Cycle GAN-Based Attack on Recaptured Images to Fool Both Human and Machine*

Wei Zhao[a,b], Pengpeng Yang[a,b], Rongrong Ni[a,b], Yao Zhao[a,b], and Wenjie Li[a,b]

a.Institute of Information Science,Beijing Jiaotong University, Beijing 100044, China.
b.Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China {rrni@bjtu.edu.cn yzhao@bjtu.edu.cn}

Abstract. Recapture is often used to hide the traces left by some operations such as JPEG compression, copy-move, etc. However, various detectors have been proposed to detect recaptured images. To counter these techniques, in this paper, we propose a method that can translate recaptured images to fake "original images" to fool both human and machines. Our method is proposed based on Cycle-GAN which is a classic framework for image translation. To obtain better results, two improvements are proposed: (1) Considering that the difference between original and recaptured images focuses on the the part of high frequency, high pass filter are used in the generator and discriminator to improve the performance. (2) In order to guarantee that the images content is not changed too much, a penalty term is added on the loss function which is the L1 norm of the difference between images before and after translation. Experimental results show that the proposed method can not only eliminate traces left by recapturing in visual effect but also change the statistical characteristics effectively.

Keywords: recaptured images · Cycle-GAN · fool human and machine

1 Introduction

Nowadays, with the popularity of digital cameras and the rapid development of Internet technology, it is an indisputable fact that digital images have become important carriers. And image editing software is widely used with the advantage of operability and practicability, which makes it easy to tamper an image. Some tampered images in the fields of politics, military and judicature will bring great harm to the society. Therefore, the identification of digital image authenticity is of particular importance.

One common type of image tampering is recapturing images. The process of recapture is as follows: firstly, the original image is projected onto a new media,

* This work was supported in part by the National Key Research and Development of China (2016YFB0800404), National NSF of China (61672090, 61332012), Fundamental Research Funds for the Central Universities (2017YJS054). We greatly acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

such as computer screen, mobile phone screen or printed paper. Then a new image can be obtained by recapturing the projection. Recaptured images may bring about bad effect on the society if they are used maliciously. For example, due to that all the tampering will leave traces on the image, attackers can eliminate these traces by recapturing the forged image. To against it, the most simple and convenient way is to make a recaptured image decision in advance.

To discriminate between the recaptured and original images, numbers of algorithms have been proposed and mainly include two branches: statistical characteristics [1-3] and deep learning based [4]. In terms of statistical features, Farid *et al*. first proposed a scheme which can to distinguish between natural and unnatural images based on high-order wavelet statistical features . Unnatural images include recaptured images and computer generated images. Cao *et al*. [2] proposed three kinds of statistical features to detect good-quality recaptured images, namely local binary pattern (LBP), multi-scale wavelet statistics (MSWS), and color features (CF). Li *et al*. [3] proposed new features based on the block effect and blurriness effect due to JPEG compression and the screen effect described by wavelet decomposition. And the deep learning based method has been proved that it has better detection performance than that statistical characteristics based. Yang *et al*. [4] proposed a laplacian convolutional neural networks (L-CNN) and improved the detection performance especially for small-size recaptured image.

On the other hand, from the point of view of an attacker, if he want to translate a recaptured image to a fake original image, two goals need to be achieved: the visual effect of LCD should be avoided and it can attack various detection schemes. Generative adversarial networks have achieved many state-of-the-art results in image translation. Generative adversarial networks include generator network and discriminator network. The generator learns the potential distribution of the real data and generates new data and the discriminator is a binary classifier that determines whether the input is real or generated. In training phase, the generator need to be continuously optimized to improve its generating ability and the discriminator need to improve its discriminating ability. The learning process is to find a Nash equilibrium between the two networks. CGAN [5] adds extra information in the generator and discriminator to guide the process of training. Pix2pix-GAN [6] can achieve image-to-image translation tasks with paired images which include an input image and a corresponding target output image. Cycle-GAN [7] used two generators and two discriminators to learn mapping functions between two domains without paired images.

