IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 May 2024, accepted 4 June 2024, date of publication 17 June 2024, date of current version 26 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3415356

RESEARCH ARTICLE

Adversarial Examples for Image Cropping: Gradient-Based and Bayesian-Optimized Approaches for Effective Adversarial Attack

MASATOMO YOSHIDA[1], (Student Member, IEEE), HARUTO NAMURA[1], AND MASAHIRO OKUDA[2], (Senior Member, IEEE)

[1]Graduate School of Science and Engineering, Doshisha University, Kyoto 610-0394, Japan
[2]Faculty of Science and Engineering, Doshisha University, Kyoto 610-0394, Japan

Corresponding author: Masatomo Yoshida (yoshida@vig.doshisha.ac.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) Support for Pioneering Research Initiated by the Next Generation (SPRING) under Grant JPMJSP2129; in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 23K11174; and in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Promotion of Distinctive Joint Research Center Program under Grant JPMXP 0621467946.

ABSTRACT In this study, we propose novel approaches for generating adversarial examples targeting machine learning-based image cropping systems. Image cropping is crucial for meeting display space restrictions and highlighting content's interest areas. However, existing image cropping systems often miss user-intended areas, have necessities to remove inherent biases in light of AI fairness, or might expose users to legal risks. To address these issues, our paper introduces approaches for effectively creating adversarial examples in both black-box and white-box settings. In the white-box approach, we utilize gradient-based perturbations focusing on the model's blurring layer and targeting effective areas. For the black-box approach, even for models where gradient information is unavailable, we levered pixel attacks with Bayesian optimization and patch attacks to effectively narrow the search space. We also introduce a novel quantitative evaluation method for image cropping by measuring shifts in gaze saliency map peak values, reflecting a typical scenario with social network services. Our results suggest that our approaches not only outperform existing methods but also exhibit the potential to be an effective solution to the problems even with models on actual platforms.

INDEX TERMS Adversarial examples, image cropping, object detection, saliency map, Twitter.

I. INTRODUCTION

Image cropping plays an important role in maximizing limited display space and emphasizing specific regions of interest within an image. Machine Learning (ML) models have become a staple in automating this process to efficiently highlight the most engaging parts of an image. For instance, Twitter (currently X),[1] for example, has announced that when displaying user posts, it uses ML models to automatically generate thumbnails for user posts by cropping images to fit appropriately within display areas. Similarly, Netflix employs an ML model to select and crop meaningful scenes from movies for thumbnail creation.[2]

However, these ML models often contain biases that need to be addressed and expose users to legal risks. The practice of automatically generating thumbnails through image cropping can sometimes distort the original meaning or identity of the image. In Japan, controversy arose following a Supreme Court decision concerning Twitter's (currently X) cropping function, which led to issues with image attribution and potentially infringed upon photographers' rights [1]. This

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

[1] https://twitter.com/

[2] https://netflixtechblog.com/a442f163af6

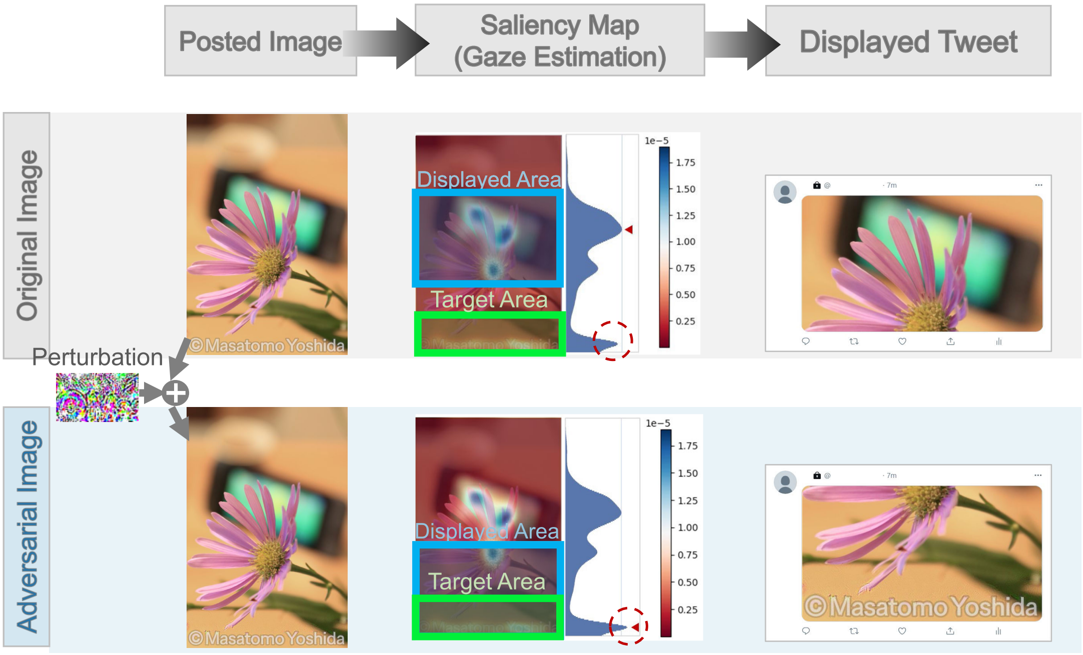


FIGURE 1. **Overview of the proposed method.[3] In the original image (upper), the copyright notice on the bottom is not displayed. However, the proposed method (lower) adds perturbations that lead to cropping functions to show the desirable area.**

controversy highlights the challenges of balancing automatic cropping with maintaining image integrity and avoiding legal risks. Thus, ML models need to be refined to ensure they respect user intentions and eliminate biases as much as possible while cropping images.

On the other hand, the research of adversarial examples for machine learning (ML) models has been actively conducted [2], [3], [4], [5], [6]. This includes not only white-box attacks, which have full access to internal model information, including gradients or model architecture but also black-box attacks, which operate without access to the target model's internal information.

In this paper, we propose a novel approach to generating adversarial examples to shift the cropping area of images (Figure 1). Unlike previous methods [2], [3], [4], [5], [7] focusing on adversarial attacks against classifiers or detectors, our method aims to perturb input images, leading the cropping model to change the intended cropping regions. First, as a white-box attack, we propose an attack that shifts the cropped area by generating gradient-based adversarial examples targeting a model that predicts gaze saliency maps. Next, considering broader model applicability, our proposed method generates adversarial examples as a black-box attack, utilizing optimization methods that reduce the search space. These two approaches introduce a fundamental approach to attacking ML models for image cropping, effectively executing attacks against the target image cropping models.

