Enhancing Remote Adversarial Patch Attacks on Face Detectors with Tiling and Scaling

Masora Okano*, Koichi Ito†, Masakatsu Nishigaki* and Tetsushi Ohki*
* Shizuoka University, Shizuoka, Japan
E-mail: okano@sec.inf.shizoka.ac.jp, {nisigaki, ohki}@inf.shizuoka.ac.jp
† Tohoku University, Miyagi, Japan
E-mail: ito@aoki.ecei.tohoku.ac.jp
‡ RIKEN AIP, Tokyo, Japan

***Abstract*—This paper discusses the attack feasibility of Remote Adversarial Patch (RAP) targeting face detectors. RAP targeting face detectors has the following difficulties compared to RAP targeting general object detectors. (1) Objects of various scales are targets for detection, and especially for small faces, the amount of convolution of features to be used as the basis for detection is small, and the range of influence on the inference results is highly restricted. (2) Also, since this is a two-class classification problem, the feature gaps between classes are large, making it difficult to attack the inference results by guiding them to another class. In this paper, we propose a new patch placement method and loss function for each problem. The patches targeting the proposed face detector showed superior detection obstruct effects compared to the patches targeting the general object detector.**

I. Introduction

Deep neural network (DNN) models are susceptible to malicious manipulation of input data. Previous researches have employed a variety of approaches, including the use of specially designed spectacles that obscure key feature points of an object by reflecting light [1], as well as the development of sophisticated Adversarial Examples (AE) that subtly alter pixel values to cause the model to misclassify inputs intentionally. The threat of DNN attacks has emerged as a significant concern in light of the increasing incorporation of DNN-based models in a multitude of systems. Additionally, researches have explored the Remote Adversarial Patch (RAP), an attack that does not depend on the specific object being targeted. Unlike methods that directly alter features, RAP can remotely influence the model's output.

RAP was defined by Mirsky et al[2]. They identified two key characterisitics in their definition of RAP. One characteristic is the semantic alteration of model inference caused by patches, and the other is the ability of patches to be remotely attacked. Among these characteristics, the remote attackability of RAP is particularly significant. Remote attackability refers to the ability of an Adversarial Patch (AP) to be exploited even when placed at a distance from the target object. In comparison to APs placed on the target object, RAP are more difficult to attack since they do not directly modify the features of the target object. For the sake of simplicity, this paper defines RAP solely based on the characteristic of remote attackability.

This study focuses on the application of RAP to face detection. Although RAP has been extensively discussed for general object detection (multi-class object detection)[3], [4], the feasibility of RAP for face detection remains underexplored. Focusing on face detection is particularly important because obstructing it significantly enhances privacy protection. For example, applying a RAP to images before publication or capture could prevent unauthorized third parties from replicating the face region.

Two key challenges in applying RAP to face detection are as follows. **(1) the scale of detectable objects is diverse.** In the field of face detection, unlike the general object detection field, there is a requirement to detect obscure and distant objects, such as surveillance camera images. As a result, there is a tendency to focus on small features so that even extremely small faces can be detected. Small features have limited characteristics in the area around the face in the image that are included in the convolution. Consequently, the area that affects face detection is also small. Thus, the restrictions on where the patches can be placed are strict and are considered difficult to attack by RAP. **(2) fewer classes to classify.** Face detection involves fewer classes compared to multi-class object detection. In addition, because RAP operates from a distance and does not alter the object's intrinsic features directly. Therefore, it is difficult to redirect a class given to a particular object to another class with similar features, which limits the approaches that can be taken in an attack.

To address this issue, this study proposes a novel RAP method which involves two main processes: scaling and tiling. The scaling process adjusts the size of patches to correspond with different face scales during training, enhancing optimization across varying scales. The **tiling process** arranges the patches in a grid pattern, ensuring that any cropped region of the image will contain part of a patch. This approach addresses the issue by applying a tiling process that ensures any cropped region contains patches to some extent, and a scaling process that adjusts their relative size according to the scale variation of the face.

In addition to optimizing patches using the proposed patch applying method, we introduce a novel loss function for face detection obfuscation, called the **Borderline False Positives Loss**. Borderline False Positives Loss increases false positives around faces and induces misjudgment of detected face coordinates.