Considering that it is difficult to get the paired images for original images and recaptured images, in this work, a method based on Cycle-GAN is proposed. Due to the fact that the difference between original and recaptured images focuses on high frequency, generator and discriminator with high-pass filter are designed to make a better image translation. Additionally, to guarantee the content of images not change a lot after being translated, a penalty term is added to the loss function which is the L1 norm of the difference between images before and

after translation. Experimental results show that the proposed method can not only fool human in visual effect but also the machine with a high probability.

The rest of the paper is organized as follows. In Section 2, proposed architecture and object function is introduced. Experiments are conducted in Section 3, and conclusions are drawn in Section 4.

2 Proposed method

Our task is to translate recaptured images to target images which is similar to original images not only in visual effect but also in statistical characteristics. It can be formulated as learning a mapping G from recaptured images X to the original images Y given training samples $\{x_i\} \in X$ and $\{y_i\} \in Y$, where $i = 1, 2, ..., m$. Note that X and Y are not corresponding one by one because it is difficult to collect recaptured images which are completely same with original images.

The overall framework of the model is shown in Fig. 1,, two generators and discriminators are used. y is present as recaptured images and x is present as original images. Generator G learns the distribution of Y and F learns the distribution of X. Discriminator D_X aims to distinguish between recaptured images $\{x\}$ and fake-recaptured images $\{F(y)\}$, and D_Y aims to distinguish between original images $\{y\}$ and fake-original images $\{G(x)\}$. To promise the mapping is meaningful, cycle network structure is used. As shown in dotted arrow, the translated images $\{G(x)\}$ and $\{F(y)\}$ are fed into the generator F and G. By limiting the difference between x and $F(G(x))$ and the difference between y and $G(F(y))$, the model can be further standardized. The training process is based on game theory and it aims at achieving the Nash equilibrium between generators and discriminators.

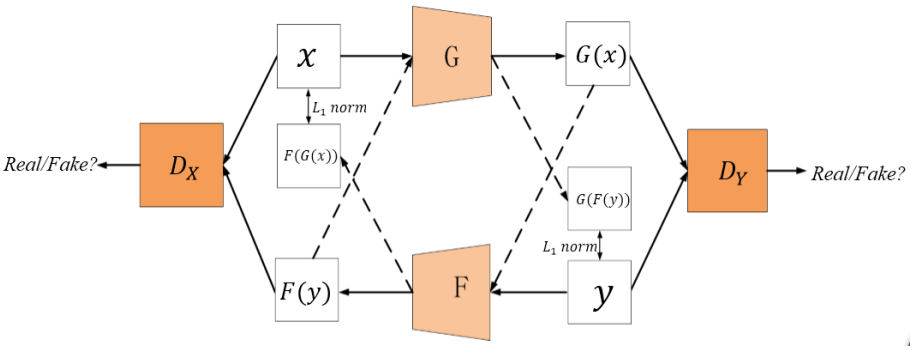

Fig. 1. The overall framework of the model.

2.1 Architecture

Generator Two generators G and F are included. The Generator G can translate X to Y and Generator F can translate Y to X. The two generators have the same architecture which is shown in Fig. 2. Considering that the difference between original and recaptured images focus on the high frequency part, so only the part of high frequency is extracted and fed into generators. Generators are only responsible for learning the difference of high frequency, which is more easier to train than reconstructing the whole image. In this work, the laplace filter are used:

$$LF = \begin{bmatrix} 0, & -1, & 0 \\ -1, & 4, & -1 \\ 0, & -1, & 0 \end{bmatrix} \tag{1}$$

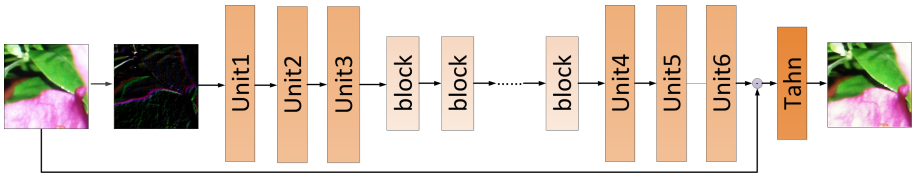

Fig. 2. The architecture of generator.

