} - Y_{I-G(I)})^2 \tag{9}$$

With these three different losses, the **CDAE** loss can ensure that the generation network generates adversarial pairs that ensure visual quality while maximizing the difference to retain the attack ability after screen capture.

### 3.4. Extension to video action recognition

Furthermore, we find that our method can be easily extended to generate adversarial examples for video action recognition. In fact, there are two main types of action recognition algorithms: (1) One treats multi-frame RGB images as 3D inputs and uses 3D convolutional networks for processing. Representative algorithms include C3D [39] and ARTNet [40]. (2) The other type is the two-stream network and its variants, which often consist of two branches fed with RGB images and optical flow images, respectively, and the recognition results of these two branches are merged. Representative methods include TSN [41], Two-Stream [32], Temporal pyramid CNNs [41], I3D [7]. To the best of our knowledge, we are the first to generate adversarial examples for the Two-Stream type action recognition network.

Specifically, as shown in the top part of Fig. 3, Two-Stream [32] has an independent spatial stream and a temporal stream, and the classification result of these two streams are combined with some weights; the result can be formulated as:

$$Result = \alpha * Score_{spatial} + \beta * Score_{temporal} \tag{10}$$

where *Result* is the classification result of the whole network. $Score_{spatial}$ and $Score_{temporal}$ are the softmax results of the spatial stream and temporal stream, respectively. $\alpha$ and $\beta$ are the coefficients of the weighted summation.

Generally, the spatial stream's weight $\alpha$ is often the same or higher than the weight of the temporal stream $\beta$ [32], which makes adversarial example generation easier for our method. If the attack on the spatial stream is strong enough, even if the classification result of the temporal stream is correct, the final result will still be incorrect.

As shown in Fig. 4, with the aforementioned color decomposed based adversarial examples method, we convert each frame of the video into two frames of adversarial examples based on the spatial stream branch. We find that this modification can destroy both the spatial stream and temporal stream. As shown in the bottom part of Fig. 3, by feeding the decomposed adversarial frames (targeted to the spatial stream) into the temporal stream, the optical flow generation method also completely fails because the newly introduced noise causes incorrect patch matching.

## 4. Experiment

In this section, we first introduce the experimental settings, and then we evaluate the visual quality of CDAE and common adversarial examples by letting the users judge by themselves. After that, we evaluate the attack capabilities of CDAE and other common adversarial attacks after screen shooting under white-box and black-box conditions and attack models with defensive capabilities. We further extend CDAE to video action recognition.