The contributions of this work are as follows:

1) **Adversarial Example for Image Cropping**: We generated adversarial examples for Deep Neural Network models for image cropping, a field with very little prior research, considering models used in an actual social networking service. In the black-box approach, we propose an iterative perturbation generation algorithm that advances FGSM. We propose an optimized method to create effective adversarial examples in the white-box approach.
2) **Reducing Attack Area / Search Space**: When attacking image cropping models, it's crucial to narrow down the necessary areas and search space for generating effective adversarial examples. In the white-box approach, we narrowed the scope of perturbation using Grad-CAM, and in the black-box approach, we reduced the search space using Bayesian optimization methods.
3) **Development of a Quantitative Evaluation Metric for Image Cropping**: Previously, a formalized evaluation framework for image cropping has been absent. In this work, we propose a new quantitative metric that assesses the extent of shift in the cropped image area, thereby providing a means to gauge the performance of our cropping technique.

II. RELATED WORKS

This paper proposes techniques that utilize adversarial examples to target ML models. Most of these models predict gaze saliency maps to adjust the cropped area. This section overviews existing research on adversarial examples in computer vision and their real-world applications. We focus

[3] The image is for illustration purposes only (not included in the dataset).

on white-box attacks in Section II-A, discuss saliency maps in Section II-B, and finally address adversarial examples in black-box attacks in Section II-C.

A. ADVERSARIAL EXAMPLES OVERVIEW AND ITS APPLICATION

Adversarial examples are inputs that introduce minimal yet impactful noise, known as perturbations, leading machine learning models to unfavorable outcomes. While adversarial examples have also been applied in NLP tasks, this section narrows the focus to computer vision, relevant to this work. Szegedy et al. [2] first introduced adversarial examples by causing state-of-the-art DNN models to misclassify through small perturbations. Goodfellow et al. [3] proposed a fundamental technique, the Fast Gradient Sign Method (FGSM), based on the gradient of the target models. Szegedy's method was effective against models with complete access to internal information, such as architecture and gradients, characterizing it as a white-box attack. Studies on adversarial training, which incorporate adversarial examples back into the target model's training data to enhance the model's robustness and resistance to adversarial attacks, are also prevalent.

Applications of adversarial examples in the real world constitute an important research area [8]. Early research on adversarial examples predominantly focused on attacks against image classification and defense strategies. Typical applications include inaudible voice commands [4], studies leading to misrecognition of traffic signs [7], and adversarial examples for facial recognition [3], [9], [10]. Additionally, Ghorbani et al. [11] introduced adversarial examples that manipulate maps representing the interpretability of ML models. Our research is closely related to this study, considering that gaze saliency maps, while not identical, share similar characteristics with interpretability maps.

Other relevant techniques include universal adversarial perturbations [12], thermometer encoding [13], defense methods using generative models [14], Jacobian regularization for robust learning [15], manipulation of low-frequency components in 3D point clouds [16], and translation-invariant attacks [17]. These techniques may be adapted to generate effective and robust adversarial examples for image cropping systems. However, to the best of our knowledge, there is very little research on attacks on the cropping models, while many of the previous researches have focused on classification and detection, as shown above.

Section II-B clarifies the similarities and differences between the gaze saliency maps we address and typical saliency maps.

B. SALIENCY MAP

Many methods including ours utilize the term "saliency map" to generate adversarial attacks, which mainly has two interpretations in computer vision. We classified these two interpretations in Table 1. One is the extent of gaze concentration measured by eye-tracking or predicted (calculated) gaze concentration. Many models including Itti et al. [18] and Ardizzone et al. [19] are used in image processing, segmentation, and object detection. Twitter, mentioned above, published in their blog that they use DNN models for gaze prediction to crop images uploaded by users when displaying posts[4], as based on the model called DeepGaze II by Kümmerer et al. [20].

TABLE 1. Variations and differences in interpreting "Saliency Maps" across studies.

Term	Meaning	Citation Instance
Saliency Map (Computer Vision)	extent of gaze concentration	[18], [19]
Saliency Map (Machine Learning)	map of feature importance	[11]

Algorithm 1 Perturbations Created by the Proposed Method (White-box approach)

```
Require: Original Image $\boldsymbol{x}$, Target area (to be displayed) $\boldsymbol{y}$,
    parameter to adjust perturbation size $\alpha$, # of iterations $N$
Ensure: Adversarial Image $\boldsymbol{x}'$
  $\boldsymbol{x}' = \boldsymbol{x}$
  for $k \leftarrow 1$ to $N$ do
      $\boldsymbol{x_p} = \boldsymbol{x}'$
      $\boldsymbol{\eta}' = \nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}', \boldsymbol{y})$   // Calc. perturbation
      $\boldsymbol{\eta} = \alpha \cdot \boldsymbol{\eta}' / ||\boldsymbol{\eta}||_2$   // Adjust size of perturbation
      $\boldsymbol{x}' = \boldsymbol{x_p} + \boldsymbol{\eta}$   // Add perturbation
  end for
  return $\boldsymbol{x}'$
```

Another role of the saliency map is to express feature importance by the map representing which feature affects the model's output. Ghorbani et al. [11] referred to the term as this meaning. Most previous studies other than Deep Gaze II made the saliency map (of feature importance) from the value of the hidden layer (not the output layer).

C. BLACK-BOX APPROACH FOR ADVERSARIAL ATTACKS

In contrast to the white-box adversarial attacks discussed in Section II-A, black-box attacks are executed without access to the model's internals, relying solely on the model's output to operate the attack [21], [22], [23], [24]. In practical scenarios, gaining access to a target model's internal information, such as gradient and architecture, identical to the original, is rare. Utilizing black-box attacks, such as those employing surrogate models or score-based approaches, leads to the generation of more realistic adversarial examples.

Below are three primary black-box approaches. (1) **Transfer attacks** involve creating a surrogate model instead of directly accessing the target model and then applying adversarial examples generated for the surrogate to the target

[4] https://blog.twitter.com/engineering/en_us/topics/infrastructure/2018/Smart-Auto-Cropping-of-Images.html

model [22], [25]. This method allows query access to the target model and, in some instances, access to the same dataset as used by the target model. The community has researched various methods to enhance transferability [26], [27]. (2) **Score-based attacks** do not require the creation of a surrogate model for the dataset and generate adversarial examples using the confidence scores for each class output by the target model. This category includes attacks that solve optimization problems independent of gradients, such as Zeroth Order Optimization (ZOO) [5], utilizing changes in output scores. Various optimization methods including Bayesian optimization [28] and evolutionary computation [29] are used in the optimization process. (3) **Decision-based attacks**, classified as attacks that only utilize the output labels without needing the output score vector [23], operate under even more stringent constraints compared to score-based attacks. These attacks generate adversarial examples by probing along the model's decision boundary and making minor adjustments to the input around the boundary. This category includes methods like the Boundary attack [30].

In this study, our black-box approach employs query-based adversarial attacks. Our black-box approach is classified under **score-based attacks**.