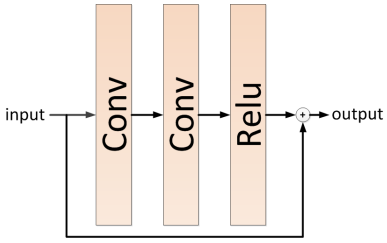

Fig. 3. The architecture of block.

In addition, six units and five residual blocks are combined together. Each of the unit 1, 2, 3 and 6 include a convolution layer, a batch normalization and a Relu function. Each of the unit 4 and 5 include a deconvolution layer, a batch normalization and a Relu function. Each residual block includes two convolution layers and a Relu function. The structure of residual block is shown in Fig. 3. In the end, a TanH activation function is used. The parameters of generator are presented in Table 1.

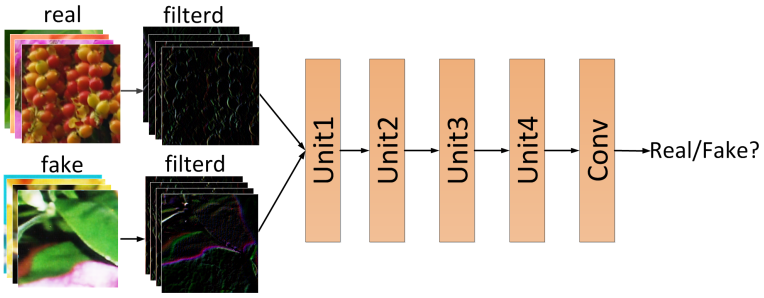

Fig. 4. The architecture of discriminator.

Discriminator Two discriminators D_X and D_Y are included. As shown in Fig. 4, the discriminator is designed to distinguish real images and fake images. Similarly, a laplace filter is used, followed by four units and one convolution layer are used. Each unit includes a convolution layer, a batch normalization and a Leaky-ReLU function. The parameters of discriminator are presented in Table 1.

Table 1. The detailed parameters of the architecture

Generator	Unit 1	Conv(7*7*32),padding=3,stride=1;batchnorm; Relu
	Unit 2	Conv(3*3*64), stride=2;batchnorm; Relu
	Unit 3	Conv(3*3*128), stride=2; batchnorm; Relu
	block	Conv(3*3*128), stride=1; Conv(3*3*128), stride=1; Relu
	Unit 4	deconv(3*3*128), stride=2; batchnorm; Relu
	Unit 5	deconv(3*3*256), stride=2; batchnorm; Relu
	Unit 6	Conv(7*7*3), stride=1; batchnorm; Relu
Discriminator	Unit 1	Conv(4*4*32), stride=2; batchnorm; Leaky-ReLU
	Unit 2	Conv(4*4*64), stride=2; batchnorm; Leaky-ReLU
	Unit 3	Conv(4*4*128), stride=2; batchnorm; Leaky-ReLU
	Unit 4	Conv(4*4*256), stride=1; batchnorm; Leaky-ReLU
	Conv	Conv(4*4*1), stride=1;

2.2 Object Function

The loss function of proposed method contains three parts: adversarial loss, cycle consistency loss and low frequency consistency loss.

