J. Vis. Commun. Image R. 102 (2024) 104210

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Full Length Article

Transferable adversarial attack on image tampering localization

Gang Cao [a,b,c,*], Yuqi Wang [a,b], Haochen Zhu [a,b], Zijie Lou [a,b], Lifang Yu [d]

[a] School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China
[b] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[c] School of Information Engineering, Changsha Medical University, Changsha 410219, China
[d] Beijing Key Laboratory of Signal and Information Processing for High-End Printing Equipments, Beijing Institute of Graphic Communication, Beijing 100026, China

ARTICLE INFO

Keywords:
Anti-forensics
Adversarial example
Optimization-based adversarial attack
Fast gradient sign method
Image tampering localization

ABSTRACT

It is significant to evaluate the security of existing digital image tampering localization algorithms in real-world applications. In this paper, we propose an adversarial attack scheme to reveal the reliability of such deep learning-based tampering localizers, which would be fooled and fail to predict altered regions correctly. Specifically, two practical adversarial example methods are presented in a unified attack framework. In the optimization-based adversarial attack, the victim image forgery is treated as the parameter to be optimized via Adam optimizer. In the gradient-based adversarial attack, the invisible perturbation yielded by Fast Gradient Sign Method (FGSM) is added to the tampered image along gradient ascent direction. The black-box attack is achieved by relying on the transferability of such adversarial examples to different localizers. Extensive experiments verify that our attacks can sharply reduce the tampering localization accuracy while preserving high visual quality for attacked images. Source code is available at https://github.com/multimediaFor/AttackITL.

1. Introduction

Since digital image editing becomes easy, image authenticity is questioned frequently. It is important to develop digital forensic techniques for detecting image forgeries [1]. Many effective image tampering localization algorithms base on deep learning [2–10] have been proposed in recent years. Such algorithms effectively learn internal forensic traces from the training data. Specifically, the local consistency-based image tampering localizers, such as Noiseprint [7], EXIF-Net [8], Forensic Similarity Graph [9] and ManTra-Net [10], regard tampering localization as an anomaly detection problem. Such localizers rely on the extraction and consistency-checking of appropriate local features. There also exists another type of more effective localizers [2–6], which regard the localization as image semantic segmentation. The tampering localization maps generated by encoder/decoder networks consist of pixel-level binary real/falsified labels.

In digital forensic applications, the authentication result of an image tampering localization system is typically influential and rather important. It is significant to evaluate the security of such tampering localization algorithms against malicious attacks in real-world applications. Different from the image classification scenario, tampering localization involves the pixel-level prediction of tampering probability. As a result, it is necessary to specially address the adversarial attack on existing tampering localizers from the view of anti-forensics.

Previous attacks on image forensic algorithms are generally based on artificial features [11–14], Generative Adversarial Network (GAN) [15,16] and adversarial examples [18–20]. Chen et al. proposed an attack method aim at SVM-based forensics technology for high-dimensional SPAM features [11]. Cao et al. proposed anti-forensic contrast enhancement operators [12], which are undetectable by the existing detectors based on the histogram peak-gap artifacts. The trace forging and hiding attacks on contrast enhancement are further proposed in [13]. Such attack methods are mostly targeted to the specific forensic detection algorithm, and fail to address the deep learning-based tampering localizers. Xie et al. proposed a GAN-based method to attack global manipulation detection schemes [15]. In the recent literature [16], forensic traces are synthesized by a two-phase GAN to deceive local consistency-based image splicing detectors and localizers [7–9]. However, such a method fails to be used for attacking the other major category of localizers, i.e., the segment-based ones [2–6]. Another query-efficient black-box attack against image forgery localization is proposed via reinforcement learning [17]. Note that adversarial examples attend to exploit the vulnerability of neural networks by adding minor perturbation to the input forgery images, resulting in forensic

[*] Corresponding author at: School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China.
E-mail address: gangcao@cuc.edu.cn (G. Cao).

https://doi.org/10.1016/j.jvcir.2024.104210
Received 7 April 2024; Received in revised form 5 June 2024; Accepted 15 June 2024
Available online 20 June 2024

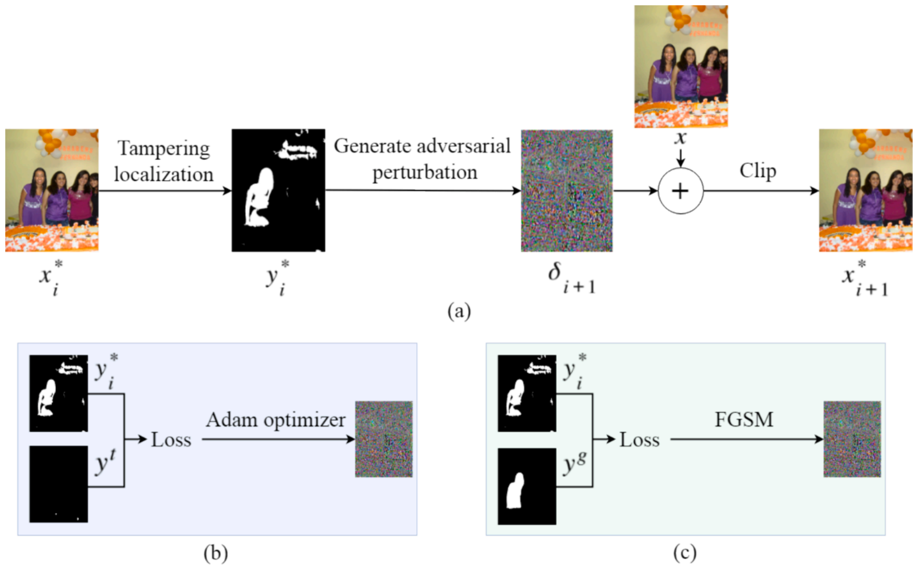


Fig. 1. (a) Overview of the proposed adversarial attack framework. (b)(c) Proposed optimization-based and gradient-based attack methods against image tampering localization algorithms, respectively.

errors [18]. In [20], the optimization-based adversarial example method is employed to attack the convolutional neural network (CNN)-based manipulation detectors. Gradient of the output score function with respect to pixel values is explored in depth. The common gradient-based adversarial attack algorithms including Fast Gradient Sign Method (FGSM) [21], Jacobian-based Saliency Map Attack (JSMA) [22] and Projected Gradient Descent (PGD) [23], which have been used to attack the manipulation detectors [18] and source camera identification models [19,24]. The prior works mainly address the white-box attack scenario, in which the parameters and the structure of the attacked model are fully or partly accessible. In contrast, the black-box attack can only assess the victim model in a nontransparent mode, and is typically achieved by the transferability of generated attacked samples.

