MTISA: Multi-Target Image-Scaling Attack

Jiaming He*, Hongwei Li†, Wenbo Jiang(Corresponding Author)†, Yuan Zhang†
*College of Computer Science and Cyber Security (Oxford Brookes College), Chengdu University of Technology, China
†School of Computer Science and Engineering, University of Electronic Science and Technology of China, China

***Abstract*—Image scaling is one of the most common operations in image processing. For instance, it is often conducted before image transferring to preserve resources, image classifiers also require images to be input at a specified size. However, potential threats may come out with the image scaling operation. A recent work called image-scaling attack can change the semantic information of the input image when it is scaled to a specific size. For example, a manipulated image of a sheep may become an image of a wolf when it scales to a specific size.**

Many works have already demonstrated the effectiveness of this attack and the security risks it poses. However, existing image-scaling attacks only focus on single target with single specific size, and are not applicable to multi-target image-scaling attack. In this paper, we present a *multi-target image-scaling attack (MTISA)*. *MTISA* can be trained with a single image performs diverse and semantically distinct outputs to fool both human vision and image classifiers. Specifically, to fool human vision, we employ SinGAN to generate semantically different but background-similar samples to serve as the attack target samples. To mislead image classifiers, we employ adversarial attacks to construct adversarial examples to serve as the attack target samples. Finally, we evaluate *MTISA* on chest X-rays dataset and ImageNet dataset, respectively. The experimental results demonstrate that *MTISA* achieves high attack success rate against both human vision and image classifiers.

***Index Terms*—Image-scaling attack, Deep learning, SinGAN**

I. Introduction

In recent years, deep learning technology has a significant progress across areas from image classification to object detection. Image scaling operations are widely used in these application areas. For instance, before the images are input into these deep neural network (DNN) models, the input images need to be scaled to the fixed input size. However, these scaling operations are vulnerable to image-scaling attack [1], where the attacker adds specific perturbations to the normal image, and the semantic information on the image will be changed after the scaling operation.

Image-scaling attacks have been demonstrated by many works [1]–[3] and pose serious security risks in the field of computer vision. However, the existing image-scaling attacks only can accord to the single specific size and can't maintain attack performance across various sizes. Hence, a natural question arises: *Can image-scaling attacks achieve multi-target on various sizes?* Compared with previous works, adding multiple targets in image-scaling attacks can enhance the flexibility and performance of the attacks. However, the challenge is that introducing multiple targets might decrease the invisibility of attack images and the quality of output images. Xiao et al. proposed a method called *probeImg* [1], which divides the image into several areas, and the corresponding area on the image will display when the attack image is scaled to different sizes. Nevertheless, the attack image after scaling cannot be fully displayed within the entire image area when employing *probeImg* as an attack mechanism.

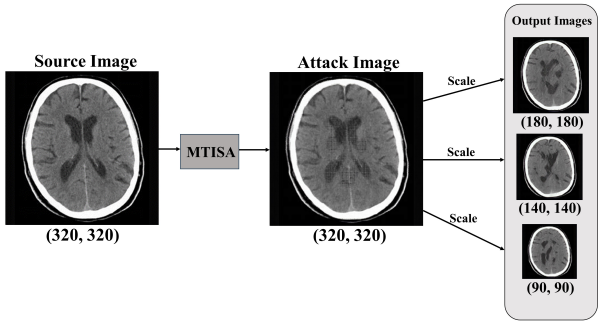


Fig. 1. Example of *MTISA* against human vision.

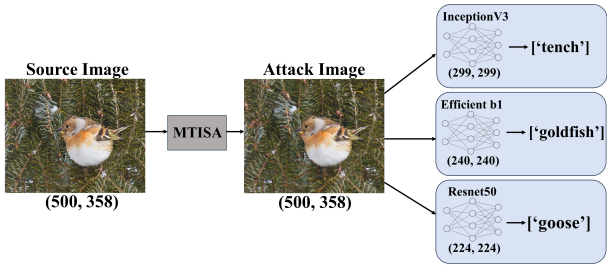


Fig. 2. Example of *MTISA* against multiple image classifiers.