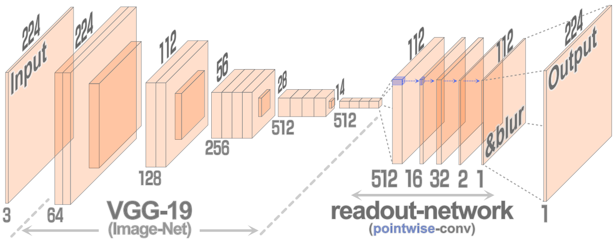


FIGURE 2. **Model architecture used in the white-box approach.**

III. PROPOSED METHOD

We propose white-box and black-box approaches. In Section III-A, we describe the white-box approach that uses gradients of the target model. In contrast, in Section III-B we detail the black-box approach which doesn't use gradient information, and alternatively, we use Bayesian optimization. Section III-C introduces our evaluation metric for image cropping.

A. WHITE-BOX APPROACH

Our white-box approach utilizes the model employed by Twitter [31], a typical neural network model designed for image cropping. This approach generates perturbations based on the model's gradients, taking an original image and a bounding box, which indicates the area an attacker wishes to display instead of the intended area. The method involves iterative calculation of perturbations using the model's gradient. This proposed method enhances traditional approaches designed for image classification tasks, increasing the effectiveness through repeated perturbations.

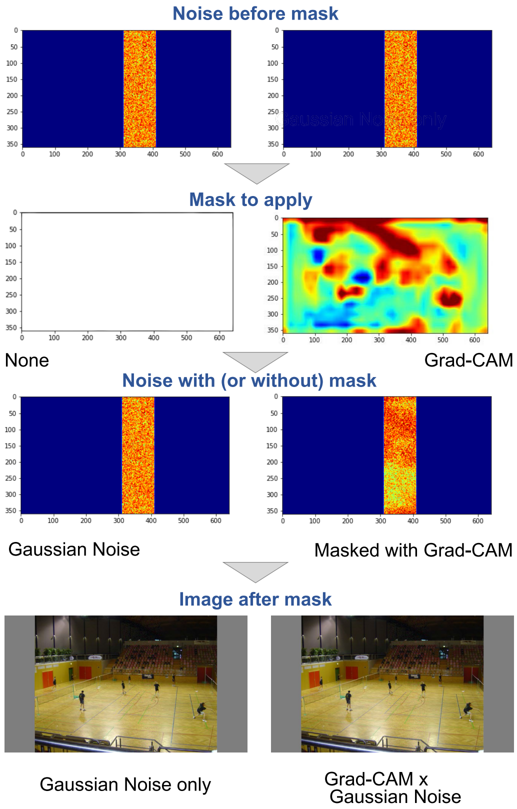


FIGURE 3. **Effect of the mask with Grad-CAM. Masking with Grad-CAM is used in the white-box approach.**

In Computer Vision, adversarial examples for DNNs are commonly generated for classifier models using datasets like ImageNet [32]. However, our study focuses on generating adversarial examples for DNNs that produce gaze saliency maps. We will explore the similarities and differences between adversarial examples for classifiers and those for gaze saliency maps predicting gaze concentration. As mentioned in related studies, while there is extensive research on adversarial examples for image classification, generating adversarial examples for the saliency maps derived from these hidden layers marks a significant departure from related work and bears unique challenges that need to be addressed, especially in light of the controversial issues specific to this task [1].

Our white-box approach combines methods that control the model's interpretation with a gaze prediction model based on the model's feature importance map. The goal of this approach is to introduce perturbations into an image to change the cropping area determined by the target model. This method calculates perturbations based on the Fast Gradient Method (FGM), a generalized form of the Fast Gradient Sign Method (FGSM) without using the sign function. The FGSM-based adversarial attack is known for its subtlety in perturbations, as represented by the following equation:

$$\boldsymbol{\eta} = \epsilon \cdot sign(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y})), \tag{1}$$

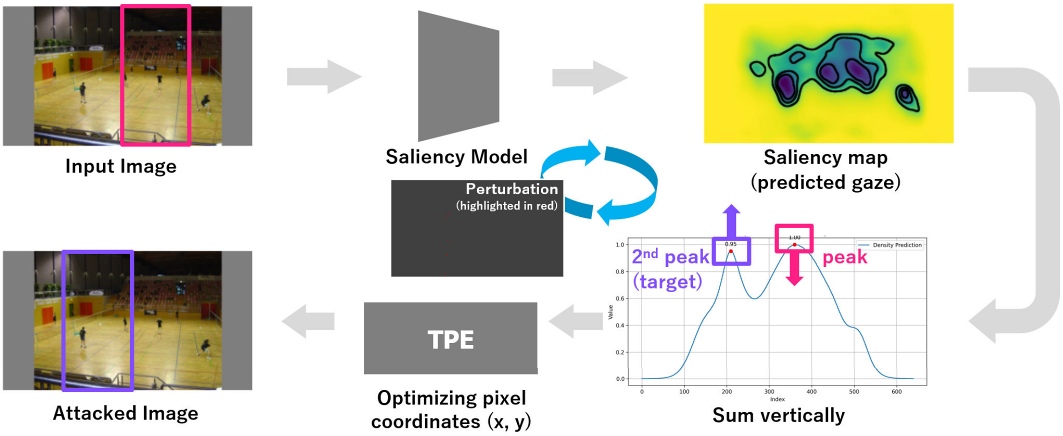


FIGURE 4. Procedure of the black-box approach.

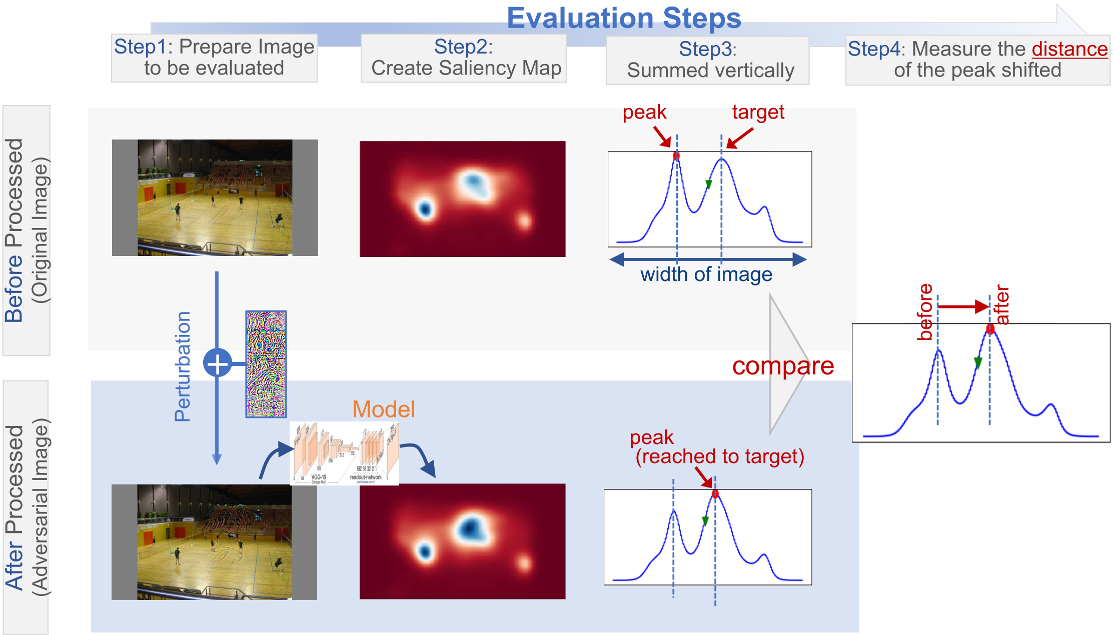


FIGURE 5. Procedure of the evaluation (Mainly used in white-box approach). In step 1, we prepare the image to be evaluated. Next, based on the image we create gaze saliency map. In step 3, we sum vertically the value of saliency map created in step 2. Finally, we measure the distance of the peak shifted.