To attenuate the deficiency of prior works, in this paper we propose effective adversarial attacks on both the local consistency-based and the segmentation-based image tampering localizers. Specifically, two practical adversarial example methods are presented in a unified attack framework. In the optimization-based attack, the attacked image forgery is treated as the parameter to be optimized via Adam optimizer [25]. In the gradient-based attack, the invisible perturbation yielded by FGSM is added to the tampered image along gradient ascent direction. The transfer-based black-box attack is achieved by applying the generated adversarial example in white-box scenario to other localizers. Extensive evaluations verify the effectiveness of our proposed attack methods.

In the rest of this paper, the detailed attack scheme is proposed in Section 2. Performance testing experiments are given in Section 3, followed by the conclusion drawn in Section 4.

2. Proposed adversarial attack scheme

In this section, we first present the attack framework on tampering localizers in Subsection 2.1. Then two specific attack methods are described in Subsections 2.2 and 2.3, respectively.

2.1. Attack framework on tampering localizers

Let the tampering localization network used for generating adversarial examples be referred to a target localizer, while the localizer to be attacked is a victim localizer. In the white-box attack, the target and victim localizers are the same one and can be formulated by $y = f_\theta(x)$. Here, $x \in [0,1]^{H\times W\times 3}$ denotes the normalized input image forgery with $H \times W$ pixels, and $y \in [0,1]^{H\times W}$ is the predicted pixel-wise tampering probability map. θ denotes model parameters. The pixel $x_{i,j,k}$ at the position (i,j) with higher $y_{i,j}$ values towards 1 signifies a higher probability for being tampered. Let $y^g \in \{0,1\}^{H\times W}$ be the ground truth of the forged image x , where the values 1, 0 mark the tampered and pristine pixels, respectively. Let x^* be the generated adversarial example image. The corresponding localization map y^* predicted by the victim localizer $f_\theta(\cdot)$ can be denoted as $y^* = f_\theta(x^*)$. Generating adversarial examples can be formulated as finding an instance $x^* = x + \delta$, which satisfies the constrains as

$$\begin{cases} y^* \to y^t \\ D(x, x+\delta) \le B \\ x+\delta \in [0,1]^{H\times W\times 3} \end{cases} \qquad (1)$$

where δ is the perturbation quantity. y^t is the target prediction probability map and $y^* \to y^t$ denotes y^* approaching y^t. The attack aims to find suitable δ that makes the forgery region hard to be detected while limiting the visual distortion $D(x, x+\delta)$ below a constant B. $D(\cdot,\cdot)$ is typically realized by L_p norm. After obtaining a suitable δ, x^* should be clipped into the range $[0,1]$ to ensure that it remains a valid image.

Within the above framework, we propose two specific white-box attack methods for generating the adversarial example x^*, i.e., optimization-based and gradient-based attacks. Fig. 1(a) shows the proposed adversarial attack framework. In the black-box attack, the adversarial examples yielded against the target localizer in the white-box scenario would be directly applied to deceive other victim localizers. The black-box attack relies on the transferability of adversarial examples due to the limited knowledge of target tampering localizers.

2.2. Optimization-based attack method

In this method, the attacked image forgery x is regarded as the objective [26] to be optimized. In terms of Eq. (1), generating adversarial examples can be approximated as the following optimization problem:

$$\begin{cases} \underset{\delta}{\mathit{minimize}}\ D(x, x+\delta) \\ \text{such that}\ \ y^* \to O^{H\times W} \\ \qquad\qquad x+\delta \in [0,1]^{H\times W\times 3} \end{cases} \qquad (2)$$

It finds δ that minimizes D and makes y^* tend to a zero matrix $y^t = O^{H\times W}$.

It means that the adversarial example is located by the localizer as the pristine image without any tampered regions. Furthermore, Eq. (2) can be reformulated as

$$\begin{cases} \underset{\delta}{minimize}\ \lambda\|\delta\|_2 + l\left(y^*, O^{H\times W}\right) \\ \text{such that} \quad x+\delta \in [0,1]^{H\times W\times 3} \end{cases} \tag{3}$$

where $l(\cdot,\cdot)$ is the loss function of the optimization process, binary cross-entropy (BCE) loss is used in our experiments. The loss function measures the distance between the prediction probability map y^* and the target map $O^{H\times W}$. The loss between the predict map and the target map is calculated on the entire image to ensure that both the tampered and the unaltered area are expected to be pristine for the target localizer. λ controls the proportion between the magnitude of perturbation and the loss value, the perturbation magnitude is measured by L_2 norm.

Fig. 1(b) provides a detailed illustration of the iterative process for obtaining δ through optimization-based attacks. In the i-th iteration, where $i = 0, 1, 2, \cdots$, the adversarial example x_i^* is input to the target localizer for generating the prediction probability map y_i^*. δ_{i+1} is gained by solving the minimization problem described in Eq. (3) via Adam optimizer. Finally, the adversarial example image is updated by

$$x_{i+1}^* = \max(\min(x+\delta_{i+1}, 1), 0) \tag{4}$$

Note that the perturbation is added globally, since the local modification to tampered regions may still leave some new inconsistency.