To address this problem, in this paper, we propose *multi-target image-scaling attack (MTISA)*, a more flexible and powerful image-scaling attack with diverse targets. Our proposed attacks can produce attack images to present diverse semantic information during the multiple scaling operations. As illustrated in Figure 1 and Figure 2, we consider two types of attack targets based on different scenarios, the *MTISA* against human vision and the *MTISA* against image classifiers. For *MTISA* against human vision, firstly, we apply SinGAN [4] to generate semantically different but background similar samples. Then we set the similarity restriction to control the variation of generated samples. Finally, we inject the generated samples as attack targets into the source images. For *MTISA* against image classifiers, our proposed attacks assisted with adversarial attacks can mislead diverse DNN-based classifiers when the attack images are input into diverse DNN models. Our attacks can be flexible to various DNN models like InceptionV3 [5], ResNet [6], EfficientNet [7], etc.

Our contributions can be summarised as follows:

- We present a multi-target image-scaling attack against human vision. Specifically, we employ SinGAN to generate semantically different but background-similar samples to serve as the attack target samples. The semantic information of the attack image is changed to diverse semantic information when the image is scaled to various sizes.
- We further present a multi-target image-scaling attack against image classifier using adversarial attack. During the inference process, the predictive result of the attack image is changed to different labels when the image is scaled to various sizes.
- We conduct extensive evaluations to demonstrate the effectiveness of *MTISA*. For *MTISA* against human vision, the average similarity between attack images and source images achieves 0.94. For *MTISA* against image classifiers, the attack success rate achieves 100%.

II. Background

A. Image-Scaling Attack

The image-scaling attack is a recently proposed attack method primarily aimed at misleading DNN-based image classifiers. Its effectiveness stems from the necessity of scaling input images to match the required input sizes of DNN models. And the image-scaling attack can be assisted with other attacks(e.g., backdoor attack [8] [9] and adversarial attack [10]). Through the image scaling operation, pixels are selectively retained or eliminated using appropriate scaling algorithms. Adversary can capitalize on these distinctive features, effectively concealing smaller-sized images within the scaled counterparts, thus incorporating diverse semantic information into the source images in various sizes. This strategic approach finally aims to mislead the classifier's interpretation and decision-making processes. The source image I_s is added the perturbation Δ_{isa} as the attack image I_{isa} = (I_s + Δ_{isa}). After scaling the image is similar to the target image I_t. The adversary can implement an attack by following the quadratic optimization problem:

$$\min \left(\|\Delta_{isa}\|_2^2\right) \ \ s.t. \ \|scale\left(I_s + \Delta_{isa}\right) - I_t\|_\infty \leqslant \varepsilon. \quad (1)$$

B. Adversarial Attack

Adversarial example proposed by Szegedy et al. [11] is designed for the image classifying task which can mislead the model to output incorrect label. The objective of the adversarial attack is to find a specific perturbation Δ_{adv} and add it to the source image I_s to get the adversarial example I_{adv}. This adversarial example I_{adv} causes the model to produce an incorrect output. (see Equation (2)).

$$\begin{cases} I_{adv} = I_s + \Delta_{adv} \\ f_{model}(I_{adv}) \neq Lable_s \end{cases} \quad (2)$$

There are various methods to generate adversarial examples, such as gradient-based methods Fast Gradient Sign Method (FGSM) [12] and iterate attack method Projected Gradient Descent (PGD) [13], are employed to efficiently discover the perturbation within a given threat model. In this work, We utilize different types of adversarial attacks to assist the multi-target image-scaling attacks.

III. Multi-target image-scaling attack

A. Threat Model

In real-world scenarios, the adversary can adapt their approach based on the specific scenario to customize distinct image-scaling attacks with multiple targeted objectives. As Figure 3 shows, the process enables the practical implementation of our proposed novel image-scaling attack with multi-target capabilities, where the adversary can adjust the attack strategy based on the unique characteristics of the scenario.

- **In the scenario of transferring a large number of digital images.** The compress operations are common in the transferring stage of large scale data (e.g., upload or storage images). The adversary can implement our proposed multi-target image-scaling attacks to reach the goal of destroying the original images or modifying the original images to mislead the image users.
- **In the scenario of image classifying.** Images need to satisfy the specific input sizes of image classifier models, so the adversary can leverage this operation to implement our proposed attacks.

Besides, the image-scaling attack also needs to satisfy the conditions that the attack image I_{isa} is visually similar to the source image I_s and the output image I_o is visually close to the target image I_t.