where $sign(\cdot)$ is the sign function, $\boldsymbol{\theta}$ represents the model parameters, $\boldsymbol{x}$ the input image, ϵ is a parameter to adjust the size of the perturbation, and $J(\cdot)$ represents the loss function. In this study, $\boldsymbol{y}$ represents the target area of the input image, that is, the area that the user intends to display. However, as our targeted model involves blur operations, the effectiveness of FGSM is reduced [33], leading us to employ FGM, represented by the following equation, which removes the sign function:

$$\boldsymbol{\eta} = \epsilon \cdot \nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}), \tag{2}$$

This modification allows for a more nuanced manipulation of the image, potentially overcoming the limitations imposed by blur operations in the target model.

Furthermore, we use the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [34] as a comparative method to generate transferable adversarial examples. MI-FGSM integrates the concept of momentum into the iterative process to stabilize update directions and escape from poor local maxima. The iterative process of MI-FGSM can be represented by the following equation:

$$\begin{aligned} \boldsymbol{g}^{t+1} &= \mu \cdot \boldsymbol{g}^{t} + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}^{t}, \boldsymbol{y})}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}^{t}, \boldsymbol{y})\|}, \\ \boldsymbol{x}^{t+1} &= \boldsymbol{x}^{t} + \epsilon \cdot sign(\boldsymbol{g}^{t+1}), \end{aligned} \tag{3}$$

where μ is the decay parameter, and $\boldsymbol{g}^{t}$ is the accumulated gradient at iteration t.

However, considering that our proposed method removes the sign function to allow for better peak shift, we also introduce a variant of MI-FGSM called Momentum Iterative Fast Gradient Method (MI-FGM) to ensure a more equitable comparison. The iterative process of MI-FGM can be formulated as:

$$\begin{aligned} \boldsymbol{g}^{t+1} &= \mu \cdot \boldsymbol{g}^{t} + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}^{t}, \boldsymbol{y})}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}^{t}, \boldsymbol{y})\|}, \\ \boldsymbol{x}^{t+1} &= \boldsymbol{x}^{t} + \epsilon \cdot \boldsymbol{g}^{t+1}, \end{aligned} \tag{4}$$

By utilizing the accumulated gradients (momentum), both MI-FGSM and MI-FGM can stabilize the update directions and escape from suboptimal local maxima more easily. This allows for the generation of adversarial examples that are potentially effective against other models. In this study, we adopt MI-FGSM as one of the comparative methods to evaluate the performance of our proposed methods and verify its effectiveness.

In our white-box approach, we compare several methods to demonstrate the effectiveness of our proposed approach. Our method applies the operation in Equation 2 several times with a smaller perturbation size. In order to generate more effective adversarial examples under the constraints of perturbation size, we employ a technique that applies perturbations only to significant areas. Specifically, we apply the output of Grad-CAM as a mask to the output of FGM. Figure 3 shows the procedure of applying Grad-CAM as a mask. As an example, we prepared images with Gaussian noise as a comparative method. As shown in the upper row, noise for the target region is prepared. Then, by applying the output of Grad-CAM as a mask, it is possible to realize the shaded noise as shown in the third row. The image with the mask applied is shown in the lower row. Detailed settings are described in Section IV-B. In our approach, perturbations are generated in three iterations, and ϵ is calculated to match the perturbation size specified in the experimental settings. The process of generating perturbations is shown in Algorithm 1.

One of the challenges faced by our approach, as indicated in Equation 2, is the need for gradient information from the target model. However, due to the inclusion of a random search in the target model for selecting the layer of the feature map, obtaining weights in the desired form was not possible. Therefore, we created a surrogate model by fixing the output of the hidden layer preceding the readout network (as illustrated in Figure 2) and retrained the model.

The method proposed in this paper differs from existing studies in two main aspects:

1) Our method generates adversarial examples for models that predict gaze by shifting the output of the final layer (output layer), whereas many previous studies focusing on saliency maps typically derive them from hidden layers (not the output layer).
2) The model used in our method includes a blur layer before the output layer. Brama and Grinshpoun [33] demonstrated that the performance of adversarial examples diminishes if the model contains strong blur effects prior to the final layer. As most previous studies use too small perturbations, their effectiveness is reduced when the model incorporates blurring effects.

As outlined above, our white-box approach introduces a new method for generating adversarial examples that shift the cropped area of images predicted by DNNs for gaze saliency maps. This task has its unique challenges and difficulties, contributing to contributions that differ from the previous research on adversarial examples for DNNs in Computer Vision.

B. BLACK-BOX APPROACH

We also generated adversarial examples for black-box models. To extend our method for a broader range of real-world models and make it applicable in various scenarios, it is beneficial to construct a black-box approach that does not rely on gradient information. In this experiment, we employed Bayesian optimization techniques to effectively target smaller regions.

The Tree-structured Parzen Estimator (TPE), introduced by Bergstra et al. [35], is a Bayesian optimization method known for its efficiency in optimizing black-box functions, often used in hyperparameter optimization within machine learning contexts.

We assume x represents the hyperparameter values and y the loss, with $y*$ being a threshold determined by a constant γ. First, the threshold $y*$ bifurcates the probability density function into two distinct sections: $P(x|y > y)$ for less favorable outcomes and $P(x|y \leq y)$ for more favorable outcomes. This division is then utilized to calculate the Expected Improvement (EI) metric, which assesses the potential of hyperparameter values. The goal is to find the hyperparameter x' that minimizes the loss function, by maximizing the following equation:

$$EI \propto \frac{P(x|y > y*)}{P(x|y \leq y*)} \tag{5}$$

TPE is especially effective in scenarios with limited evaluations, often outperforming methods based on evolutionary computation.

The workflow of our black-box approach is summarized in Figure 4. Initially, an input image is prepared and a gaze saliency map is created, similar to the white-box approach. Then, using the Optuna library, TPE is utilized to select pixel coordinates that alter the saliency based on the peak height, as illustrated in Figure 6. Unlike the gradient-based perturbations applied in the white-box approach, this method involves selecting individual pixel coordinates and placing white pixels. While the white-box approach allows for the calculation of multiple perturbation pixels in bulk, the black-box approach is challenging due to its computational demands. However, TPE reduces computation load, suggesting that our method can alter gaze peaks with minimal image modification. The loss function in this approach is based on the difference in peak heights.