2.3. Gradient-based attack method

Inspired by [17], the popular gradient-based adversarial example method FGSM [21] is used to attack tampering localizers. Such attack method is built on the assumption that a linear approximation can be used to calculate the change in the loss function of the neural network when the perturbation values are very small. FGSM takes advantage of the linear approximation of target localizers for fast generation. Adding a small perturbation in the gradient ascent direction can enlarge the loss value of the target localizer dramatically. It makes the localizer predict the opposite output $1 - y^g$. Although FGSM can reduce the prediction accuracy of target localizers due to the non-linearity of loss functions, it will not cause the localizer to predict opposite localization results. Meanwhile, as the perturbation increases, the image visual quality gradually deteriorates. The perturbation magnitude also restricts localization outputs to the opposite. Thus, gradient-based attacks still adhere to the constraints of Eq. (1).

Generating the adversarial example x^* via FGSM can be formulated as

$$x^* = \max(\min(x + \varepsilon\cdot sign(\nabla_x l(y, y^g)), 1), 0) \tag{5}$$

where $l(y, y^g)$ is the loss function of the target localizer at training phase.$sign(x) = 1$ if $x > 0$, otherwise$sign(x) = 0$. When generating adversarial examples to attack the target localizer, the loss value is expected to be as large as possible. Such attack method utilizes the gradient of the loss with respect to the input image forgery x , and the perturbation is added in the direction of gradient rising denoted by $sign(\nabla_x l)$. The magnitude of the perturbation is constrained by $\|\delta\|_\infty \le \varepsilon$ and the adversarial example is cropped to a valid range. Fig. 1(c) shows the detailed gradient-based attack process.

3. Experiments

In this section, the performance evaluation experiments for the proposed attack methods are presented in detail. First, the optimization and gradient based white-box attacks on CASIAv1 dataset are performed to find a suitable target localizer to generate adversarial examples. Then the black-box attacks take advantage of transferability to attack other victim localizers on more different datasets. Further, a comparison of performance with other attacks shows that the proposed attack methods perform well. Finally, the robustness of the adversarial examples to different common manipulations is tested.

3.1. Experimental setting

Datasets and localization algorithms. Test datasets include CASIAv1 [27], Columbia [28], Coverage [29], DSO [30] and IMD [31] with 920, 160, 100, 100 and 2010 forged images, respectively. Due to limited computing resources, we follow the prior work [6] to crop some oversized images to 1096 × 1440 pixels for preparing the test images.

The attack performance is tested against six state-of-the-art image tampering localization algorithms: OSN [2], MVSS-Net [3], PSCC-Net [4], CAT-Net [5], TruFor [6] and Noiseprint [7]. The first three are used as target localizers to generate adversarial examples. All the above localizers use officially published models. Note that Noiseprint can not work on the CASIAv1 and IMD images due to too uniform content or small resolution.

Evaluation metrics. The localization accuracy metrics, i.e., F1 and Intersection over Union (IoU) are widely adopted by the existing works on tampering localization. F1 is defined as the harmonic average of precision and recall rate. That is,

$$\mathrm{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

where $Precision = TP/(TP + FP)$ and Recall $= TP/(TP + FN)$. FP denotes the number of wrongly classified pristine pixels in a test image. TP and FN denote the numbers of correctly and wrongly classified forged pixels, respectively. IoU measures the similarity between the predicted area and the ground truth as

$$\mathrm{IoU} \ = \ TP/(TP + FP + FN) \tag{7}$$

The threshold for calculating F1 and IoU is set to 0.5, which is a common setting for most tampering localization algorithms. Such metric values before and after the attack, and their decrease rate are computed to evaluate the performance of attack methods. The decrease rates of F1 and IoU scores, denoted by d_{F1} and d_{IoU} respectively, are defined as the ratio between the decrement value and the measurement before attacks. That is,

$$\begin{cases} d_{\mathrm{F1}} = \left(\mathrm{F1}_{before} - \mathrm{F1}_{after}\right)/\mathrm{F1}_{before} \\ d_{\mathrm{IoU}} = \left(\mathrm{IoU}_{before} - \mathrm{IoU}_{after}\right)/\mathrm{IoU}_{before} \end{cases} \tag{8}$$

Meanwhile, Peak Signal-to-Noise Ratio (PSNR) [32] and Structural Similarity (SSIM) [33] are used to evaluate the visual quality of attacked images.

Parameter Setting. In both two attacks, the adversarial example is initialized with the forged image, i.e., $x_0^* = x$. In the optimization-based attack, the Adam optimizer is implemented with a learning rate of 0.003 and the number of iterations is set to 30 epochs. $\lambda = 0, 1e^{-6}, 1e^{-5}$ are respectively set in opt-OSN, opt-MVSS and opt-PSCC to achieve a good attack performance while maintaining a certain level of visual quality. When $\lambda = 0$, the loss function does not limit the magnitude of the perturbation. It is pointed out that when x^* is initialized with x , even if the size of the perturbation is not limited in the loss function, the adversarial example generated by the Adam optimizer is still around the input image forgery x [34]. In the gradient-based attack, the step size ε is set as 0.02, 0.02, 0.001, 0.01 and 0.01 for the five datasets, respectively. It is set to achieve higher attack performance while maintaining similar PSNR values. The parameters are searched through a series of preliminary experiments. All the PSNR values are greater than 34 dB, and no obvious perturbation is visible to the human eyes.

Table 1
Localization accuracy and visual quality comparison before and after the optimization (Opt) and gradient (Grad) based attack on CASIAv1 dataset. The results of white-box attacks are in bold. d_{F1} and d_{IoU} values are in percentage.