B. Multi-Target Image-Scaling Attacks against Human Vision

1) ***Challenges to introduce multiple targets in image-scaling attacks****:* We try introducing more targets with existing image-scaling attacks and find that the existing image-scaling attacks do not enable to introduce multiple targets. If the adversary repeatedly implements attacks on the source images using multiple target images of distinct sizes sourced from the same dataset, it is observed that the resulting attack image will exhibit blurriness, and the quality of the image significantly declines when it is scaled to the set size. This degradation is attributed to the introduction of multiple target images with huge dissimilarities in pixels. Furthermore, one simple method can be employed to improve the invisibility of attack is to increase the size of the source images while decreasing the sizes of the target images at the same time. However, this approach results in a significant disparity of size between the attack image and the output image. Consequently, the effectiveness of the attack decreases significantly as a consequence of this disparity of size and the low-resolution target image.

2) ***Insight of our MTISA****:* In real-world scenarios, there is no need to set the target images as completely different real images. Changing the semantic information in image is key to fooling human vision (e.g., Changing the position of shadow in chest X-rays). We present a novel multi-target image-scaling attack that assisted with SinGAN, which is an unconditional generative model that can be trained using a

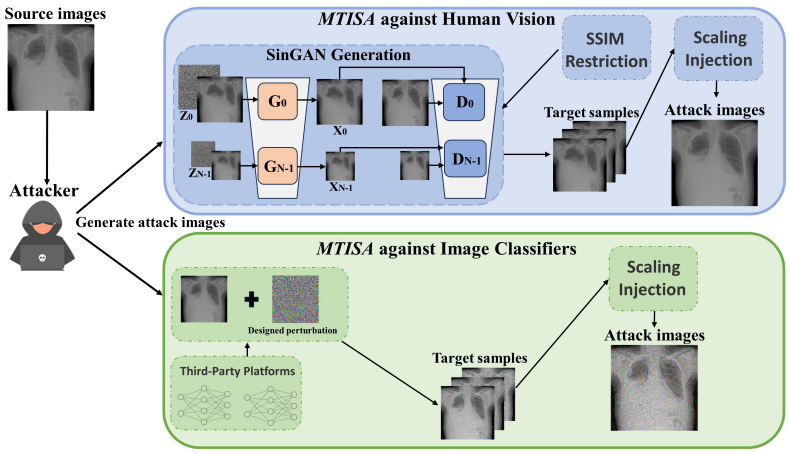


Fig. 3. The workflow of *MTISA* against human vision and image classifiers.

single natural image. By leveraging SinGAN's capabilities, we aim to bridge the mutual influence between different target images. Following the multiple times scaling operation to the attack image, the resulting output images may exhibit diverse semantic information compared to the source image. Consequently, the original semantic information present in the source image may become changed or destroyed in the multiple scaling processes. After multiple times attack, The perturbation Δ_{isa} will be minimal if the disparity σ between source image I_s and target image I_t is as possible few.

$$\begin{cases} \sigma = scale(I_s) - \sum_{t=1}^{N_t} I_t \\ \Delta_{isa} \propto \sigma \end{cases} \tag{3}$$

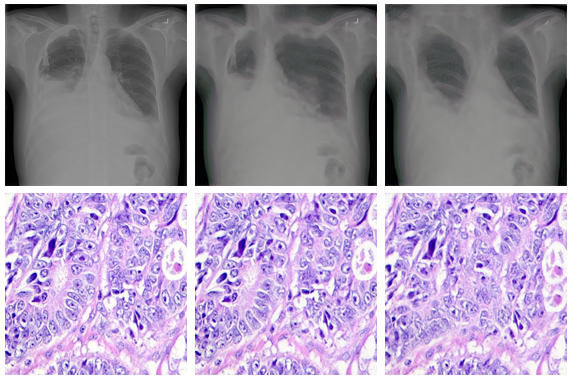
With the perturbation Δ_{isa} getting reduced, the mutual influence between different target images I_t will decrease.

*3) **The process of MTISA based on SinGAN**:* For reducing perturbation Δ_{isa}, we propose the multi-target image-scaling attacks assisted with SinGAN. The generated samples are the result of applying modifications to the source image I_s, then we can inject the samples into the source image I_s. The attack images generation can be formulated as follows:

$$\underset{scale}{arg\ min} \|I_t\|_{scale=0}^{scale=1}\ s.t.\ scale(I_s) - I_t \leqslant SSIM_{min} \tag{4}$$

We consider that target images I_t should have high similarity in pixels with the source image I_s. So we set a minimal score SSIM $SSIM_{min}$ to restrict the scaling inject process.