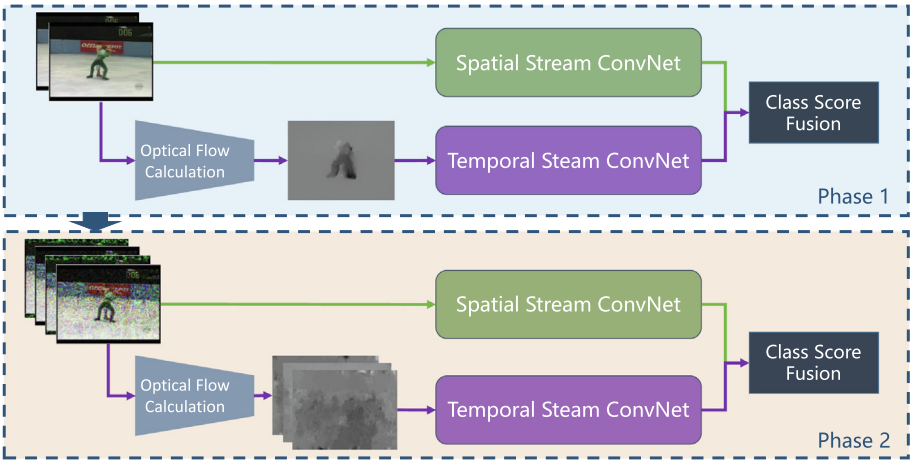


**Fig. 3. Overview of the proposed framework for generating video adversarial examples..**

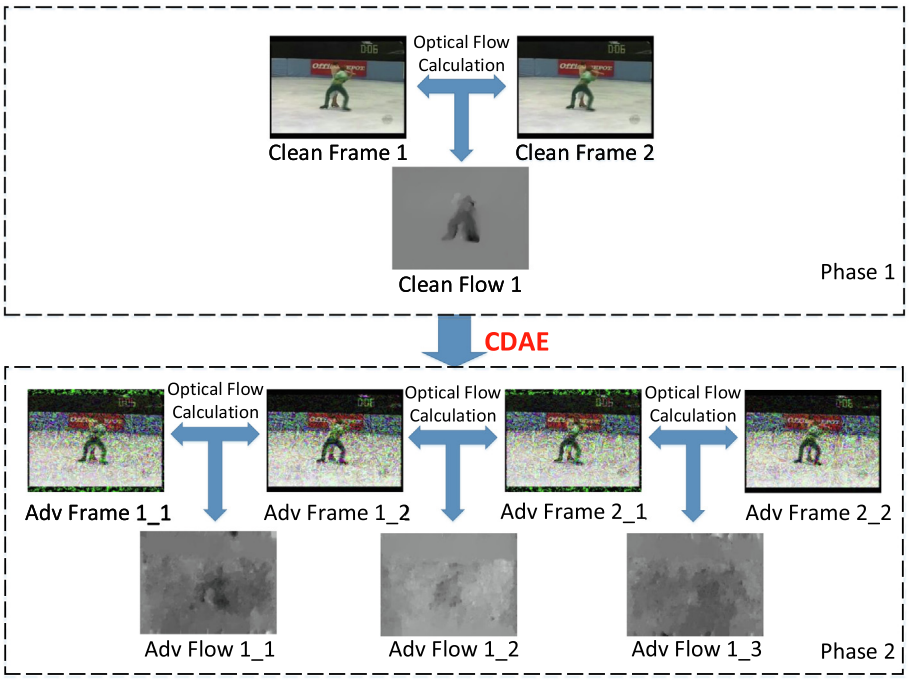


**Fig. 4. Process of generating video adversarial examples.** Phase 1 generates one optical flow frame from two adjacent natural frames. From clean images *Clean Frame* 1 and *Clean Frame* 2, we can obtain very clear motion information: *Clean Flow* 1. Phase 2 uses ***CDAE*** to change the two frames of clean images *Clean Frame* 1 and *Clean Frame* 2 into two pairs of adversarial examples: *Adv Frame* 1_1, *Adv Frame* 1_2, *Adv Frame* 2_1 and *Adv Frame* 2_2. The corresponding optical flow also becomes three: *Adv Flow* 1_1, *Adv Flow* 1_2 and *Adv Flow* 1_3 . After **CDAE**, the motion information on the optical flow is completely disturbed, as *Adv Flow* 1_1, *Adv Flow* 1_2 and *Adv Flow* 1_3 have no useful motion information.

### *4.1. Experiment setup*

**Dataset.** For visual quality comparisons, we randomly select 100 images from the ImageNet validation dataset to form an evaluation set. For white-box attack, black-box attack and the attack of adversarial training model, all results are reported on 1000 randomly selected images from the ImageNet validation dataset. For video action recognition, we generate adversarial examples on the UCF101 dataset [34], which contains 101 action classes and 13,320 video clips. Due to the limited computational resources, one-thousand video clips are randomly extracted as the test set.

**Baseline attack settings.** We compare the capabilities of a series of adversarial example generation methods, including the gradient-based method, optimization-based method and generation-based method, as baselines with CDAE's attack capabilities after screen shooting. We adopt I-FGSM, PGD, ATN, GAP under the constraint of the $L_\infty$ norm and C&W under the constraint of the $L_2$ norm. By default, we use classification accuracy as an evaluation indicator in all the experiments below.