Attack Method		Before	**Opt-OSN**	**Opt-MVSS**	**Opt-PSCC**	**Grad-OSN**	**Grad-MVSS**	**Grad-PSCC**
F1	OSN [2]	0.51	**0.05**	0.23	0.33	**0.16**	0.20	0.30
(d_{F1})			**(90)**	(55)	(35)	**(69)**	(61)	(41)
	MVSS-Net [3]	0.45	0.13	**0.03**	0.19	0.17	**0.09**	0.18
			(71)	**(93)**	(58)	(62)	**(81)**	(60)
	PSCCNet [4]	0.46	0.23	0.23	**0.13**	0.26	0.24	**0.15**
			(50)	(50)	**(72)**	(43)	(47)	**(68)**
	Average	0.47	0.14	0.16	0.22	0.20	0.18	0.21
			(70)	(66)	(53)	(57)	(62)	(55)
IoU	OSN [2]	0.47	**0.04**	0.20	0.29	**0.12**	0.17	0.26
(d_{IoU})			**(91)**	(57)	(37)	**(74)**	(64)	(44)
	MVSS-Net [3]	0.40	0.10	**0.02**	0.15	0.14	**0.06**	0.15
			(75)	**(95)**	(61)	(65)	**(84)**	(63)
	PSCC-Net [4]	0.41	0.20	0.20	**0.11**	0.22	0.21	**0.10**
			(51)	(51)	**(73)**	(46)	(49)	**(76)**
	Average	0.43	0.11	0.14	0.18	0.16	0.15	0.17
			(74)	(67)	(58)	(62)	(66)	(60)
PSNR		—	35.54	35.20	37.62	35.04	35.31	35.24
SSIM		—	0.95	0.95	0.97	0.94	0.95	0.94

Table 2
Optimization and gradient-based attacks performance on localization accuracy and visual quality on more other datasets and localizers. The results of white-box attacks are in bold. d_{F1} and d_{IoU} values are in percentage.

Dataset			Columbia			Coverage			DSO			IMD	
Attack Method		Before	Opt	Grad	Before	Opt	Grad	Before	Opt	Grad	Before	Opt	Grad
F1 (d_{F1})	OSN[2]	0.71	**0.12(83)**	**0.34(52)**	0.26	**0.11(58)**	**0.13(52)**	0.47	**0.01(98)**	**0.06(87)**	0.50	**0.04(92)**	**0.11(78)**
	MVSS-Net[3]	0.64	0.55(14)	0.56(13)	0.45	0.21(54)	0.24(48)	0.30	0.17(43)	0.22(28)	0.27	0.11(59)	0.16(40)
	PSCC-Net[4]	0.62	0.36(41)	0.50(19)	0.44	0.13(71)	0.13(70)	0.53	0.00(100)	0.00(100)	0.16	0.01(94)	0.02(89)
	CAT-Net[5]	0.79	0.91(-15)	0.92(-16)	0.29	0.34(-17)	0.33(-15)	0.33	0.04(88)	0.07(80)	0.67	0.20(70)	0.27(60)
	TruFor[6]	0.81	0.73(10)	0.71(12)	0.53	0.34(35)	0.36(32)	0.90	0.35(62)	0.41(55)	0.72	0.43(41)	0.46(37)
	Noiseprint[7]	0.36	0.16(56)	0.13(63)	0.15	0.12(20)	0.15(-3)	0.29	0.04(86)	0.05(84)	—	—	—
	Average	0.65	0.47(28)	0.53(20)	0.35	0.21(40)	0.22(37)	0.47	0.10(79)	0.13(72)	0.46	0.16(65)	0.20(57)
IoU (d_{IoU})	OSN[2]	0.61	**0.09(85)**	**0.25(60)**	0.18	**0.08(55)**	**0.09(49)**	0.34	**0.00(100)**	**0.03(90)**	0.40	**0.03(93)**	**0.07(83)**
	MVSS-Net[3]	0.60	0.45(24)	0.46(24)	0.38	0.17(56)	0.19(51)	0.22	0.12(45)	0.16(29)	0.21	0.08(62)	0.12(44)
	PSCC-Net[4]	0.48	0.27(44)	0.39(18)	0.34	0.11(67)	0.11(69)	0.42	0.00(100)	0.00(100)	0.13	0.01(92)	0.01(92)
	CAT-Net[5]	0.75	0.88(-18)	0.90(-20)	0.23	0.26(-13)	0.26(-12)	0.28	0.03(89)	0.04(85)	0.59	0.15(75)	0.21(64)
	TruForp[6]	0.75	0.64(15)	0.62(16)	0.45	0.28(39)	0.29(36)	0.85	0.27(68)	0.31(63)	0.63	0.34(46)	0.37(42)
	Noiseprint[7]	0.26	0.09(65)	0.08(70)	0.09	0.07(21)	0.09(-4)	0.21	0.02(90)	0.02(90)	—	—	—
	Average	0.57	0.40(29)	0.45(21)	0.28	0.16(43)	0.17(39)	0.39	0.07(82)	0.10(74)	0.39	0.12(69)	0.15(62)
PSNR		—	35.34	34.93	—	34.19	34.20	—	35.02	36.51	—	35.00	36.90
SSIM		—	0.87	0.85	—	0.94	0.94	—	0.87	0.91	—	0.89	0.94

3.2. Influence of target localizer in white-box attack

Firstly, the attacks use MVSS-Net, OSN and PSCC-Net as the target localizer respectively to geneate adversarial examples on CASIAv1 dataset. Table 1 shows the F1, IoU and decrease rate before and after the optimization-based (Opt) and gradient-based (Grad) white-box attacks and their transferability to other victim localizers. It can be observed that both attacks can significantly reduce the image tampering localization accuracy. The maximum decrease rate of F1 and IoU can exceed 93 % and 95 % respectively in the white-box scenario. The knowledge about the victim localizer can be fully accessed, so that white-box attacks can significantly degrade the performance of the victim localizers.