For adjusting the generated samples' degree of variation to improve and reach the best performance of the attack's effectiveness and invisibility, the adversary can control the starting scale in SinGAN's architecture. The starting scale can decide the detected range of the image. With an increasing number of scales, SinGAN becomes capable of capturing larger structures and the overall spatial arrangement of objects within the source image. In other words, the variation of generated samples can be controlled by adjusting the starting scale. For different attack targets (e.g., medical images, social media images), the adversary can improve the attack to have a great performance by controlling the starting scale. Figure 4 shows the comparison of different kinds of real images and the different starting scales of the generated images.



Real image **Starting scale = 1** **Starting scale = 0**

Fig. 4. The examples of providing images of chest X-rays and tumor biopsies in two starting scales, showcase the contrasting attributes between the real images and generated images of different starting scales.

The "Real/Fake" user unpaired study conducted on Amazon Mechanical Turk (AMT) serves as a metric for evaluating users' ability to discern between generated and real images. Higher scores on this metric indicate increased difficulty in distinguishing between the two image types, thus implying improved invisibility of the attack.

TABLE I
THE CONFUSION RATES OF DIFFERENT STARTING SCALES

Starting Scale	Diversity	Survey	Confusion
0	0.5	unpaired	42.9% ± 0.9%
1	0.35	unpaired	47.04% ± 0.8%

Starting scale 0 refers that the generation starts from the first scale.
Starting scale 1 refers that the generation starts from the second scale.

In Table 1, there are the confusion rates for two different protocols: Starting from the coarsest scale 0, where samples exhibit significant diversity, and subsequently starting from

the second coarsest scale 1, which focuses on preserving the global structure of the original image. The unpaired study which shows the fake or real images altogether shows that the starting scale 1 has a higher confusion rate. It is evident that as the diversity of generated images increases, there is a discernible decrease in the confusion rate according to the AMT test. Notably, the exceptional score achieved by the starting scale 1 in the AMT test signifies its ability to enhance the invisibility of attacks.

As demonstrated in Algorithm 1, the attack algorithm will iterate every source image s in the dataset D and implement the attack on the source images. Moreover, the selection of target images for this attack is limited by the comparison of the score $SSIM(sample, s)$ and the minimum SSIM score $SSIM_{min}$. The primary objective of this constraint is to optimize the invisibility of the attack. Specifically, if the score exceeds the minimum score, the starting scale parameter p is set to 1, implying less degradation or change of the semantic information within the image after scaling.

Algorithm 1 Multi-target image-scaling attacks against human vision

Input: source image dataset D, samples generating function with SinGAN $Gen()$, SSIM metrics function $SSIM()$, starting scale parameter p, number of targets i, The minimal SSIM limited score $SSIM_{min}$, scaling function $ScaleFunc()$, image-scaling attacks with multiple images $ReAttack()$.
Output: attack images A.

$Initialize\ p \leftarrow 0$
for s in D **do**
  for 0 to i **do**
    $sample \leftarrow gen(s, p)$
    if $SSIM(sample, s) \leq SSIM_{min}$ **then**
      $p \leftarrow 1$
      $sample \leftarrow gen(s, p)$
    end if
    $samples[i] \leftarrow sample$
  end for
  $A \leftarrow ReAttack(s, samples)$
  return A
end for

C. Multi-target image-scaling attacks against image classifier

In this scenario, the adversary can accord to different DNN classifiers to make the corresponding sizes of adversarial examples and implement the image-scaling attacks with these adversarial examples. The proposed attack augmented by the adversarial attack methodology provides flexibility in the selection of target labels, allowing for either a random or specific target label to be chosen with the misleading of victim models. This choice depends on the attacker's objective, and the target label can be easily altered through the implementation of diverse attack strategies.

In Algorithm 2, the adversarial examples $advs$ for victim image classifiers are produced with different adversarial attacks, then the adversary can implement the attack assisted with these generated adversarial examples $advs$.