**Threat model.** For the experiment of image adversarial examples, we generate untargeted attacks and test after the screen capturing process. In the white-box attack, the source network and the target network are the same, and four different networks are selected for experiments. In the black-box attack, the source network and the target network are different networks. We choose ResNet_101 as the source network, and the other four models are tested for the target network. When attacking the adversarial training model, we selected five very powerful adversarial training models that were published on the Internet as the target model for testing. For video action recognition adversarial examples, two of the most common two-stream networks are selected as the target networks.

**Shooting setting.** As shown in Fig. 5, all the shooting experiments are conducted in a bright indoor office environment. A commonly used office monitor is used to show all the adversarial examples in the experiment, and the camera of a commonly used smartphone is used to shoot adversarial examples on the screen. The model of the display is ViewSonic VA2465, and the resolution of the screen is 1920×1080. All image adversarial examples are displayed at a frequency of 60 Hz (over CFF and a common frequency for displays). To prevent the change in the size of the adversarial examples from affecting the attack ability, we display them on the display with their original size. The shooting device is the camera of the smartphone Xiaomi Mix2. In the experiment, the default application of the mobile phone is used to shoot the screen in automatic mode. We mounted the phone on a tripod and set the distance between the screen and the lens of the phone to 30 cm.

**Fig. 5.** Experimental settings.

### 4.2. Experiment results

**Visual Quality Comparison.** Visual quality is an important but subjective metric, so we use the same metric as previous methods, such as SRGAN [20] to let the users judge by themselves. The mean opinion score (MOS) is proposed to measure the visual quality of images. To compare the visual quality of different kinds of adversarial example methods, we also adopt MOS as the evaluation metric by default. Specifically, we randomly selected twenty volunteers and asked them to rate the visual quality of adversarial examples from 1 (bad quality) to 5 (excellent quality). The evaluation set is randomly selected from the ImageNet dataset.

The final visual comparison results are shown in Table 1. It can be seen that our method achieves the same MOS score as C&W [6], which is much better than those achieved by the other methods. The reason why C&W also achieves such a high MOS is that their method only adds very small and invisible perturbations onto the original image.

For a more intuitive comparison, some visual examples are given in Fig. 6. When the frequency of the video displayed on the display exceeds the critical flicker frequency, the human visual system merges the two frames into one, and then people observe only a single picture. Since the common camera equipment records the average value of the objects captured during the exposure time, if the shutter speed of the common camera equipment is reduced to a frequency that is lower than that of the video displayed on the screen, the fusion of multiple frames will be taken, which is similar to the tendency of the human eye to merge frames. Therefore, we use cameras to take long-exposure photos by reducing the shutter speed to simulate the fusion process of the human eye, as shown in Fig. 6(h). Obviously, to have a strong attack ability, I-FGSM_32 and PGD_32 need to use very large adversarial perturbations and destroy the original visual quality considerably. By contrast, although our ***CDAE*** method has an even larger modification, it still theoretically guarantees that the adversarial examples observed by human eyes are completely consistent with the original images. To ensure fairness in the following attack ability comparison, we compare our method to these methods with small perturbations to ensure similar visual quality.

**White-box Attack.** For white-box attack, we tried four different recognition models (resnet_v2_101 [16], inception_resnet_v2 [35], inception_v3 [36], and VGG [33]) and compared our method with four state-of-the-art adversarial attacking methods on the ImageNet dataset. For each generated adversarial example, we first use the same camera to capture it from the screen display and then test their attack abilities. In this experiment, without a loss of generality, we use I-FGSM [19] as our basic adversarial attack algorithm in the implementation of our gradient-based ***CDAE*** method. As shown in Table 2, gradient-based CDAE and generation-based CDAE achieve comparable results, and our gradient-based and generation-based CDAE both outperform the baseline I-FGSM algorithm by approximately $5\% \sim 24\%$ and even outperform PGD_4.