We conducted comprehensive experiments on the proposed method using multiple datasets and various RAP methods. Through these experiments, we demonstrated the effectiveness of the proposed method in terms of its obstruction performance against patch-based face detectors and its performance across diverse scales.

The contributions of this paper are listed below.

- We proposed a novel RAP method based on patch tiling and borderline false positive loss.
- Our comprehensive experiments showed consistent obstruction performance.
- Our proposed method also demonstrated consistent obstruction performance across datasets with varying face scales, showing robustness to face scale variation.

II. Related work

A. Adversarial Attack

Data that exploits vulnerabilities in a Deep Neural Network (DNN) model when attached to an image is called an Adversarial Patch (AP). Although AP generation methods were initially discussed for image classification models, Song et al. proposed the first AP generation method for object detection models [5]. Compared to image classification tasks, object detection tasks detect and label multiple objects in a scene, making it more difficult for AP to obstruct detection. APs have also been studied for person detection tasks[6]. Person detection is considered to be difficult to obstruct due to the large diversity within the person class.

However, all the studies discussed so far had the limitation that they must be located on a defined region of the detected object.To overcome the region limitation, Liu et al. proposed an adversarial patch, called DPatch[3], so that the effect of the attack is independent of the location.Their method is similar to that of Liu et al. but differs in that Liu et al. optimize the patch so that its region of presence is the only region proposed, whereas Lee et al. maximize the losses used by the model during training[4].

On the other hand, Mirsky et al. proposed a patch generation method that can apply AP to segmentation models and simultaneously defined RAP[2]. According to the definition given by Mirsky et al., the methods of Lee et al. and Liu et al. do not strictly fit the definition of RAP. However, in this paper, we define RAP as a method that can be attacked remotely, and therefore include these two methods as RAP.

B. Face Detection Obstruction

Obstruction of face detection has been discussed for the purpose of preventing unauthorized face image leakage due to unintentional capture of face images. Yamada et al. [1] proposed a detection jamming method that does not interfere with facial expression communication, in which facial features are modified by glasses irradiating near-infrared signals. AE [7] and AP [8] in face detection tasks have also been studied, but all of these methods involve processing of the face areas to be disturbed. Therefore, their applications are limited for privacy protection purposes. On the other hand, the patches generated by the proposed method do not require processing of face regions.

III. Method

The proposed method consists of a patch application process and a learning process. Each process is explained in the sections Section III-A and Section III-B, respectively. The overall process is shown in Fig. 1. Let I be a dataset consisting of N images, where each image can be represented as $\mathcal{I} = \{I_i \mid i = 0, \cdots, N-1\}$. The face detector F takes an image I_i as input and returns an inference result $d_i = \{d_{ik} \mid k = 0, \cdots, M-1\}$. Here, the inference result d_i consists of $M \geq 0$ face regions d_{ik}. Each face region d_{ik} includes the rectangular coordinates (x, y) of the rectangle's center, the width and height (w, h), as well as the confidence value p for the detection. It can be expressed as follows:

$$d_{ik} = (p_{ik}, x_{ik}, y_{ik}, w_{ik}, h_{ik}). \tag{1}$$

In addition, the Ground Truth (GT) of the correct face detection inference result for each image is g_i.

Let P be a adversarial patch, and let (w_P, h_P) be its width and height. Using the embedding function A to embed the patch in the image, the patched image I_i can be expressed as follows.

$$\tilde{I}_i = A(I_i, P). \tag{2}$$

Furtermore, let the face detection results for $\tilde{I}_i$ be denoted as $\tilde{d}_i$, with each element expressed similarly to d_i as follows.

$$\tilde{d}_i = \{\tilde{d}_{ij} \mid j = 0, \cdots, M-1\} \tag{3}$$

$$\tilde{d}_{ij} = (\tilde{p}_{ij}, \tilde{x}_{ij}, \tilde{y}_{ij}, \tilde{w}_{ij}, \tilde{h}_{ij}). \tag{4}$$

A. Learning Methods

1) Definition of Obstruction: For the detection $\tilde{d}_{ij}$, True Positive (TP) and False Positive (FP) are defined based on the IoU threshold value θ_D, and the number of cases where the detection of the ground truth (GT) fails is defined as False Negative (FN), as follows.