Algorithm 2 Multi-target image-scaling attacks against image classifier

Input: source image dataset D, number of targets i, adversarial attack $AdvAttack()$, pretrained classifier models $models$, image-scaling attacks with multiple images $ReAttack()$.
Output: attack images A.

for s in D **do**
  for 0 to i **do**
    $advs[i] \leftarrow AdvAttack(models[i], s)$
  end for
  $A \leftarrow ReAttack(s, advs)$
  return A
end for

For the first goal of misleading the classifier with a random target label, the adversary can produce adversarial examples by implementing the attack strategy as follows:

$$\max(\Delta_{adv} \leqslant \epsilon)\ s.t.\ f_{model}(I_s + \Delta_{adv}) \neq Lable_s. \quad (5)$$

The perturbation Δ_{adv} is produced to add on the source image to be the adversarial example. Specifically, the design of perturbation Δ_{adv} aims to ensure that it remains below a predefined threshold ϵ. This limit on Δ_{adv} serves to decrease the detectability of the perturbation Δ_{adv}.

For the secondary goal of misleading the classifier with a specific target label, the adversary intends to mislead the model by modifying the original attack towards a specific target label as follows:

$$\max(\Delta_{adv} \leqslant \epsilon)\ s.t.\ f_{model}(I_s + \Delta_{adv}) = Lable_t. \quad (6)$$

The target of attack is modified to guide the generation of the adversarial example with a specific target label. The adversary possesses the capability to set different target labels $Lable_t$ for various image classifiers of varying sizes with different targets to mislead.

IV. Experiments

In this section, We will implement and evaluate our attack by following experiments:

- The invisibility and attack effectiveness of the image-scaling attacks with multiple human vision targets.
- The stability and transferability of the image-scaling attacks with multiple image classifier targets.

A. Dataset & Setup

In our experiments, we implement our multi-target image-scaling attacks with the three different targets in three scaling sizes. For our attacks against human vision, we randomly choose the images from the chest X-rays dataset [14]. We utilize the 400 generated samples of two starting scales to implement the attacks.

For our attacks against the image classifiers, we randomly choose 600 source images and respective labels from the ImageNet [15] dataset. Furthermore, we choose three pretrained models which are InceptionV3, EfficientNet b1, and ResNet50 as our target models. These models have established

TABLE II
THE INPUT SIZES OF MODELS

Model	Input size(pixels * pixels)
InceptionV3	(299, 299)
EfficientNet b1	(240, 240)
ResNet50, VGG16, GoogleNet	(224, 224)

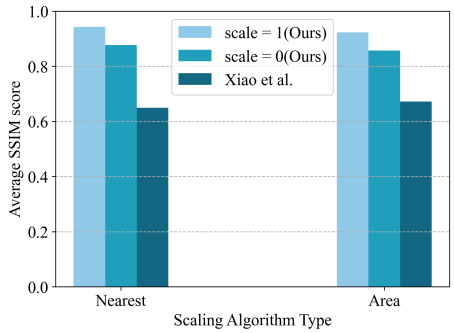


Fig. 5. Invisibility of *MTISA* for different starting scales comparing with Xiao et al. [1].

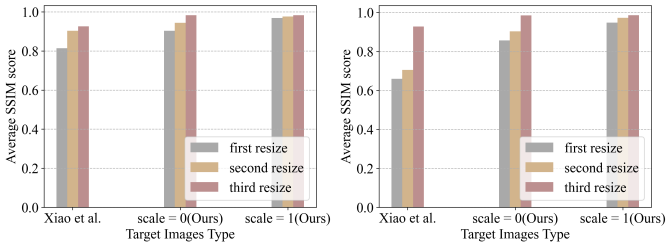
themselves as benchmarks in the field of image recognition and have different input sizes as Table 2 shows.

In the experiments, we choose the Quadratic strategy as the chosen attack strategy and OpenCV's Nearest and Area algorithm as the scaling algorithms in our attack framework.

B. Multi-target image-scaling attacks against human vision

Structural Similarity Index Measure (SSIM) is used as metric to assess the similarity between images. In our experiments, we implement our attack to fool human vision and measure its invisibility by recording the SSIM scores of the attack images and source images. Additionally, we assess the attack effectiveness by analyzing the SSIM scores between three output images and respective target images. And we compare these two metrics with using the image-scaling attack proposed by Xiao et al. [1].

As shown in Figure 5, by observing the SSIM scores, the invisibility of our attacks exhibits superior performance in both scaling algorithms when the starting scale parameter is set to 1 and 0.