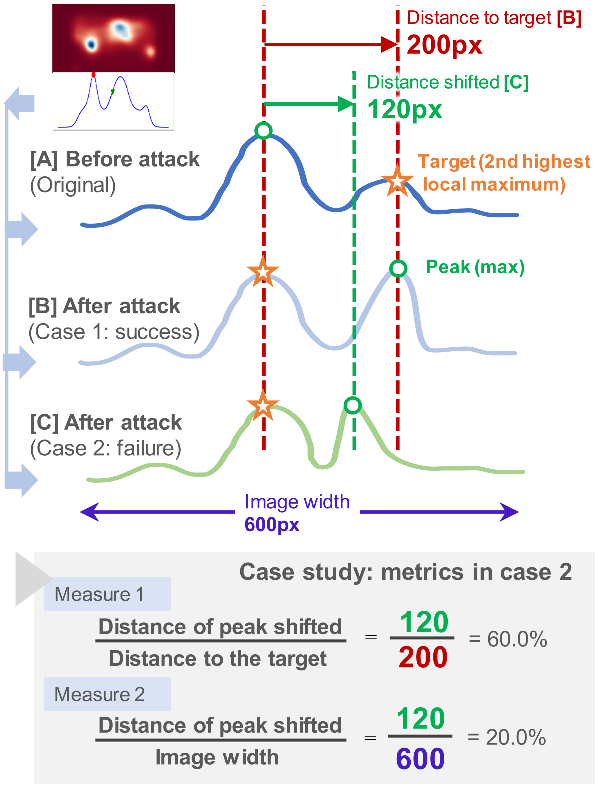

Distance to target [B]
200px
Distance shifted [C]
120px
[A] Before attack
(Original)
Target (2nd highest local maximum)
Peak (max)
[B] After attack
(Case 1: success)
[C] After attack
(Case 2: failure)
Image width
600px
Case study: metrics in case 2
Measure 1
Distance of peak shifted / Distance to the target = 120 / 200 = 60.0%
Measure 2
Distance of peak shifted / Image width = 120 / 600 = 20.0%

FIGURE 6. Illustration of two distance metrics, comparing successful and unsuccessful cases.

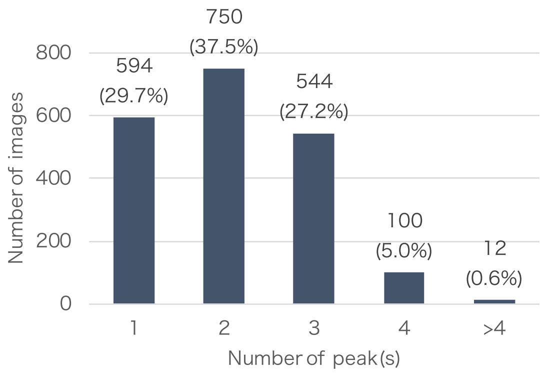


FIGURE 7. Distribution of the number of peaks in the CAT2000 dataset.

C. EVALUATION METHOD

To the best of our knowledge, the research field focusing on shifting the cropped area of an image is relatively unexplored, and thus, a standard evaluation method does not yet exist for this task. Herein, we introduce a novel method to quantitatively assess the extent to which the cropped area of an image has been shifted, utilizing the saliency map generated from gaze prediction, which contains critical information for guiding image cropping.

We evaluate the effectiveness of our algorithm objectively by analyzing the change in the saliency map before and after the perturbation. Pixels of the saliency map are added vertically to form a one-dimensional sequence, and we refer to its maximum value as a "peak." In each image, our objective is to shift the focus to the second highest local maximum of the saliency map. The success of our perturbation is quantified by the movement of this peak within the saliency map.

TABLE 2. Averaged Measure 1 $M_{\text{target}}(i)$ for proposed and baseline methods.

Perturbation size (L2 norm)	10	20	30
🟩 Gaussian Noise only	4.2%	9.6%	12.5%
🟩 Grad-CAM x Gaussian Noise	5.5%	9.5%	11.2%
⬜ Grad-CAM x Gradient (MI-FGM)	42.1%	60.7%	71.1%
🟦 Grad-CAM x Gradient (FGM)	40.9%	61.9%	69.2%
🟦 **Proposed Method**	**50.0%**	**65.8%**	**74.4%**

The evaluation process, as illustrated in Figure 5, begins with generating a gaze saliency map for the images both before and after perturbation application. We quantify how the peak of the saliency map aligns with our targeted area, thus assessing the method's impact based on the peak's displacement.

In our study, we evaluate the effects of adversarial attacks on the saliency maps by utilizing two specific metrics for each image labeled as i. Initially, we introduce a measure to gauge the effectiveness of the perturbation. This measure, referred to as Measure 1 $M_{\text{target}}(i)$, is calculated based on how much the peak of the saliency map has moved. A greater Measure 1 value indicates a perturbation that successfully moves the peak closer to our predetermined target, signifying a significant change in the saliency area. Following this, we present another metric, Measure 2 $M_{\text{width}}(i)$, which standardizes the distance moved by the saliency peak across images of different widths, allowing for a consistent comparison across various image sizes.

$$\text{Measure 1:} \quad M_{\text{target}}(i) = \frac{D_{\text{shifted},i}}{D_{\text{target},i}}, \tag{6}$$

$$\text{Measure 2:} \quad M_{\text{width}}(i) = \frac{D_{\text{shifted},i}}{W_{\text{image},i}}, \tag{7}$$

where $D_{\text{shifted},i}$ represents the distance the saliency peak has moved for image i, $D_{\text{target},i}$ is the distance from the original peak position to the target for image i, and $W_{\text{image},i}$ denotes the width of image i.

IV. EXPERIMENT

A. SETTING AND DATASET

We used the images from the CAT2000 dataset [36], which is known for its diverse range of images and original saliency maps created from actual human gaze data. This dataset is ideal for our experiments as it includes images with varying numbers of local maxima in saliency, which is crucial for testing the effectiveness of our proposed method in shifting predicted gaze concentration peaks. To focus our study, we selected images with 2 to 4 local maxima, excluding those

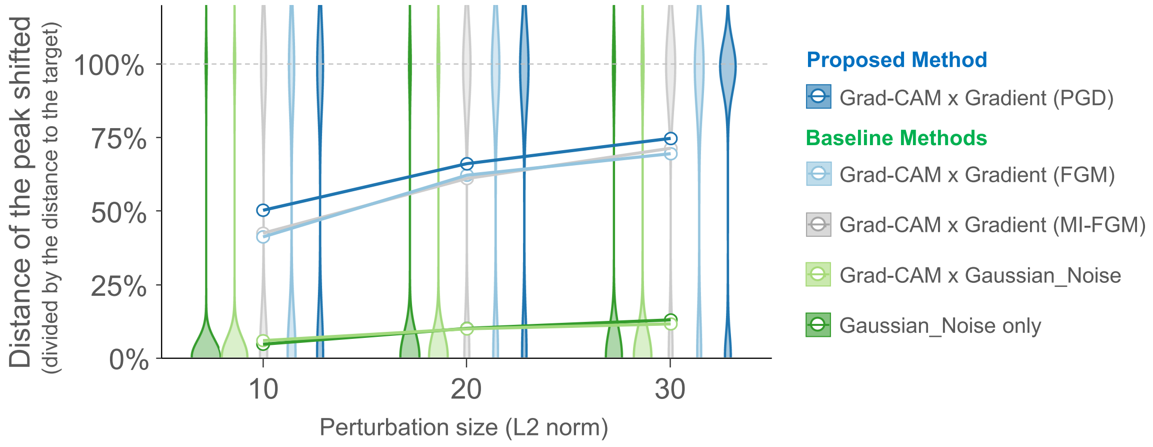


FIGURE 8. Distribution of the distance of the peak shifted.