In addition, both attacks show strong transferability. F1 and IoU are reduced by at least 35 % and 37 % respectively in the optimization-based attack, and by at least 41 % and 44 % respectively in the gradient-based attack. As can be seen from Table 1, the average F1 after the optimization-based attack is 0.14, 0.16 and 0.22. The average F1 after gradient-based attack is 0.20, 0.18, and 0.21. The selection of the target localizer has no obvious impact on the attack effect. The adversarial examples generated against OSN and MVSS-Net have the equal attack performance, while the adversarial examples generated against PSCC-Net are slightly less effective. Adversarial examples generated against PSCC-Net decrease the F1 by an average of 53 % and 55 % in each attack. The adversarial examples generated against OSN and MVSS-Net can decrease the F1 by an average of 70 % and 66 % in the optimization-based attack, 57 % and 62 % in the gradient-based attack. For the sake of consistency and without loss of generality, OSN is chosen as the target localizer in all the following experiments to generate adversarial examples.

3.3. Transferability in black-box attack

In the above subsection, the adversarial examples generated against the target localizers are proven to be effective and exhibit a certain level of transferability. To further demonstrate the transferability of our proposed attack method, we evaluate the effectiveness of optimization and gradient-based attacks on other datasets and victim localizers. The attack performance against six tampering localization algorithms on four different datasets is presented in Table 2. The adversarial examples generated against the target localizer OSN transfer well to other victim localizers, demonstrating a good attack performance in the black-box scenario.

We can see from Table 2 that the optimization-based attack can reduce F1 scores by an average of 28 %, 40 %, 79 % and 65 % on different datasets, respectively. The black-box attacks on the Columbia dataset reduce the F1 scores of PSCC-Net from 0.62 to 0.36 and 0.50. The adversarial examples generated on the IMD dataset reduce the IoU of PSCC-Net from 0.13 to 0.01. As for the DSO dataset, optimization and

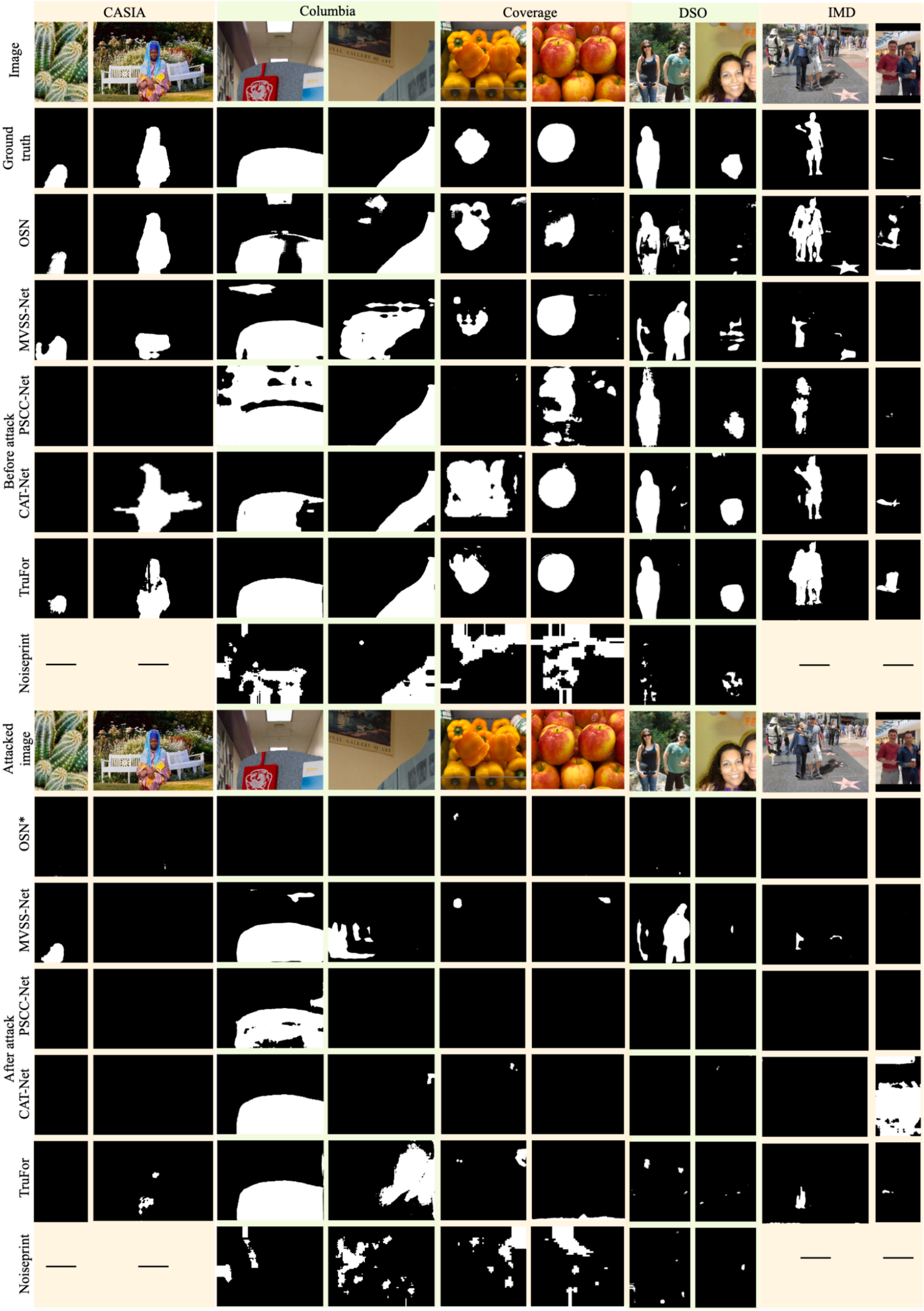


Fig. 2. Forensic results of different tampering localization algorithms before and after applying the optimization-based attack to ten example forged images from five datasets. Here, '*' denotes white-box attacks. '-' means inapplicable due to computation restriction.