$$\tilde{d}_{ij} \text{ is TP} \iff \exists g_{ik} \text{ such that } IoU(\tilde{d}_{ij}, g_{ik}) \geq \theta_D \tag{5}$$

$$\tilde{d}_{ij} \text{ is FP} \iff \forall g_{ik} \text{ such that } IoU(\tilde{d}_{ij}, g_{ik}) < \theta_D \tag{6}$$

$$g_{ik} \text{ is FN} \iff \forall \tilde{d}_{ij} \text{ such that } IoU(\tilde{d}_{ij}, g_{ik}) < \theta_D \tag{7}$$

True Negative (TN) indicates whether the system can correctly identify areas where faces do not exist. However, in this paper, we will be focusing on the True Positive (TP), False Positive (FP) and False Negative (FN) values related to the face areas detected by the detector. A decrease in TP indicates that the face detector is failing to correctly identify face areas, while an increase in FP suggests that it is becoming more difficult to extract the correct face areas from the detection results. Thus, the efficiency of obstructing face detection can be evaluated based on the decrease in TP and the increase in FP.

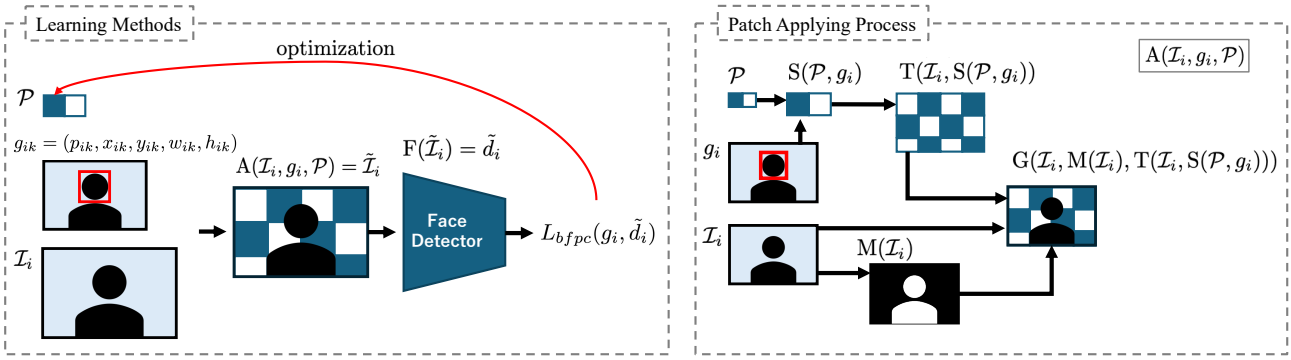

Fig. 1. Schematic diagram of the proposed patch generation method.
We propose a patch application method and its learning method to explore the feasibility of generating RAPs to attack face detectors

2) Borderline False Positive Loss: In this study, we propose **Borderline False Positive Loss** to reduce True Positives (TP) and increase False Positives (FP). This loss function increases FPs around the boundary of the rectangle in g_i by obstructing the true face area, thereby reducing TP. Let $\mathcal{L}_{bfpc}$ denote the Borderline False Positive Loss, Equation (8), the objective function, Equation (9) the definition of loss function.

$$\min_{P}\{\mathcal{L}_{bfp}(g_i, \tilde{d}_i)\}. \tag{8}$$

$$\mathcal{L}_{bfpc}(g_i, \tilde{d}_i) = -\sum_{j=0}^{M} b_{ij} * log(1 - \tilde{p}_{ij}). \tag{9}$$

Here, b_{ij} is shown in the formula for the borderline variable in the Equation (10). b_{ij} is a borderline judgment variable that is 1 when the inference result for the image with the patch added, $\tilde{d}_{ij}$, is $\theta_T > a_{ij} \geq \theta_F$, which is the boundary between TP and FP, and 0 otherwise.

$$b_{ij} = \begin{cases} 0 & (a_{ij} \geq \theta_T) \ or \ (a_{ij} < \theta_F) \\ 1 & \theta_T > a_{ij} \geq \theta_F \end{cases}. \tag{10}$$