**Black-box Attack.** For the black-box attack, we adopt the same configuration as the white-box attack; the only difference is that we generate adversarial examples based on the recognition model resnet_101 but test them on the above four recognition models. As we can see in Table 3, compared to our baseline I-FGSM, the attack transferability of our method is still better by $2\% \sim 6\%$. Except for inception_resnet_v2, our method achieved the best results. Note that the performance of CDAE is influenced by the baseline attack method, which is based on (default: I-FGSM); however, it can always bring relative

**Table 1**
The MOS of different methods.

| | C&W | I-FGSM_4 | I-FGSM_32 | PGD_4 | PGD_32 | ***Gradient-based CDAE*** | ***Generation-based CDAE*** |
|---|---|---|---|---|---|---|---|
| MOS | 4.5 | 3.6 | 1.5 | 3.9 | 2.0 | **4.9** | 4.5 |

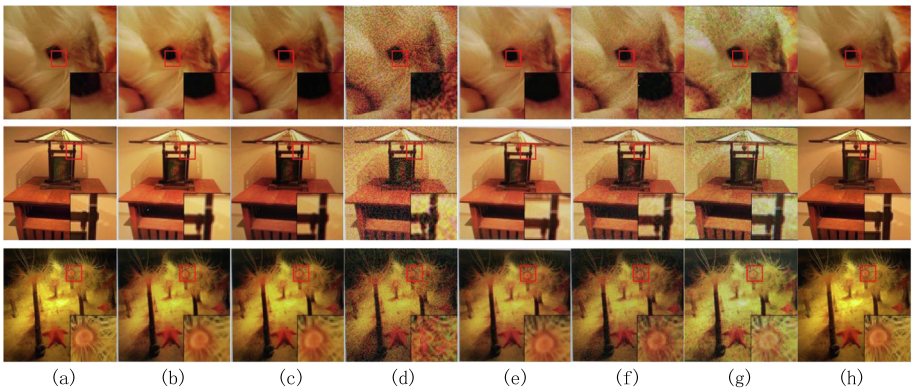


**Fig. 6. Picture adversarial examples captured by a camera.** (a) Clean image. (b) C&W [6]. (c) I-FGSM [19] at 4. (d) I-FGSM [19] at 32. (e) PGD [27] at 4. (f) PGD [27] at 32. (g) Our method: ***CDAE*** captured by a camera. (h) **CDAE** evaluated by HVS.

**Table 2**
Classification accuracy of white-box adversarial attacks with networks by different methods after capturing.

| | resnet_v2_101 [16] | inception_resnet_v2 [35] | inception_v3 [36] | VGG [33] |
|---|---|---|---|---|
| C&W [6] | 34.6% | 58.6% | 40.0% | 29.7% |
| I-FGSM_4 [19] | 12.7% | 55.8% | 34.3% | 15.1% |
| PGD_4 [27] | 6.9% | 52.4% | 27.6% | 15.5% |
| ATN_4 [4] | 15.6% | 56.9% | 36.2% | 25.2% |
| GAP_4 [30] | 22.1% | 52.7% | 37.7% | 28.2% |
| ***Gradient-based CDAE*** | **6.0%** | **46.0%** | 10.0% | 5.6% |
| ***Generation-based CDAE*** | 7.2% | 48.5% | **9.5%** | **4.5%** |

**Table 3**
Classification accuracy of black-box adversarial attacks with networks by different methods after capturing.

| | resnet_v2_101 [16] | inception_resnet_v2 [35] | inception_v3 [36] | alexnet [18] | VGG [33] |
|---|---|---|---|---|---|
| C&W [6] | 34.6% | 54.3% | 38.7% | 24.9% | 29.3% |
| I-FGSM_4 [19] | 12.7% | 26.6% | 15.9% | 9.8% | 11.1% |
| PGD_4 [27] | 6.9% | **19.4%** | 10.5% | 5.3% | 6.4% |
| ATN_4 [4] | 15.6% | 42.1% | 25.8% | 18.9% | 23.7% |
| GAP_4 [30] | 22.1% | 35.6% | 20.6% | 12.6% | 18.3% |
| ***Gradient-based CDAE*** | **6.0%** | 24.0% | **10.0%** | 4.8% | **5.7%** |
| ***Generation-based CDAE*** | 7.2% | 23.4% | 12.4% | **4.2%** | 6.2% |

performance gain. If CDAE is based on PGD, the performance on inception_resnet_v2 will be 16.8%, which is better than the original PGD's performance (19.4%).