Table 3
Performance comparison with other attack methods on CASIAv1. The results of white-box attacks are in bold. d_{F1} and d_{IoU} values are in percentage.

Attack Method		Before	**JPEG** [35]	**Median** [35]	Median- JPEG	Opt	Grad
F1	OSN[2]	0.51	0.26	0.37	0.18	**0.06**	**0.16**
(d_{F1})			(48)	(28)	(64)	**(88)**	**(69)**
	MVSS-Net[3]	0.45	0.15	0.39	0.18	0.12	0.17
			(68)	(14)	(59)	(73)	(62)
	PSCC-Net[4]	0.46	0.18	0.26	0.03	0.24	0.26
			(62)	(44)	(93)	(48)	(43)
	CAT-Net[5]	0.72	0.29	0.12	0.17	0.41	0.43
			(59)	(83)	(76)	(43)	(40)
	TruFor[6]	0.69	0.57	0.51	0.44	0.44	0.49
			(17)	(27)	(37)	(36)	(29)
	Average	0.57	0.29	0.33	0.20	0.25	0.30
			(49)	(42)	(65)	(56)	(47)
IoU	OSN[2]	0.47	0.23	0.31	0.15	**0.04**	**0.12**
(d_{IoU})			(51)	(33)	(67)	**(91)**	**(74)**
	MVSS-Net[3]	0.40	0.12	0.33	0.14	0.09	0.14
			(71)	(17)	(64)	(78)	(65)
	PSCC-Net[4]	0.41	0.14	0.19	0.03	0.21	0.22
			(66)	(53)	(94)	(49)	(46)
	CAT-Net[5]	0.64	0.24	0.09	0.13	0.35	0.37
			(62)	(86)	(80)	(45)	(42)
	TruFor[6]	0.63	0.50	0.45	0.38	0.39	0.43
			(20)	(28)	(40)	(38)	(32)
	Average	0.51	0.25	0.27	0.17	0.22	0.26
			(51)	(47)	(67)	(57)	(49)
PSNR		—	30.43	26.87	26.06	35.54	35.04
SSIM		—	0.93	0.83	0.80	0.95	0.94

gradient-based black-box attacks can reduce the F1 scores of TruFor from 0.90 to 0.35 and 0.41, respectively.

It can be concluded that in most cases, the adversarial examples generated against OSN based on optimization and gradient perform well in black-box scenario. The black-box attack only uses the adversarial examples generated in the white-box scenario, and the victim localizer is not similar to the target localizer. That's why the attack performance is not as obviously in black-box scenarios when compared to white-box attacks. Qualitative evaluation results of the proposed optimization-based attack against different localization algorithms are shown in Fig. 2. Note that such localization algorithms mostly perform well on the ten forged example images before the attack, except some failure cases. However, the localizers fail to find and locate the altered regions accurately, which can be validated by the almost black prediction masks. Meanwhile, the subtle visual distortion incurred by adversarial perturbation is imperceptible. Such visual evaluation results further verify the effectiveness of our proposed attack methods.

3.4. *Performance comparison with other attacks*

The proposed attack methods are compared with common post-processing attacks on CASIAv1 dataset. JPEG compression [35] with the factor of 55, median filter [35] with the kernel 3 × 3 and JPEG compression after median filtering are tested. The comparison results are shown in Table 3.

Compared with JPEG compression and median filtering attacks, adversarial attack can reduce the accuracy of the tampering localization algorithms while maintaining better visual quality. In the optimization

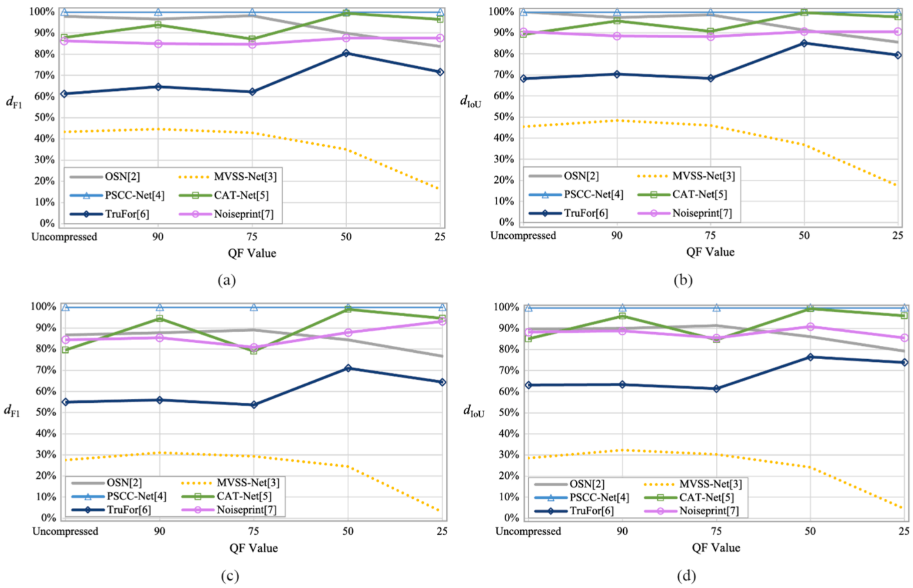
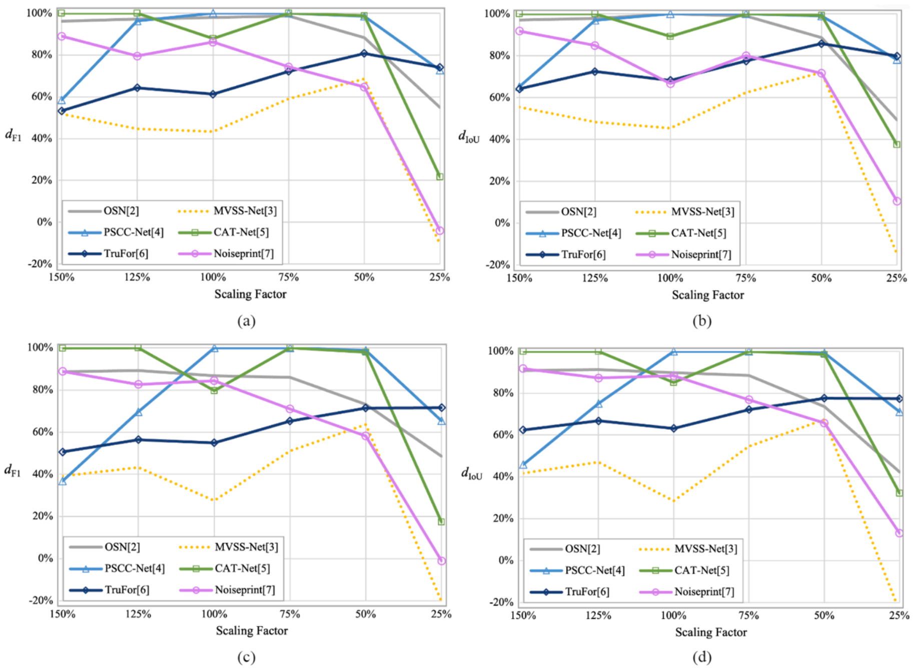