Adversarial Loss The optimization process of GAN is actually a game between two competing networks: the generator is responsible for generating data which is similar to the real data, and the discriminator is responsible for distinguishing the generated data from the real data. Formally, the game between the generator G and the discriminator D has the minimax objective. Note that the distribution of recaptured images is $P_{data}(x)$ and the distribution of original images is $P_{data}(y)$. We need to translate a recaptured image x to a target image $G(x)$ which follows the distribution of $P_{data}(y)$. Therefore, for the mapping function $G : X \rightarrow Y$ and its discriminator D_Y:

$$\begin{aligned} L_{adv}(G, D_Y, X, Y) =& E_{y \sim p_{data}(y)}(\log D_Y(y)) \\ &+ E_{x \sim p_{data}(x)}(1 - \log D_Y(G(x))), \end{aligned} \tag{2}$$

where G tries to generate images $G(x)$ that look similar to images from domain Y, while D_Y aims to distinguish between translated samples $G(x)$ and real samples y. G tries to minimize this objective and D tries to maximum it.

Due to it is meaningless to learn the translation from original images to recaptured images, so the results of $F : Y \rightarrow X$ is not involved in our experiment. But it's a essential part in the entire framework for cycle consistency. So, for the mapping function $F : Y \rightarrow X$ and its discriminator D_X, there is another constraint:

$$\begin{aligned} L_{adv}(F, D_X, X, Y) =& E_{x \sim p_{data}(x)}(\log D_X(x)) \\ &+ E_{y \sim p_{data}(y)}(1 - \log D_X(F(y))), \end{aligned} \tag{3}$$

Cycle Consistency Loss Compared with other generation models, the greatest advantage of GAN is that it doesn't need to formulate a target distribution, but to learn the distribution directly using two group of images. However, this mechanism also brings a shortcoming that the model is too free and uncontrollable. A generator can map the input images to any random permutation of images in the target domain, which may cause there is not any semantic links between input images and output images. Thus, it's difficult to guarantee that the learned function can map input X to desired output Y. To ensure the mapping is practical, cycle consistency loss is introduced.

For each image x from domain X, the image translation cycle should be able to bring x back to itself: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. And for each image y from domain Y, the image translation cycle also need to bring y back to itself: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$:

$$\begin{aligned} L_{cyc}(G, F) =& E_{x \sim p_{data}(x)}(\|F(G(x)) - x\|_1) \\ &+ E_{y \sim p_{data}(y)}(\|G(F(y)) - y\|_1), \end{aligned} \tag{4}$$

Low Frequency Consistency Loss It has been noted that the dataset is unpaired which is convenient to be collected. But it also bring a disadvantage that no groundtruth for recaptured images to constrain the model when training. Due to the deep learning model is driven by data, the generator may learn the difference between original image and recaptured image in training dataset. So if the dataset is not rich enough, the model is likely to be overfitting. The generator can not only learn the difference left by recapturing, but also other difference between two group of data, such as: color distribution, the content of images and so on. In that cases, the translated images may have large chromatic differences from target images. Considering the characteristic of recaptured images, the content of images is similar with original images and the main difference is focus on the high frequency. So an extra term need to be added to ensure that the low frequency part is not changed. In proposed method, median filtering is used to extract the part of low frequency:

$$\begin{aligned} L_{Low}(G,F) =& E_{x\sim p_{data}(x)}(\|f(G(x)) - f(x)\|_1) \\ &+ E_{y\sim p_{data}(y)}(\|f(F(y)) - f(y)\|_1), \end{aligned} \tag{5}$$

where, $f(.)$ is a median filter function which can reserve the low frequency part.

In total, the full objective is:

$$\begin{aligned} L(G,F,D_X,D_Y) =& L_{adv}(G,D_Y,X,Y) \\ &+ L_{adv}(F,D_X,X,Y) \\ &+ \alpha L_{cyc}(G,F) \\ &+ \beta L_{low}(G,F), \end{aligned} \tag{6}$$

where α, β are weight coefficients. In the experiments, α is set as 10 and β is set as 5.

Finally, by optimizing the loss function according to Eq. (7), we can get the well-trained generators and discriminators. According to the purpose of this work, only generator G is needed.

$$G^* = \min_{G,F} \max_{D_X,D_Y} L(G,F,D_X,D_Y), \tag{7}$$

where, G^* presents the well-trained generator G.