TABLE 3. Averaged results of Measure 2 $M_{\text{width}}(i)$ for proposed and baseline methods.

Perturbation size (L2 norm)	10	20	30
🟩 Gaussian Noise only	1.2%	2.7%	3.6%
🟩 Grad-CAM x Gaussian Noise	1.6%	2.7%	3.2%
⬜ Grad-CAM x Gradient (MI-FGM)	10.9%	15.7%	18.6%
🟦 Grad-CAM x Gradient (FGM)	10.9%	16.2%	18.6%
🟦 **Proposed Method**	**12.9%**	**17.2%**	**19.0%**

TABLE 4. Aggregated minimum size of perturbations needed to shift the peak to the target in each method.

	Perturbation size (L2)	Baseline 10	20	30	N/S	Total
Proposed	10	18	18	14	134	184
	20	4	2	1	55	62
	30		1	1	27	29
	N/S				60	60
	Total	22	21	16	276	335

*N/S = Not Shifted to the target.
**When calculating, minimum size in proposed/baseline methods was used.

with a single or more than four peaks, to ensure suitability for the task measurement (Figure 7).

Other major restrictions applied to exclude images that are inappropriate for the experiments are as follows (if an image meets at least one of these restrictions, it is excluded).

1) Vertical photos[5]: photos in which the filled pixels are more than 25% (which corresponds to the aspect ratio of 4:3). 0% corresponds to 16:9. We should note that this method is applied only to horizontal images in the experiments, but it can be applied to images of any aspect ratio.
2) distance of the x-axis between the peak and the target is more than 15%, compared to the width of the images.
3) distance of the y-axis (height) between the summed value of the peak and that of the target is more than 40%, compared to the summed value of the peak.

After all, we used 335 images in this experiment and resized them to 640 pixels in width.

B. EXPERIMENT 1: WHITE-BOX APPROACH

In this experiment, we assessed the impact of applying perturbations guided by the Grad-CAM output, a technique that highlights regions of interest, thereby optimizing the perturbation process. Grad-CAM, proposed by Selvaraju et al. [37], helps in identifying significant areas for applying perturbations, reducing the likelihood of applying on ineffective areas.

In our white-box experiments, we used a model based on the VGG19 architecture [38] pre-trained on the SALICON dataset [39] and fine-tuned on the MIT1003 dataset [40]. The model's hidden layer output preceding the readout network was fixed during the fine-tuning process. For generating perturbations, we employed the Fast Gradient Method (FGM), which removes the sign function from the Fast Gradient Sign Method (FGSM) to enhance the effectiveness to the image cropping model. The perturbations were applied iteratively with a smaller size (ϵ) in each iteration, calculated to match the total perturbation size specified in the experimental settings (10, 20, 30 in L_2 norm). Because the gradient is re-calculated in every step in our method, ϵ is 1.77 times smaller than that for FGM. We used the Momentum Iterative Fast Gradient Method (MI-FGM), a variant of MI-FGSM [34] without the sign function, as a comparative method. In both the proposed method and MI-FGM, the decay factor μ is set to 1.0, and the number of iterations N is set to 3.

The Grad-CAM output was normalized to a 0-1 scale and used as a mask to focus our method's perturbations on

[5] Note the difference with Figure 1 (vertical photo). As 67% of the dataset are horizontal photos, we used horizontal photos in the experiment.

areas with higher importance scores. This approach aims to enhance the effectiveness of perturbations by concentrating on regions more likely to influence the model's gaze prediction. The effectiveness of each method was evaluated by measuring the saliency peak's movement using the DeepGaze II model [20], which closely mirrors Twitter's image cropping system.

We compared our proposed method with four baseline methods: (1) Gaussian noise only, (2) Gaussian noise combined with Grad-CAM, (3) the gradient-based method using FGM for saliency transfer, and (4) the gradient-based method using MI-FGM. For Gaussian noise applications, we targeted a specific area around the intended peak shift, with a width set at 100 pixels (15.6% of the image width) and height set at matching the image's one.

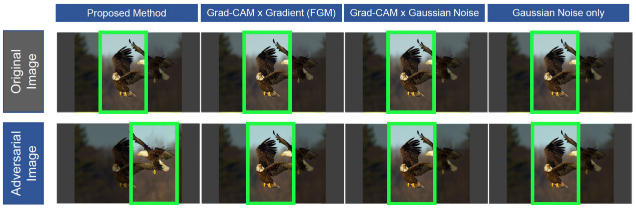


FIGURE 9. Experimental Results: Area surrounded by a square indicates the cropped area.: (upper row) original image, (bottom row) adversarial examples created by several methods.

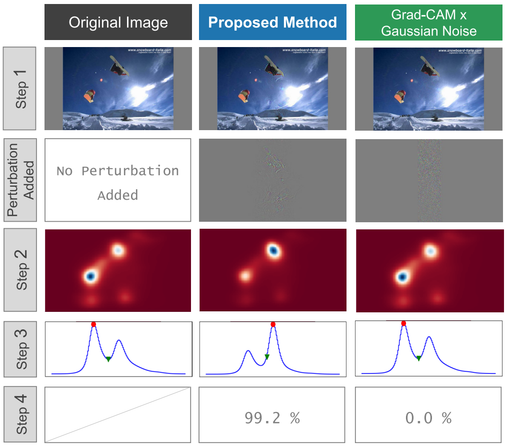


FIGURE 10. Comparison among original, proposed, and baseline methods. The step numbers (Step n) correspond to the numbers in Figure 5.

C. EXPERIMENT2: BLACK-BOX APPROACH

In Experiment 2, we extend the scope of our adversarial example generation to encompass black-box models, thereby broadening the applicability of our methodology across a diverse scenarios where direct access to model gradients is not feasible. To this end, we leveraged the capabilities of Bayesian optimization for optimizing objective functions without requiring gradient information. We utilized the same image dataset as in Experiment 1. The optimization was performed using Optuna TPE (version 3.3.0). The code was executed on a workstation equipped with an AMD EPYC 7232P (8-Core) CPU and 4 x NVIDIA RTX A4000 GPUs, and measured the metrics. We investigated 4 metrics such as Success Rate, PSNR, loss function value, and process time. We evaluated these metrics across five different numbers of attacked pixels: 5, 15, 30, 40, and 50. The loss function was set as the difference between the peak value and the value of the second maxima. The number of trials for pixel attacks was set at 100. Additionally, to select pixels more efficiently, we considered attacking multiple pixels in patches. The patches were compared in sizes of 1 × 1 and 3 × 3. Regarding the peak, as in Experiment 1, we measured the movement distance of the peak in a graph that represented the gaze saliency map in one dimension.