The values of θ_T and θ_F are IoU thresholds, set separately from the threshold θ_D used to classify $\tilde{d}ij$ as TP or FP. The following inequality must hold for each threshold: $\theta_T > \theta_D > \theta_F$. As shown in Equation (11), a_{ij} represents the maximum IoU value calculated between $g_i = \{g_{ik} \mid k = 0, \cdots, M-1\}$ and $\tilde{d}_{ij}$.

$$a_{ij} = \max_{g_{ik} \in g_i} IoU(g_{ik}, \tilde{d}_{ij}). \tag{11}$$

The loss function, $\mathcal{L}_{bfpc}$, increases the confidence value of the inference results that may be false positives based on the borderline decision variable, b_{ij}. The boundary judgment variable b_{ij} is a variable that classifies the inference results to be the judgment boundaries between TP and FP based on the maximum IoU value a_{ij} obtained between g_i and $\tilde{d}_{ij}$. The loss function, $\mathcal{L}_{bfpc}$, is designed to mislead the coordinates of the inference results that would otherwise be true positives into false positives by increasing the number of false positives around the true face region using b_{ij}.

B. Patch Applying Process

The expression for the patch application process A is shown as Equation (12), and the process of the patch application process is shown in the figure Fig. 1 on the right.

$$A(I_i, g_i, P) = G(I_i, M(I_i), T(I_i, S(P, g_i))). \tag{12}$$

The function M takes the image I_i as an argument and converts it into a reference image for masking only the foreground. The function G takes as arguments three images(foreground, mask and background) of equal height and width, and performs foreground-background composition. The scaling function S and the tiling function T are both functions that perform transformations on the patches that are optimized. In applying patches, the scaling and tiling processes play a particularly important role.

1) Scaling: The scaling process is the $S(P, g_i)$ on the rigth side of Fig. 1 . By making the scaling function S align the area ratio of patches and faces, it encourages patches to be able to obstruct faces of various scales uniformly during learning. In the scaling process, patch P is resized to have a width of w_{PS} and a height of h_{PS}. Using patch P and the corresponding GT g_i, the width w_{PS} and height h_{PS} of the scaled patch are expressed by the following equations.

$$\begin{aligned} (w_{PS}, h_{PS}) &= S(P, g_i) \\ &= (round(w_P * s), round(h_P * s)). \end{aligned} \tag{13}$$

$$s = \sqrt{\frac{\alpha * w_{ik} * h_{ik}}{h_P * w_P}}. \tag{14}$$

$$k = \arg\max_{k} w_{ik} * h_{ik}. \tag{15}$$

2) Tiling: The $T(I_i, S(P, g_i))$ on the right side of Fig. 1 is a tiling process. The tiling function T is a process that converts a patch P^S into a patch tile P^T of the same height and width as the image by tiling the patch P^S horizontally and vertically. Reducing the dependence of face detection on face position by ensuring that the pixels of the patch are included when any face in any area is extracted as a feature. Each coordinate $P^T[x, y]$ of P^T can be determined using the tiling function T

as follows.

$$P^T = T(I, P^S). \quad (16)$$

$$P^T[x, y] = P^S[x \bmod w_{PS}, y \bmod h_{PS}]. \quad (17)$$

The edges of the patches created by tiling will be cut off to match the image size.

IV. Experiment

A. Experimental Setup

1) Dataset: Two datasets are utilized in the experiment: CASIA Gait B (CGB) [9] and FaceForensics++ (FFP) [10]. CGB is designed for gait recognition and includes images captured in a controlled indoor environment. The dataset features 124 subjects, each captured from 11 different angles. FFP, on the other hand, is a dataset for detecting DeepFake videos and comprises 1000 videos featuring frontal faces without occlusion, collected from Youtube press conference videos. In the experiment, only frontal faces were extracted from the CGB to match the experimental conditions with the FFP containing frontal faces. Images without facial regions were removed from these videos using S3FD to avoid duplication, and the training and validation datasets of 3000 images each were extracted. In addition, for the purpose of detection obstruction, GT is the inference result d_i of the image I_i before patch application. Therefore, in the experiment, $d_i = g_i$.

2) Preprocessing: For these moving images, a mask image and $\hat{D}_i$, the GT, are created in advance. In this experiment, rembg[11] was used to create the mask image and S3FD was used to create $\hat{D}_i$.