The generative model and the target model are tightly coupled, and the GAN needs to be retrained when switching target models. But our method is still efficient. First, our proposed generative model, i.e., CDAE, is tightly coupled with the target model, but due to the transferability of most common adversarial examples, when switching to a different target model, even without retraining, CDAE still has attack ability but with a weaker performance. Second, we have given the comparative results under black-box conditions in Table 3. The results show that CDAE achieves a better attack capability than the state-of-the-art methods even without retraining when attacking different target models, which means that CDAE has better transferability than other methods. Third, the generative model we proposed requires more training costs during the training period, but during the testing period, it can generate a large number of adversarial examples efficiently, which is more advantageous in practical applications.

**Attack Adversarial Training Model.** Many different adversarial defense methods have also been proposed [38,43,45,22,8,25,49]. Adversarial training is by far one of the most effective adversarial defense methods [38,43]. Here we try to attack some very powerful defense models proposed in [38]. For fairness, we directly adopt their released pre-training models on the Internet. It can be seen from Table 4 that our method is even able to attack the networks that are adversarially trained and achieves better performance than state-of-the-art methods.

**Video Action Recognition Adversarial Attack.** We use the most popular action recognition algorithms, TSN, Two-Stream and I3D, as the target networks to attack. Since UP [21] only works for the C3D [39] network, we only compare our method with the image insertion-based video attack method [17], which proposes an attack method for video classifiers and exploits the vulnerability in the Google video classification API. Because the underlying API algorithm only processes the first frame of

**Table 4**
Classification accuracy of adversarial attacks with adversarial training networks by different methods after capturing.

| | ens_incp_res_v2 [38] | incp_res_v2 [38] | incp_v3 [38] | ens3_incp_v3 [38] | ens4_incp_v3 [38] |
|---|---|---|---|---|---|
| C&W [6] | 31.6% | 36.2% | 15.5% | 11.8% | 17.3% |
| I-FGSM_4 [19] | 23.5% | 31.6% | 20.3% | 15.2% | 15.6% |
| PGD_4 [27] | 18.6% | 15.5% | 7.7% | 8.3% | 5.0% |
| GAP_4 [30] | 28.4% | 33.5% | 18.3% | 13.6% | 14.8% |
| ***Gradient-based CDAE*** | 9.1% | **10.6%** | **5.4%** | 7.1% | **4.4%** |
| ***Generation-based CDAE*** | **8.6%** | 12.3% | 8.5% | **6.3%** | 7.0% |