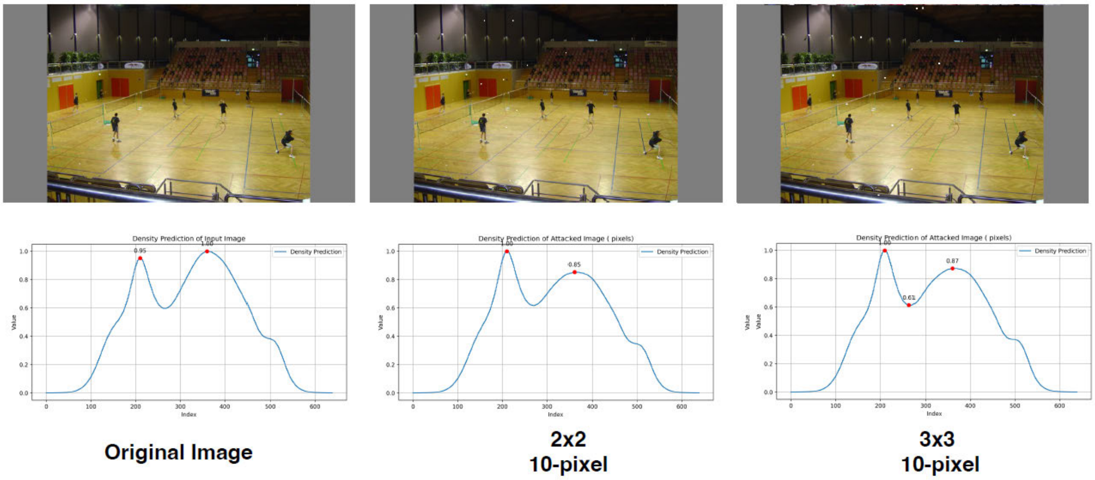
TABLE 5. Result of pixel attacks (1 × 1).

# Attacked Pixels	Success Rate	PSNR (dB)	Loss	Time (sec)
5	0.18	51.74	0.19	76.77
15	0.24	46.66	0.16	144.61
30	0.30	43.56	0.14	243.99
40	0.35	42.24	0.12	311.13
50	0.36	41.31	0.11	380.88

D. RESULTS

The results of **Experiment 1** are summarized in Tables 2 to 4 and Figures 8 to 10. Table 2 shows the Measure 1 $M_{\text{target}}(i)$, with percentages indicating the normalized distance relative to the target. The proposed method outperformed the baseline for all perturbation sizes. Gaussian Noise with Grad-CAM outperformed Gaussian Noise only in a small perturbation. In contrast, Gaussian noise only (without Grad-CAM) outperformed in a large perturbation. The gradient-based baseline method (FGM) outperformed the other three baseline methods including MI-FGM in all perturbation sizes. For reference, we also implemented the evaluation by the Measure 2 $M_{\text{width}}(i)$ replacing the divisor $D_{\text{target, i}}$ with $W_{\text{image, i}}$ (Table 3) and obtained similar results. Figure 8 displays the distribution of peak shifts, revealing a binary pattern of shifts clustering at 0% or 100%. This pattern seems to be due to the task's requirement that the cropping area only changes when one peak takes over another as the maximum value. Unlike other tasks involving adversarial attacks, where the impact often gradually increases as the size of perturbations expands, this task shows a unique behavior not seen in other contexts.

We also conducted cross-tabulation analysis between the baseline and proposed methods to examine the instances where the peak shifted towards the target. Upon closer examination of the images in which the peak shifted to the target for both the proposed and baseline methods, it was found that, among comparable instances, the average minimum size (L2) of perturbation required for the peak shift was 11.9 for the proposed method and 19.0 for the baseline method. Consequently, the minimum perturbation



Original Image

2x2
10-pixel

3x3
10-pixel

FIGURE 11. Results among different patch sizes.

TABLE 6. Result of pixel attacks in various patch sizes.

Patch Size	# Attacked Patches	Success Rate	PSNR (dB)
1x1	5	0.18	51.74
	10	0.22	48.38
2x2	5	0.27	44.79
	10	0.32	42.03
3x3	5	0.36	41.39
	10	0.39	38.50

size required by the proposed method was 62.5% smaller than that required by the baseline method on average among comparable instances (Table 4). This result implies that, when considering non-comparable instances (shown as "N/S" in the table), the proposed method likely exceeds the baseline method in terms of effectiveness, indicating that the proposed method can generate perturbations that are more challenging to detect (smaller) than those created by the baseline method under identical constraints. Experimental results are shown in Figure 9. Figure 10 shows the comparison of the proposed and baseline methods in each step. As shown in the image, it contains specific patterns rather than the baseline method, but they are still less perceivable.

The results of **Experiment 2** are presented in Figure 11 and Table 5. Table 5 details the outcomes of attacks with 1 × 1 pixel perturbations over scenarios with 5 to 50 pixels, covering success rates, PSNR, loss values, and processing time. As the pixel count increased from 5 to 30, success rates, calculated by the instances the loss function (the gap between the top peak value and the subsequent peak) dropped below zero, enhanced. Yet, with more than 30 pixels, these improvements did not extend. Conversely, the time required for processing escalated, especially for scenarios involving over 30 pixels, where increased time did not correlate with higher success rates. This suggests a diminishing return on success rate beyond 30 pixels, necessitating additional strategies for further improvement. Figure 11 represents examples from the experiment, illustrating the visual impact of 3 × 3 versus 2 × 2 patches accompanied with the original image. The visual analysis of the perturbations highlights the balance between detectability and the stealthiness of the attack. While 3 × 3 patches result in perturbations that are visually more noticeable and potentially more disruptive, the subtler 2 × 2 patches suggest a preferable shift in saliency for scenarios where a less perceptible attack is required.

V. CONCLUSION AND FUTURE WORKS

In our study, we developed novel approaches for generating adversarial examples that address three main challenges: accurately cropping user-intended areas, eliminating biases in the context of AI fairness, and reducing legal risks associated with image cropping systems on social network services. Through our approaches, we crafted adversarial examples aimed at challenging the machine learning-based image cropping systems, assuming real-world platforms.

The core aspects of our approach include leveraging the baseline method to adversarial examples to effectively apply to new ones targeting the saliency detection model, thereby aligning with addressing issues of the task. We demonstrate the efficacy of our method against models even including a blur layer, a condition previously identified as challenging. Our white-box approach highlights substantial efficiency, achieving a 62.5% reduction in the L2 norm compared to baseline methods under identical constraints. Furthermore, our black-box strategy not only further reduces computational demands but also verifies its applicability across a broader range of models.