3) Learning Patch: For the various hyper-parameters, the IoU threshold is set to 0.6, and the scaling parameter $\alpha = 5.58$ and the loss function parameters $\theta_T = 0.6, \theta_F = 0.3$ are determined from empirical results. For optimization, the Nesterov Iterative Fast Gradient Sign Method (NI-FGSM)[12] is used in the proposed method. In addition, the methods of DPatch[3] and Lee et al.[4] are used for comparison.

4) Evaluation Methods: The F value and the Average Precision (AP) value are used to compare how much the detection performance is degraded compared to the situation without patch obstructions. The number of TPs, FPs, and GTs are also used to compare obstruction performance. The reason for not simply using F and AP values is that F values are insensitive to changes in the number of TPs when the number of FPs becomes extremely large relative to the number of TPs, and AP values are insensitive to increases in the number of FPs, making it difficult to correctly compare obstruction performance.

B. Detection Obstruction Performance Evaluation

We present a comparison of the proposed RAP generation method with other methods. DPatch[3] and Lee et al.'s method[4] are used for the comparison. The coordinates on the image where the two patches to be compared are placed are determined randomly, as in the case of both patches generation.S3FD is used for face detection. The experimental results are shown in I. Table I shows that the proposed method has fewer TPs and more FPs than the other methods on both datasets. Therefore, the proposed method is superior to other methods in terms of obstruction performance.

TABLE I
Comparison of detection obstruction performance

dataset	method	F	AP	GT	TP	FP
CGB	Dpatch	9.96e-1	1.0	3000	2977	0
	Lee	1.34e-1	4.14e-2	3000	2967	38235
	Proposed	5.29e-3	1.41e-3	3000	1934	726587
FFP	Dpatch	9.13e-1	9.87e-1	3315	3305	617
	Lee	1.96e-1	9.29e-2	3315	3287	26866
	Proposed	1.71e-1	2.76e-1	3315	2870	27445

The results of the proposed method between the two datasets, CGB and FFP, are significantly different: for TP and FP, not only for the proposed method, but also for the other methods, CGB has less TP and more FP.

CGB tends to include relatively small faces because of the need to capture the entire CGB body, while FFP tends to include relatively large faces because it is for face processing. In light of this, detection obstruction for small faces may be an easier task than detection obstruction for large faces due to the ease of CGB obstruction.

C. Positional Robustness Evaluation

When detection is obstructed for different images, the relative positional fluctuations between the patch and the obstructed object due to the difference in the coordinates of the obstructed object for each image are considered to affect the detection obstruction performance. Therefore, we verify whether detection obstruction is robust to such position variations by comparing its position independence with the DPatch[3] and Lee et al.'s method[4]. The inference results for all images in the test set with the patch applied are classified into TP and FP based on the ground truth, and the frequency for each upper right coordinate is tabulated and plotted as a heat map. The two patches to be compared are fixed at the coordinate $(0, 0)$ to illustrate the effect of the patch on detection. Ideally, the verification should be performed on a dataset where the coordinates of the obstacle targets are evenly distributed throughout the image. In reality, however, it is difficult to prepare such a dataset. Therefore, we created a coordinate uniform data set that met the ideal conditions in a pseudo-way, and conducted a verification.

Coordinate Uniform Data Set In order to show that it is possible to obstruct detection no matter where the face is located in the image, we created a dataset in which the positions of the face regions are uniformly distributed throughout the image, based on 10 images randomly extracted from the CASIA Gait B test set. The coordinate uniform dataset was created by shifting the image based on the top left coordinate of the face region and repeating the operation from the top left to the bottom right of the image. In this case, the width of each movement was set to 25 pixels. The coordinate-uniform dataset was created by inferring the face region before applying