3 Experimental results and analysis

Image database in the experiments includes 20000 images: 10000 original images and 10000 recapture images. The size of the images is 256×256. The images derive from the image databases provided in [2]. We crop the block with size of 1024×1024 from the center of the images. Then the images are cut into non-overlapping blocks of 512×512. Finally, 256×256 images are got by center

clipping. And training dataset, validation and test dataset are randomly divided by percent 40/10/50. Hyper-parameter setting in the experiment is as follows: the learning rate is 0.0001 and iteration epoch is 15. And the Adam optimizer with $\beta = 0.5$ are used. All the results shown in this section are averaged over 6 random experiments.

In the experiment, three recaptured image detection methods are involved. They include the method based on statistical characteristics: LBP feature [3] and wavelet statistical feature [3] and based on deep learning: L-CNN [4]. Firstly, these three methods are well-trained to get the different accuracies for different images. Furthermore, to analyze the validity of model modification, a contrast experiment is performed in which the model is original Cycle-GAN without any modification. Finally, in order to verify the effectiveness of proposed method, it was trained using the training dataset and the recaptured images in test set are fed into the model to be transferred to a fake images.

In Table 2, the detection accuracies using three methods for different images are presented. Noted that $IMAGE_{nor}$ means the recaptured images in test dataset without any translation. $IMAGE_{cyc}$ means the images translated by original Cycle-GAN and $IMAGE_{prop}$ means the images translated by proposed method. It can be seen that there detection methods mentioned above can all detect the recaptured images effectively. And the the ability of original Cycle-GAN to attack the detection methods is worse than the proposed method. It can be seen that after being transferred by proposed method, the classic schemes will be fooled with great probability. At the same time, it is noticed that the attack effect is different for three detection methods, and the proposed method can attack the L-CNN effectively but don't perform very well in attacking the methods of LBP and wavelet. We guess it is because the method of L-CNN is more similar to the discriminator of proposed method.

In the aspect of visual effects, six group of images are shown in Fig. 5. In each group, recaptured image is on the top and the translated image is on the bottom. From these images, we can find that proposed method can remove the traces of texture left by recapturing LCD screen effectively.

In conclusion, the proposed method can not only eliminate of traces left by recapturing in visual effect but also change the statistical characteristics to attack the detection methods effectively.

4 Conclusion

In this paper, we proposed a method to translate recaptured images to fake "original images" based on Cycle-GAN. According to the characteristics of recaptured images, generator and discriminator with high-pass filter are designed to make a better image translation. Additionally, to guarantee the content of images don't change a lot after being translated, a penalty term is added to the loss function which is the L1 norm of the difference between images before and after translation. Experimental results show that the proposed method can not only fool human in visual effect but also the machine with a high probability.