Fig. 3. Robustness against post JPEG compression with different quality factors (QFs) on DSO dataset. F1 and IoU decrease rates are for the optimization-based (a)(b) and gradient-based (c)(d) attacks, respectively.



(a) (b) (c) (d)

Fig. 4. Robustness against post resizing with different scaling factors on DSO dataset. F1 and IoU decrease rates are for the optimization-based (a)(b) and gradient-based (c)(d) attacks, respectively.

and gradient based attack, the average decrease rate of F1 and IoU can reach to 56 %, 47 % and 57 %, 49 %. JPEG compression after median filtering has the best attack performance. The average F1 score decrease from 0.57 to 0.20 and the IoU from 0.51 to 0.17. The localization accuracy of TruFor has reduced by about 40 %, all the other localizers have reduced by at least 59 %. However, such combined post-processing attack also leads to severely degraded images. The average PSNR of attacked images is only 26.06 dB, while the PSNR of the adversarial examples can reach to 35 dB. This indicates that the method sacrifices too much visual quality in order to improve attack performance. As a result, comparing with the existing attacks, our proposed optimization-based and gradient-based methods achieve higher performance.

3.5. *Robustness analysis*

To further demonstrate the robustness of the adversarial examples generated by the proposed attack methods, the attack performance of the adversarial examples after certain post-processing operations is tested in this subsection. Specifically, JPEG compression and resizing are applied to the generated adversarial examples. The DSO dataset is chosen for this part of the experiment due to the proposed attack algorithm demonstrating the best performance on this dataset.

First, the adversarial examples are subjected to post JPEG compression with Quality Factors (QF) of 90, 75 and 50, respectively. The attack performance after JPEG compression is illustrated in Fig. 3. In order to demonstrate the attack effectiveness more clearly, this figure only shows the decrease rates of F1 and IoU scores. A higher decrease rate demonstrates better attack performance. As can be seen from the figure, the effectiveness of the attack does not decrease significantly after JPEG compression. In the optimization-based attack, the attack performance on PSCC-Net, CAT-Net and Noiseprint after JPEG compression almost does not decrease. However, after JPEG compression, the attack performance on OSN and MVSS-Net is reduced. As QF increases, the attack performance becomes worse. A similar result can be observed for the gradient-based attack.

Next, the attack performance after post resizing operation is tested. The adversarial example images are scaled via bilinear interpolation to 150 %, 125 %, 75 % and 50 % of their original size while keeping the aspect ratio, respectively. The results of the attack performance after post resizing are detailed in Fig. 4. It can be observed that after resizing, the attack performance on OSN, CAT-Net, and TurFor initially increases, followed by a decreasing trend. Conversely, the attack performance on PSCC-Net and Noiseprint decreases as the image is resized larger or smaller. Compared to JPEG compression, resizing will lead to significant fluctuations in attack performance. The robustness of adversarial examples to resizing is somewhat lower compared to the robustness to JPEG compression.

4. Conclusion

In this work, we propose an effective adversarial attack scheme to evaluate the security of state-of-the-art image tampering localization algorithms. Both local consistency-based and segmentation-based tampering localizers are used as victim localizers. The attack on tampering localizers is first formulated formally, then two specific adversarial attack methods, optimization-based and gradient-based attacks, are presented under the unified attack framework. Adversarial examples are generated against the target localizer in the white-box attack. The black-box attack exploit the transferability of adversarial examples. In both white and black-box scenarios, the accuracies of the victim localizers are significantly reduced by the proposed attacks. Meanwhile, the adversarial example images enjoy good transferability and visual transparency. Our attack methods also outperform other existing attacks and demonstrate a degree of robustness to JPEG compression and resizing operations.

CRediT authorship contribution statement

Gang Cao: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Yuqi Wang:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Haochen Zhu:** Validation, Software, Methodology, Data curation. **Zijie Lou:** Visualization, Software, Investigation, Formal analysis. **Lifang Yu:** Writing – review & editing, Resources, Methodology, Conceptualization.

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.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62071434, 62172005), Fundamental Research Funds for the Central Universities (CUC24GT01), CUC Public Computing Cloud.

References

[1] L. Zheng, Y. Zhang, V.L. Thing, A survey on image tampering and its detection in real-world photos, Journal of Visual Communication and Image Representation 58 (2019) 380–399.