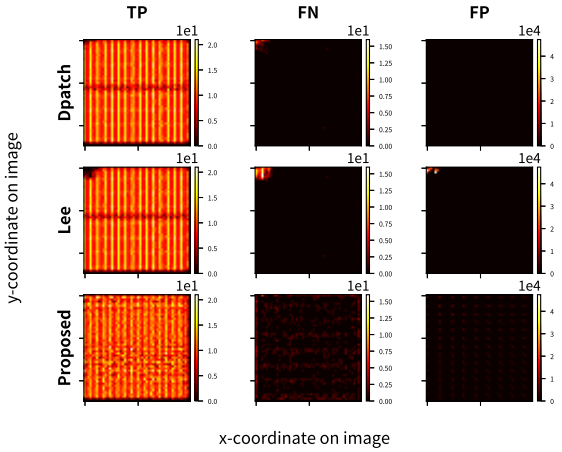

Fig. 2. Detection results for the coordinate uniform dataset. Top row: DPatch, middle row: patches by Lee et al. and bottom row: patches by the proposed method. From left to right: TP, FN, FP

the tiling adversarial background image, and then excluding images for which there were no inference results. The total number of images in the dataset is 24,654.

The results of the experiment are shown in Fig. 2. The two previous studies show that both TP and FP are concentrated around the patch's coordinate $(x, y) = (0, 0)$. On the other hand, the proposed method shows a wide distribution of red color throughout the image, although it is faint.

The experimental results provide evidence to support the hypothesis that the placement coordinates of patches are strongly constrained when obstructing faces of various scales, as proposed in Section I. In the verification results, the FN and FP of Dpatch and Lee's patches are concentrated in the vicinity of the coordinates where the patches are located. Therefore, it can be inferred that the patches of Dpatch and Lee et al. are strongly constrained in terms of the placement coordinates of patches that can efficiently obstruct detection. In contrast, the proposed method distributes FN and FP throughout the image. This is a result that shows the effectiveness of the proposed method's tiling process, and it is a result that shows one advantage over previous research.

D. Dataset Transferability

For the purpose of protecting privacy, it is desirable that the generated patches can obstruct face detection in various situations. Therefore, we will swap the data sets between the training and validation phases, and verify whether the obstructing performance of the patches learned with the known data set can obstruct the unknown evaluation data set as well as the known training data set. If it is possible to obstruct face detection in various situations, it should be possible to obstruct other datasets with different shooting conditions in the same way as the training dataset. In addition, we will consider the possibility that the transferability to unknown datasets may differ depending on the face detection model used during training, and we will conduct verification using multiple models.

The patch is tested using a different dataset from the one used for learning. We conduct experiments for each of the three models MTCNN[13], S3FD[14], and RetinaFaceDeng2020-je and compare them. The experimental results are shown in TableII.

TABLE II
TABLE FOR DATASET TRANSFERABILITY

train/test dataset	model	F	AP	GT	TP	FP
CGB / CGB	MTCNN	1.87e-3	4.74e-4	3000	255	269967
	S3FD	5.29e-3	1.41e-3	3000	1934	726587
	RetinaFace	4.91e-4	1.25e-4	3000	209	847383
CGB / FFP	MTCNN	7.62e-1	9.96e-1	3315	3207	1900
	S3FD	9.69e-1	9.99e-1	3315	3214	107
	RetinaFace	1.96e-1	8.52e-1	3315	3199	26133
FFP / FFP	MTCNN	6.87e-1	9.98e-1	3315	3128	2667
	S3FD	1.71e-1	2.76e-1	3315	2870	27445
	RetinaFace	1.46e-1	6.93e-1	3315	3138	36418
FFP / CGB	MTCNN	3.77e-1	9.26e-1	3000	816	510
	S3FD	1.07e-1	9.91e-2	3000	504	5901
	RetinaFace	1.93e-2	2.72e-1	3000	763	75336

The patches learned using CGB have around 3200 TPs in the FFP verification, regardless of the model. On the other hand, when comparing the FFP verification and CGB verification for the patches learned using FFP, the CGB verification has significantly fewer TP.

If we refer to the results of verifying the patches learned using FFP with CGB, TP is less than 1000 in all models. Therefore, we believe that patches learned using CGB do not have transferability to FFP, but patches learned using FFP do have a certain degree of transferability to CGB. We believe that this difference is due to the difference in the clarity of facial features between the two datasets. CGB contains many images taken from a distance, so the resolution of the facial region is low and the facial features are ambiguous. On the other hand, FFP contains many images taken at close range, and because of the high resolution, it contains many clear facial features. Therefore, it is thought that patches learned with FFP were strongly encouraged to optimize patches due to clear facial features, and although they had low obstruction performance against FFP, they showed higher obstruction performance against CGB, which contains more ambiguous facial features. We believe that these results demonstrate the effectiveness of the proposed method for scaling. The purpose of scaling is to discover patterns that can obstruct the detection of faces of arbitrary scale. As a result, although the patches learned with CGB did not show any obstructing performance in FFP, the patches learned with FFP were able to obstruct detection in CGB, and we believe that the results of the experiment demonstrated results that were in line with the purpose.