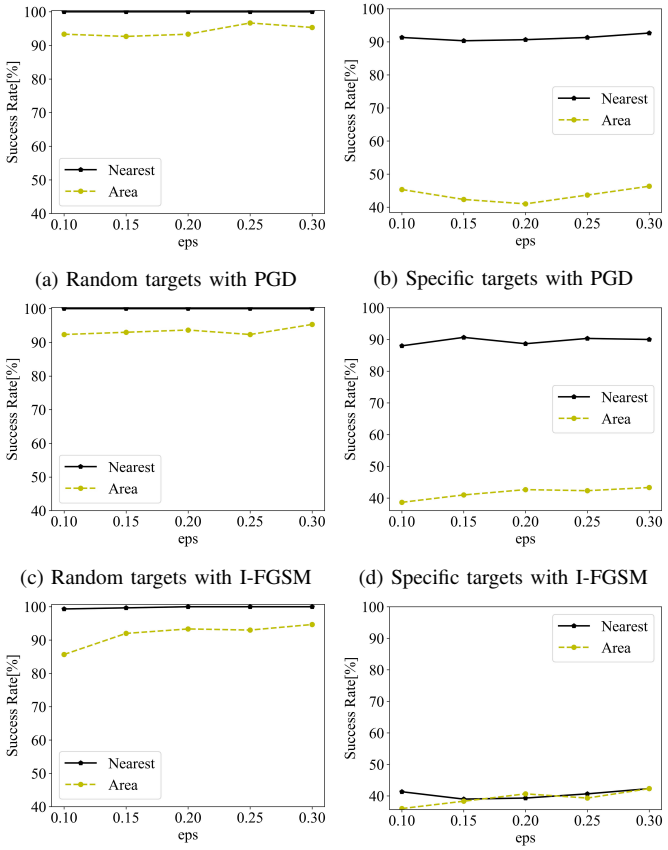
(a) Nearest (b) Area

Fig. 6. Attack effectiveness of *MTISA* for two scaling algorithms comparing with Xiao et al. [1].

In Figure 7, we present a comparison of the SSIM scores between downscaled images and their corresponding target images. Our analysis refers that when the parameter is set to 1 and 0, the SSIM scores remain high and the difference in SSIM scores is relatively insignificant. This observation emphasizes the notable impact of employing generated samples in the attack, indicating a substantial increase in invisibility of the attack.

C. Multi-target image-scaling attacks against image classifier

1) Stability of the attack: In order to evaluate the stability of our attack assisted with PGD [13], I-FGSM [16] and FGSM [12], we assess the success rates of our attacks across three models. For iterate attacks I-FGSM and PGD, we set the parameter for controlling the size of perturbation α = 0.007 and iteration = 40. This evaluation allows us to prove the reliability and flexibility of our attacks in varying model architectures.



(a) Random targets with PGD (b) Specific targets with PGD

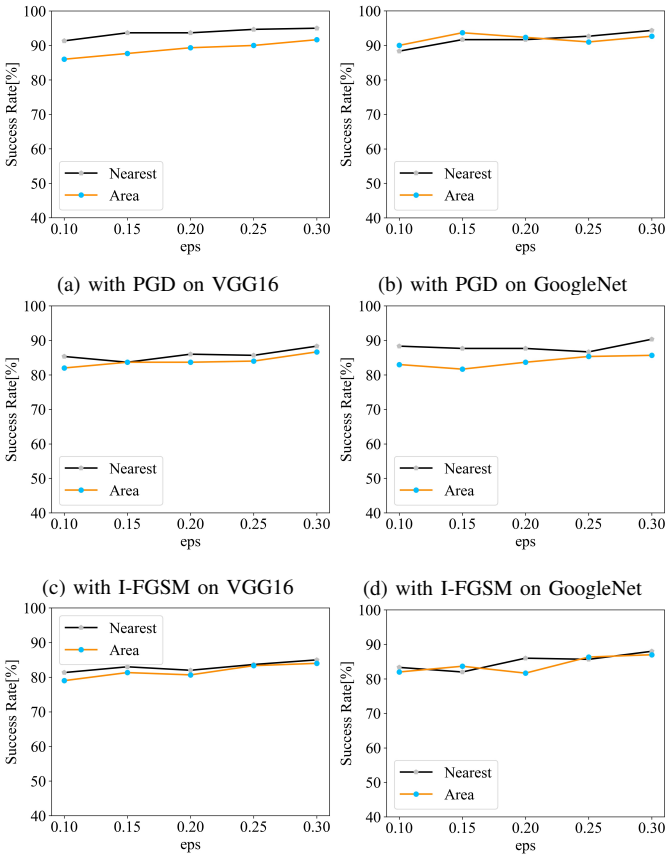
(c) Random targets with I-FGSM (d) Specific targets with I-FGSM