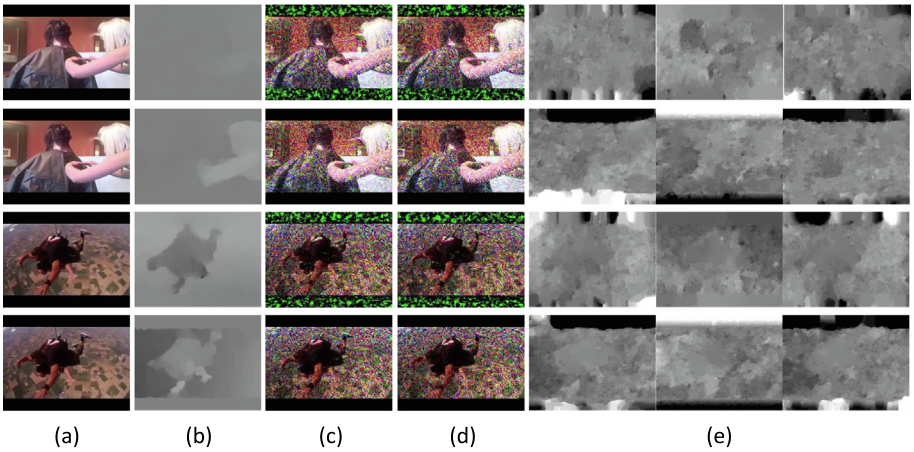


**Fig. 7. Video adversarial examples by *CDAE*.** (a) Clean image. (b) Optical flow in the $x$ direction and $y$ direction. (c) Adversarial example for the first frame. (d) Adversarial example for the second frame. (e) Adversarial examples' optical flow in the $x$ direction and $y$ direction.

**Table 5**
Video adversarial examples' classification accuracy of adversarial attacks by different methods.

| | TSN [41] | Two-Stream [32] | I3D [7] |
|---|---|---|---|
| Clean video | 93.5% | 88.0% | 97.6% |
| Image insertion attack | 93.5% | 88.0% | 97.6% |
| ***CDAE*** (our method) | **4.5%** | **0%** | **0%** |

a video every second, by inserting an attack image into the video every second, the API only outputs the tags associated with the inserted attack image.

Fig. 7 shows some video adversarial examples generated by ***CDAE***. It is easy to find that the optical flow of the original video is greatly attacked by our method in both the $x$ and $y$ directions. As shown in Table 5, our method achieves excellent attack results. It is nearly 100% successful when attacking the Two-Stream network and I3D, and 95.5% when attacking the TSN network. By comparison, the image insert attack baseline [17] almost fails.

## 5. Ablation study

**Classification Accuracy vs Iteration $n$.** In this part, we explore the influence of iteration $n$ on the success rate of gradient-based CDAE. We use I-FGSM as the basic algorithm and set the step size to 5. The top part of Fig. 8 shows the relationship curve between the classification accuracy of **CDAE** on the Imagenet dataset and iteration number $n$. We can see that as $n$ increases, the classification accuracy decreases, *i.e.,* the attack ability grows. When $n$ reaches 30, the classification accuracy stops decreasing, which means our method converges at this stage.

**Modification Amount $\delta$ vs Iteration $n$.** Since we are also curious about whether the modification amount $\delta$ is able to converge in **CDAE**, the relationship between the modification amount $\delta$ and the iteration number $n$ on the ImageNet dataset is further studied in this part. As shown in the bottom of Fig. 8, at the early stage of optimization, as $n$ increases, the modification amount $\delta$ increases. However, when the modification amount $\delta$ reaches 70, it almost stops increasing, which means that the maximum $\delta$ is obtained. This demonstrates that the proposed alternative optimization algorithm can approximate the objective function Eq. (2) quite well and at least reach a local optimal solution.