Below are the limitations and future works of this study:

- To confirm the effectiveness of our approach, we excluded some images that are not suitable for experiments and obtained stable results. However, future research may require validation on a larger set of images to ensure the robustness of the proposed method.

- Many existing adversarial example generation methods have been proposed for models without blur layers. When applying existing methods to the model in this study, the same level of perturbation may be insufficient, necessitating the further development of approaches.
- Generating less perceptible perturbations is also a crucial aspect to be addressed in future work.
- While our proposed method applies perturbations only to the target region, it can be applied to other regions as well. Investigating the application on other areas, such as image edges (boundaries), is also included in the future research [41], [42].
- In the black-box approach, attacking a large number of pixels proved to be time-consuming without yielding significant improvements. Future research should focus on developing effective attacks that can target a large number of pixels efficiently.

ACKNOWLEDGMENT

An earlier version of this paper was presented at the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022), Singapore, May 22–27, 2022.

REFERENCES

[1] S. Umeda, "Japan: Supreme court says retweeting tweet with photo infringes right of attribution," in *Proc. Global Legal Monitor Library Congr.*, 2020, Accessed: Feb. 5, 2024. [Online]. Available: https://www.loc.gov/item/global-legal-monitor/2020-09-23/japan-supreme-court-says-retweeting-tweet-with-photo-infringes-right-of-attribution/

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Represent. ICLR Conf. Track*, 2014.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Represent. ICLR Conf. Track*, 2015.

[4] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "DolphinAttack: Inaudible voice commands," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Oct. 2017, pp. 103–117.

[5] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.* New York, NY, USA: Association for Computing Machinery, Nov. 2017, pp. 15–26.

[6] J.-H. Jacobsen, J. Behrmann, R. Zemel, and M. Bethge, "Excessive invariance causes adversarial vulnerability," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[7] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1625–1634.

[8] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. 5th Int. Conf. Learn. Represent. ICLR Workshop Track*, Toulon, France, 2017.

[9] Q. Li, Y. Hu, Y. Liu, D. Zhang, X. Jin, and Y. Chen, "Discrete point-wise attack is not enough: Generalized manifold adversarial attack for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20575–20584.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[11] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3681–3688.

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 86–94.

[13] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[14] A. Jalal, A. Ilyas, C. Daskalakis, and A. G. Dimakis, "The robust manifold defense: Adversarial training using generative models," 2017, *arXiv:1712.09196*.

[15] J. Hoffman, D. A. Roberts, and S. Yaida, "Robust learning with Jacobian regularization," 2019, *arXiv:1908.02729*.

[16] B. Liu, J. Zhang, and J. Zhu, "Boosting 3D adversarial attacks with attacking on frequency," *IEEE Access*, vol. 10, pp. 50974–50984, 2022.

[17] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4307–4316.

[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[19] E. Ardizzone, A. Bruno, and G. Mazzola, "Saliency based image cropping," in *Proc. Image Anal. Process. ICIAP*, vol. 8156, 2013, pp. 773–782.

[20] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding Low- and high-level contributions to fixation prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4799–4808.

[21] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

[22] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.* New York, NY, USA: Association for Computing Machinery, Apr. 2017, p. 506.

[23] K. Mahmood, R. Mahmood, E. Rathbun, and M. van Dijk, "Back in black: A comparative evaluation of recent state-of-the-art black-box attacks," *IEEE Access*, vol. 10, pp. 998–1019, 2022.

[24] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021.

[25] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu, "DaST: Data-free substitute training for adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 231–240.

[26] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, "Towards transferable targeted attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 638–646.

[27] J. Byun, S. Cho, M.-J. Kwon, H.-S. Kim, and C. Kim, "Improving the transferability of targeted adversarial examples through object-based diverse input," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15223–15232.

[28] B. Ru, A. Cobb, A. Blaas, and Y. Gal, "BayesOpt adversarial attack," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–16.

[29] T. Suzuki, S. Takeshita, and S. Ono, "Adversarial example generation using evolutionary multi-objective optimization," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2019, pp. 2136–2144.

[30] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," 2017, *arXiv:1712.04248*.

[31] L. Theis, I. Korshunova, A. Tejani, and F. Huszár, "Faster gaze prediction with dense networks and Fisher pruning," 2018, *arXiv:1801.05787*.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[33] H. Brama and T. Grinshpoun, "Heat and blur: An effective and fast defense against adversarial examples," 2020, *arXiv:2003.07573*.

[34] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[35] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 2546–2554.

[36] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," in *Proc. CVPR Workshop Future Datasets*, 2015.

[37] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[39] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1072–1080.

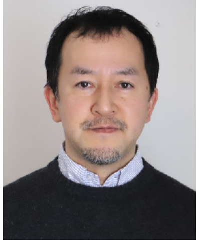
[40] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[41] M. Zajac, K. Zołna, N. Rostamzadeh, and P. O. Pinheiro, "Adversarial framing for image and video classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 10077–10078.

[42] F. Alrasheedi and X. Zhong, "Imperceptible adversarial attack on deep neural networks from image boundary," 2023, *arXiv:2308.15344*.

MASATOMO YOSHIDA (Student Member, IEEE) received the B.E. and M.E. degrees from Doshisha University, Kyoto, Japan, in 2021 and 2023, respectively, where he is currently pursuing the Ph.D. degree in engineering with the Graduate School of Science and Engineering. His research interests include analyzing spatio-temporal time series data, image processing, deep learning, and adversarial examples. He received the Support for Pioneering Research Initiated by the Next Generation (SPRING) Scholarship by Japan Science and Technology Agency (JST). He has also received an award in a local photography contest held in Kyoto, in 2015.

HARUTO NAMURA received the B.E. degree from the Faculty of Science and Engineering, Doshisha University, Kyoto, Japan, in 2022, and the M.E. degree from the Graduate School of Science and Engineering, Doshisha University, in 2024. In April 2024, he was with Sony Corporation. His research interests include image processing, including medical images and the study of adversarial examples.



MASAHIRO OKUDA (Senior Member, IEEE) received the B.E., M.E., and Dr.Eng. degrees from Keio University, Yokohama, Japan, in 1993, 1995, and 1998, respectively. From 1996 to 2000, he was a Research Fellow with Japan Society for the Promotion of Science. He was a Visiting Scholar with the University of California at Santa Barbara, Santa Barbara, CA, USA, in 1998, and Carnegie Mellon University, Pittsburgh, PA, USA, in 1999. From 2000 to 2020, he was with the Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu, Japan. He is currently a Professor with the Faculty of Science and Engineering, Doshisha University. His research interests include image restoration, high dynamic range imaging, multiple image fusion, and digital filter design. He received the SIP Distinguished Contribution Award, in 2013; the IE Award; and the Contribution Award from IEICE, in 2017.