(e) Random targets with FGSM (f) Specific targets with FGSM

Fig. 7. Success rates of *MTISA* on InceptionV3, EfficientNet b1 and Resnet50 in different eps.

The parameter epsilon (eps) is to decide the maximum distortion of the adversarial example compared to the original input image. In Figure 6, we evaluate the success rate of the attack with the random and specific label targets when given different eps. When eps increases, the success rate of Area scaling attacks with random label targets decreases significantly. Furthermore, the attacks assisted with PGD and I-FGSM with random label targets can achieve an accuracy of 100% when using the Nearest algorithm. For the goal of misleading the image classifiers to the specific targets, this attack strategy also achieves a high attack success rate.

2) Transferability of the attack: We measure the success rates of the attacks using only a single model for training purposes. Our proposed attack demonstrates its effectiveness even in scenarios where the adversary possesses no prior knowledge about the target model. To assess the transferability of our attack, we generate attack images using a single model for training and evaluate their success rates across multiple models which share the same input size as the single model.



(a) with PGD on VGG16 (b) with PGD on GoogleNet

(c) with I-FGSM on VGG16 (d) with I-FGSM on GoogleNet

(e) with FGSM on VGG16 (f) with FGSM on GoogleNet

Fig. 8. Success rates of *MTISA* which trained with resnet50 are implemented on VGG16 and GoogleNet.

For enhancing the transferability of the attacks, we employ distribution relevant attack method [17] to improve the adversarial attacks. Figure 8 shows the result of the transfer attacks with random label targets on other models. When the attack is only trained with ResNet50, we can observe that the attack remains attack performance against models like VGG16 and GoogleNet.

V. Conclusion

In this paper, we present a multi-target image-scaling attack which can fool both human vision and machine vision systems. The adversary can implement our attack according to multiple target objects by resizing the image to different sizes. Specifically, We employ SinGAN to achieve the multi-target image-scaling attack and employ adversarial attacks to further fool image classifiers. Extensive experiments demonstrate that our attack is effective across various models and input sizes. It also has good transferability and remains effective for unseen victim models.

Acknowledgment

This work is supported by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B0101360001, National Natural Science Foundation of China under Grants 62020106013, 61972454, and 61802051, Sichuan Science and Technology Program under Grants 2020JDTD0007 and 2020YFG0298, the Fundamental Research Funds for Chinese Central Universities under Grant ZYGX2020ZB027.

References

[1] Q. Xiao, Y. Chen, C. Shen, Y. Chen, and K. Li, "Seeing is not believing: Camouflage attacks on image scaling algorithms," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 443–460.

[2] E. Quiring and K. Rieck, "Backdooring and poisoning neural networks with image-scaling attacks," in *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2020, pp. 41–47.

[3] B. Kim, A. Abuadbba, Y. Gao, Y. Zheng, M. E. Ahmed, S. Nepal, and H. Kim, "Decamouflage: A framework to detect image-scaling attacks on cnn," in *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 2021, pp. 63–74.

[4] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4570–4580.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[8] W. Jiang, H. Li, G. Xu, and T. Zhang, "Color backdoor: A robust poisoning attack in color space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8133–8142.

[9] W. Jiang, T. Zhang, H. Qiu, H. Li, and G. Xu, "Incremental learning, incremental backdoor threats," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[10] W. Jiang, H. Li, G. Xu, T. Zhang, and R. Lu, "Physical black-box adversarial attacks through transformations," *IEEE Transactions on Big Data*, 2022.

[11] 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.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[13] 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.

[14] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[16] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[17] Y. Zhu, Y. Chen, X. Li, K. Chen, Y. He, X. Tian, B. Zheng, Y. Chen, and Q. Huang, "Toward understanding and boosting adversarial transferability from a distribution perspective," *IEEE Transactions on Image Processing*, vol. 31, pp. 6487–6501, 2022.