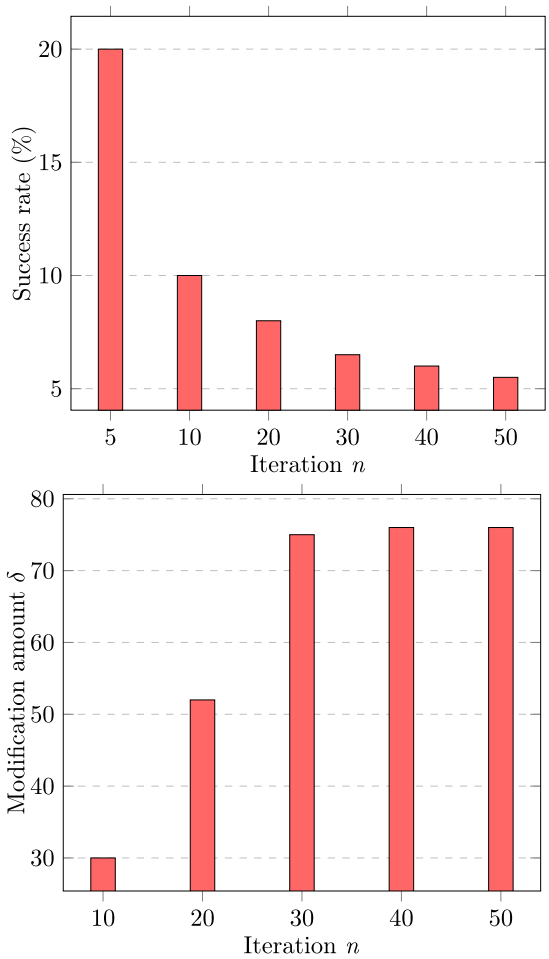


**Fig. 8.** Classification accuracy vs iteration $n$ (top), and modification amount $\delta$ vs iteration $n$ (bottom).

## 6. Conclusion

Inspired by the property of the human visual system, we proposed the first color decomposition-based adversarial example generation framework ***CDAE*** for screen devices. It successfully circumvents the self-contradiction problem between visual quality and attack ability. More specifically, it can theoretically guarantee the visual quality perceived by human observers while maximizing the adversarial perturbations to achieve a better attack ability. Equipped with I-FGSM and GAP, we propose both gradient-based CDAE and generation-based CDAE. The extensive experiments demonstrate that it can significantly improve the attack ability and transferability.

Despite the power of our method, there are many facets that are worth investigating in the future. For example, we will study robust adversarial examples that are insensitive to capturing variations (e.g., distance change, angle change, and camera type). In addition, the transferability of adversarial examples among different models should also be further boosted.

## CRediT authorship contribution statement

**Huanyu Bian:** Conceptualization, Methodology, Writing - original draft. **Hao Cui:** Software. **Kunlin Liu:** Data curation. **Hang Zhou:** Formal analysis. **Dongdong Chen:** Writing - review & editing. **Nenghai Yu:** Funding acquisition. **Wenbo Zhou:** Visualization. **Weiming Zhang:** Supervision, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62072421 and 62002334, Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001, and by Fundamental Research Funds for the Central Universities under Grant WK2100000011.