[2] H. Wu, J. Zhou, J. Tian, J. Liu, Y. Qiao, Robust image forgery detection against transmission over online social networks, IEEE Trans. Inf. Forensics Secur. 15 (2022) 443–456.

[3] X. Chen, C. Dong, J. Ji, J. Cao, X. Li, "Image manipulation detection by multi-view multi-scale supervision", in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14185–14193.

[4] X. Liu, Y. Liu, J. Chen, X. Liu, PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization, IEEE Trans. Circuits Syst. Video Technol. 32 (11) (2022) 7505–7517.

[5] M.J. Kwon, S.H. Nam, I.J. Yu, H.K. Lee, C. Kim, Learning JPEG compression artifacts for image manipulation detection and localization, Int. J. Comput. Vis. 130 (8) (2022) 1875–1895.

[6] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, L. Verdoliva, "TruFor: Leveraging all-round clues for trust-worthy image forgery detection and localization", in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20606–20615.

[7] D. Cozzolino, L. Verdoliva, Noiseprint: A CNN-based camera model fingerprint, IEEE Trans. Inf. Forensics Secur. 15 (2019) 144–159.

[8] M. Huh, A. Liu, A. Owens, A.A. Efros, "Fighting fake news: Image splice detection via learned self-consistency", in: *Proceedings of the European Conference on Computer Vision*, 2018, pp. 101–117.

[9] O. Mayer, M.C. Stamm, Exposing fake images with forensic similarity graphs, IEEE J. Sel. Top. Signal Process. 14 (5) (2020) 1049–1064.

[10] Y. Wu, W. AbdAlmageed, P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features", in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9543–9552.

[11] Z. Chen, B. Tondi, X. Li, R. Ni, Y. Zhao, M. Barni, "A gradient-based pixel-domain attack against SVM detection of global image manipulations", in: *Proceedings of the IEEE Workshop on Information Forensics and Security*, 2017, pp. 1–6.

[12] G. Cao, Y. Zhao, R. Ni, H. Tian, "Anti-forensics of contrast enhancement in digital images", in: *Proceedings of the 12th ACM Workshop on Multimedia and Security*, 2010, pp. 25–34.

[13] G. Cao, Y. Zhao, R. Ni, H. Tian, L. Yu, Attacking contrast enhancement forensics in digital images, SCIENCE CHINA Inf. Sci. 57 (2014) 1–13.

[14] G. Cao, Y. Wang, Y. Zhao, R. Ni, C. Lin, "On the security of image manipulation forensics", in: *Proceedings of the 16th Pacific-Rim Conference on Multimedia*, 2015, pp. 97–105.

[15] H. Xie, J. Ni, Y.Q. Shi, Dual-domain generative adversarial network for digital image operation anti-forensics, IEEE Trans. Circuits Syst. Video Technol. 32 (3) (2021) 1701–1706.

[16] S. Fang, and M. C. Stamm, "Attacking image splicing detection and localization algorithms using synthetic traces," *arXiv preprint arXiv*:2211.12314, 2022.

[17] X. Mo, S. Tan, B. Li, J. Huang, "Poster: Query-efficient black-box attack for image forgery localization via reinforcement learning", in: *Proceedings of ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3552–3554.

[18] M. Barni, K. Kallas, E. Nowroozi, B. Tondi, "On the transferability of adversarial examples against CNN-based image forensics", in: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 8286–8290.

[19] D. Güera, Y. Wang, L. Bondi, P. Bestagini, S. Tubaro, E.J. Delp, "A counter-forensic method for CNN-based camera model identification", in: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 1840–1847.

[20] B. Tondi, Pixel-domain adversarial examples against CNN-based manipulation detectors, Electron. Lett 54 (21) (2018) 1220–1222.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv*: 1412. 6572, 2014.

[22] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, "The limitations of deep learning in adversarial settings", in: *Proceedings of IEEE European Symposium on Security and Privacy*, 2016, pp. 372–387.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv*: 1706.06083, 2017.

[24] F. Marra, D. Gragnaniello, L. Verdoliva, On the vulnerability of deep learning to adversarial attacks for camera model identification, Signal Process. Image Commun. 65 (2018) 240–248.

[25] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv*: 1412.6980, 2014.

[26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv*: 1312.6199, 2013.

[27] J. Dong, W. Wang, T. Tan, "Casia image tampering detection evaluation database", in: *Proceedings of IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 422–426.

[28] T.T. Ng, S.F. Chang, Q. Sun, A data set of authentic and spliced image blocks, Columbia University, ADVENT Technical Report, 2004.

[29] B. Wen, Y. Zhu, R. Subramanian, T.T. Ng, X. Shen, S. Winkler, "COVERAGE—A novel database for copy-move forgery detection", in: *Proceedings of IEEE International Conference on Image Processing*, 2016, pp. 161–165.

[30] D. Carvalho, T. Jose, C. Riess, E. Angelopoulou, H. Pedrini, A. de Rezende Rocha, Exposing digital image forgeries by illumination color classification, IEEE Trans. Inf. Forensics Secur. 8 (7) (2013) 1182–1194.

[31] A. Novozamsky, B. Mahdian, S. Saic, "IMD2020: A large-scale annotated dataset tailored for detecting manipulated images", in: *Proceedings of IEEE Winter Conference on Applications of Computer Vision Workshops*, 2020, pp. 71–80.

[32] I. Avcibas, B. Sankur, K. Sayood, Statistical evaluation of image quality measures, J. Electron. Imaging 11 (2) (2002) 206–223.

[33] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[34] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *arXiv preprint arXiv*:1611.02770, 2016.

[35] H. Li, W. Luo, X. Qiu, J. Huang, Identification of various image operations using residual-based features, IEEE Trans. Circuits Syst. Video Technol. 28 (1) (2016) 31–45.