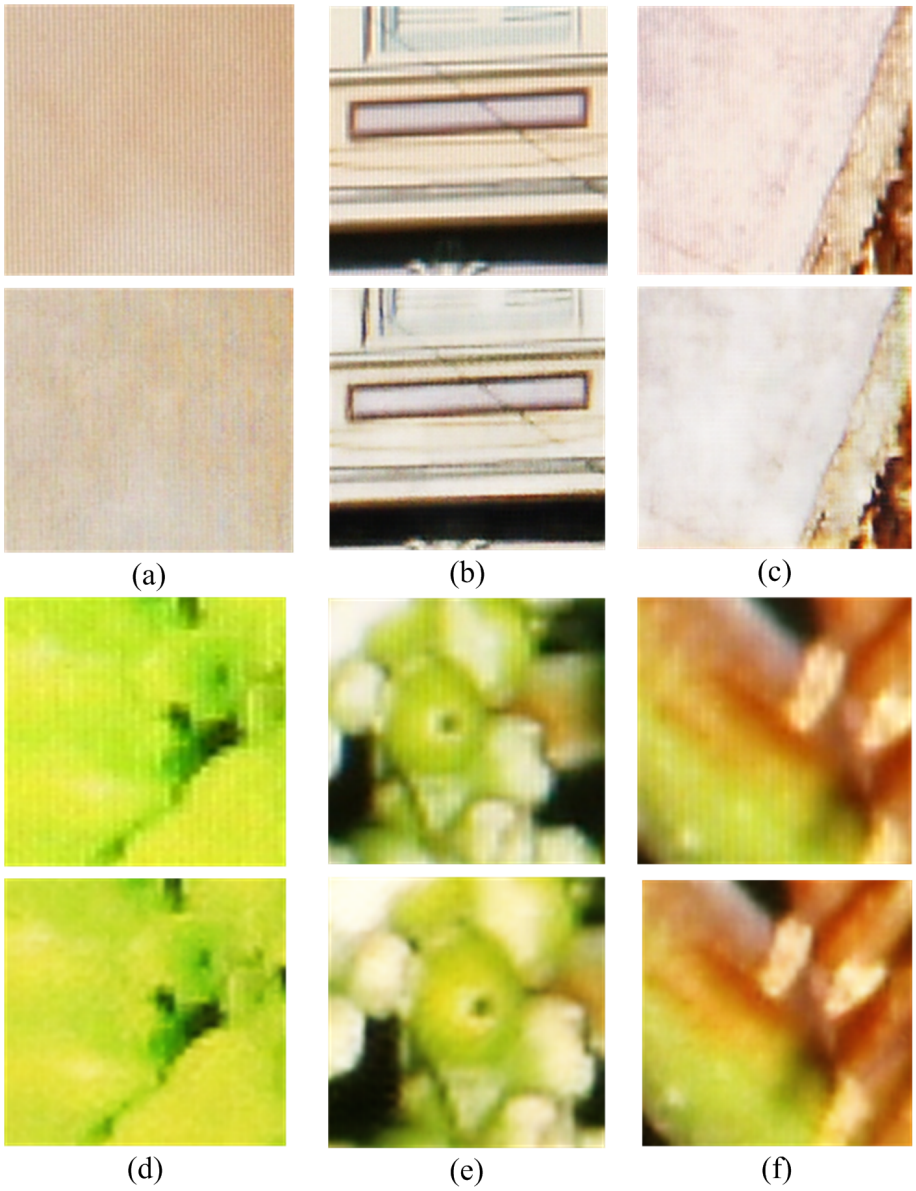

Fig. 5. The visual effect of recaptured images and corresponding translated images. In each group, recaptured image is on the top and the translated image is on the bottom.

Table 2. The classification accuracy using three methods for different images

method / image	$L-CNN$	LBP	$Wavlet$
$IMAGE_{nor}$	99.0%	95.6%	82.0%
$IMAGE_{cyc}$	34.83%	70.96%	53.3%
$IMAGE_{prop}$	9.4%	32.85%	39.44%

References

1. S. Lyu and H. Farid: How Realistic is Photorealistic?, IEEE Trans. on Signal Processing, 845-850(2005).
2. Cao. H, Alex, K.C:Identification of recaptured photographs on LCD screens. IEEE International Conference on Acoustics,Speech and Signal Processing, 1790–1793 (2010)
3. Li, R., Ni, R., Zhao, Y.: An effective detection method based on physical traits of recaptured images on LCD screens. International Workshop on Digital-forensics and Watermaking,107-116 (2015)
4. Yang Pengpeng, Rongrong Ni, Yao Zhao: Recapture image forensics based on laplacian convolutional neural networks. InternationalWorkshop on DigitalWatermarking, Springer, Cham, (2016).
5. Mirza, Mehdi, and Simon Osindero: Conditional generative adversarial nets. arXiv preprint arXiv, 1411–1784 (2014).
6. Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A: Image-to-image translation with conditional adversarial networks. arXiv preprint.(2017).
7. Zhu, J. Y., Park, T., Isola, P., Efros, A. A: Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint arXiv:1703.10593 (2017)