## References

[1] Satoshi Abe, Atsuro Arami, Takefumi Hiraki, Shogo Fukushima, Takeshi Naemura, Imperceptible color vibration for embedding pixel-by-pixel data into lcd images, in: CHI EA, ACM, 2017.
[2] Stephen J Anderson, David C Burr, Spatial and temporal selectivity of the human motion detection system, Vision Res. (1985).
[3] Anish Athalye and Ilya Sutskever. Synthesizing robust adversarial examples. arXiv, 2017..
[4] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. arXiv, 2017..
[5] Tom B Brown, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. Adversarial patch. arXiv: Computer Vision and Pattern Recognition, 2018. .
[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. arXiv, 2016..
[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In CVPR, pages 4724–4733. IEEE, 2017..
[8] Kejiang Chen, Yuefeng Chen, Hang Zhou, Xiaofeng Mao, Yuhong Li, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Self-supervised adversarial training. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 2218–2222. IEEE, 2020..
[9] C CIE. Commission internationale de l'eclairage proceedings, 1931. Cambridge University Press Cambridge, 1932. .
[10] Xiaoyi Dong, Jiangfan Han, Dongdong Chen, Jiayang Liu, Huanyu Bian, Zehua Ma, Hongsheng Li, Xiaogang Wang, Weiming Zhang, and Nenghai Yu. Robust superpixel-guided attentional adversarial attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12895–12904, 2020..
[11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. pages 9185–9193, 2018. .
[12] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. arXiv, 1, 2017. .
[13] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. arXiv, 2018. .
[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples (2014). arXiv, 2014. .
[15] Jiangfan Han, Xiaoyi Dong, Ruimao Zhang, Dongdong Chen, Weiming Zhang, Nenghai Yu, Ping Luo, and Xiaogang Wang. Once a man: Towards multi-target attack via learning multi-target adversarial network once. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5158–5167, 2019..
[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. .
[17] Hossein Hosseini, Baicen Xiao, Andrew Clark, and Radha Poovendran. Attacking automatic video analysis algorithms: A case study of google cloud video intelligence api. In MPS, pages 21–32. ACM, 2017. .
[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, pages 1097–1105, 2012. .
[19] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv, 2016..
[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In CVPR, volume 2, page 4, 2017. .
[21] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. arXiv, 2018. .
[22] Tianlin Li, Aishan Liu, Xianglong Liu, Xu. Yitao, Chongzhi Zhang, Xiaofei Xie, Understanding adversarial robustness via critical attacking route, Inf. Sci. 547 (2021) 568–578.
[23] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 1028–1035, 2019..
[24] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In Proc. Eur. Conf. Comput. Vis., pages 395–410. Springer, 2020. .
[25] Jiayang Liu, Weiming Zhang, Yiwei Zhang, Dongdong Hou, Yujia Liu, Hongyue Zha, and Nenghai Yu. Detection based defense against adversarial examples from the steganalysis point of view. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4825–4834, 2019..
[26] M Ronnier Luo, Guihua Cui, and Bryan Rigg. The development of the cie 2000 colour-difference formula: Ciede 2000. Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 26(5), 340–350, 2001. .
[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv, 2017. .
[28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In CVPR, pages 2574–2582, 2016. .
[29] Viet Nguyen, Yaqin Tang, Ashwin Ashok, Marco Gruteser, Kristin Dana, Wenjun Hu, Eric Wengrowski, and Narayan Mandayam. High-rate flicker-free screen-camera communication with spatially adaptive embedding. In INFOCOM, pages 1–9. IEEE, 2016. .
[30] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. arXiv, 2017. .
[31] Ernst Simonson, Josef Brozek, Flicker fusion frequency: background and applications, Physiological reviews 32 (3) (1952) 349–378.
[32] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014. .
[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv, 2014..
[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012. .
[35] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alexander A Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, AAAI 4 (2017) page 12.
[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In CVPR, pages 2818–2826, 2016. .
[37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXiv, 2013..
[38] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. arXiv, 2017..
[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, pages 4489–4497, 2015. .

[40] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In CVPR, pages 1430–1439, 2018. .
[41] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In ECCV, pages 20–36. Springer, 2016. .
[42] Grace Woo, Andrew Lippman, and Ramesh Raskar. Vrcodes: Unobtrusive and active visual codes for interaction by exploiting rolling shutter. In ISMAR, pages 59–64. IEEE, 2012..
[43] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 501–509, 2019..
[44] Hongyue Zha, Weiming Zhang, Chuan Qin, and Nenghai Yu. Direct adversarial attack on stego sandwiched between black boxes. In 2019 IEEE International Conference on Image Processing (ICIP), pages 2284–2288. IEEE, 2019..
[45] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. arXiv, pages arXiv–1909, 2019. .
[46] Lan Zhang, Cheng Bo, Jiahui Hou, Xiang-Yang Li, Yu Wang, Kebin Liu, and Yunhao Liu. Kaleido: You can watch it but cannot record it. In MobiCom, pages 372–385. ACM, 2015. .
[47] Yiwei Zhang, Weiming Zhang, Kejiang Chen, Jiayang Liu, Yujia Liu, and Nenghai Yu. Adversarial examples against deep neural network based steganalysis, in: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, 2018, pp. 67–72.
[48] Hang Zhou, Dongdong Chen, Jing Liao, Kejiang Chen, Xiaoyi Dong, Kunlin Liu, Weiming Zhang, Gang Hua, and Nenghai Yu. Lg-gan: Label guided adversarial network for flexible targeted attack of point cloud based deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10356–10365.
[49] Hang Zhou, Kejiang Chen, Weiming Zhang, Han Fang, Wenbo Zhou, Yu. Nenghai, Dup-net: Denoiser and upsampler network for 3d adversarial point clouds defense, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1961–1970.