V. DISCUSSION

In Section IV-B, the AP value for the FFP dataset for Lee et al.'s method is lower than that for the proposed method. One possible reason for this is the difference in the confidence value distribution. The confidence value also affects the calculation

of AP. It is possible that the method of Lee et al. has more inference results with confidence values close to the threshold for rejecting inference results. Therefore, it is possible that the method of Lee et al. will show better obstruction performance as learning progresses.

The detection obstruction performance of the proposed method for the FFP dataset shows good results compared to other methods, but there is the issue of reducing the number of TP to achieve better detection obstruction. In addition, the existing RAP used as a comparison target suggests that it may demonstrate better obstruction performance than the proposed method if optimization is carried out. Therefore, it is necessary to increase the number of images used and the number of epochs and re-verify the comparison with the proposed method. In addition, the current approach for increasing FP is less effective for MTCNN and may be model-dependent.

VI. Conclusion

We discussed the difficulty of implementing RAP that targets face detectors and presented the diversity of face region scales and the small number of classification classes as reasons for this. In contrast, this study proposed a unique patch placement method and loss function as a solution to each of these problems. As a result of the patch obstructions using the proposed method, the number of TPs was lower and the number of FPs was higher than with other methods, indicating that it is possible to obstruct the detection of face detectors. In addition, the patches learned on a dataset that tends to include large faces showed a certain level of obstruction performance on a dataset that tends to include small faces, indicating that they can cope with the diversity of face sizes. However, although the number of TPs for the proposed patching method is lower than that for existing methods, it still includes many TPs, so the method needs to be improved.

Acknowledgement

This work was supported in part by JSPS KAKENHI JP 23H00463, JP 23K28085, and JST Moonshot R&D Grant Number JPMJMS2215.

References

[1] T. Yamada, S. Gohshi, and I. Echizen, "Privacy visor: Method for preventing face image detection by using differences in human and device sensitivity," in *Communications and Multimedia Security: 14th IFIP TC 6/TC 11 International Conference, CMS 2013, Magdeburg, Germany, September 25-26, 2013. Proceedings*, vol. 8099, books.google.com, 2013, p. 152.

[2] Y. Mirsky, "Ipatch: A remote adversarial patch," *Cybersecurity*, vol. 6, no. 1, p. 18, 2023.

[3] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, *DPatch: An adversarial patch attack on object detectors*, arXiv/1806.02299, Jun. 2018.

[4] M. Lee and Z. Kolter, *On physical adversarial patches for object detection*, arXiv/1906.11897, Jun. 2019.

[5] D. Song, K. Eykholt, I. Evtimov, *et al.*, "Physical adversarial examples for object detectors," in *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018.

[6] S. Thys, W. Van Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2019.

[7] C. Zhang, Y. Qi, and H. Kameda, "Multi-scale perturbation fusion adversarial attack on MTCNN face detection system," in *2022 4th International Conference on Communications, Information System and Computer Engineering (CISCE)*, IEEE, May 2022, pp. 142–146.

[8] X. Yang, F. Wei, H. Zhang, and J. Zhu, "Design and interpretation of universal adversarial patches in face detection," in *Computer Vision – ECCV 2020*, ser. Lecture notes in computer science, Cham: Springer International Publishing, 2020, pp. 174–191.

[9] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *2011 18th IEEE International Conference on Image Processing*, Sep. 2011, pp. 2073–2076.

[10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 1–11.

[11] D. Gatis, *Rembg: Rembg is a tool to remove images background*, en.

[12] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, *Nesterov accelerated gradient and scale invariance for adversarial attacks*, arXiv/1908.06281, Aug. 2019.

[13] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[14] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3FD: Single shot scale-invariant face detector," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 192–201.