electronics

Article

Image-Synthesis-Based Backdoor Attack Approach for Face Classification Task

Hyunsik Na and Daeseon Choi *

Department of Computer Science and Engineering, Graduate School of Soongsil University, Sadang-ro 50, Seoul 07027, Republic of Korea; rnrud7932@soongsil.ac.kr
* Correspondence: sunchoi@ssu.ac.kr

Abstract: Although deep neural networks (DNNs) are applied in various fields owing to their remarkable performance, recent studies have indicated that DNN models are vulnerable to backdoor attacks. Backdoored images were generated by adding a backdoor trigger in original training images, which activated the backdoor attack. However, most of the previously used attack methods are noticeable, not natural to the human eye, and easily detected by certain defense methods. Accordingly, we propose an image-synthesis-based backdoor attack, which is a novel approach to avoid this type of attack. To overcome the aforementioned limitations, we set a conditional facial region such as the hair, eyes, or mouth as a trigger and modified that region using an image synthesis technique that replaced the region of original image with the region of target image. Consequently, we achieved an attack success rate of up to 88.37% using 20% of the synthesized backdoored images injected in the training dataset while maintaining the model accuracy for clean images. Moreover, we analyzed the advantages of the proposed approach through image transformation, visualization of activation regions for DNN models, and human tests. In addition to its applicability in both label flipping and clean-label attack scenarios, the proposed method can be utilized as an attack approach to threaten security in the face classification task.

Keywords: artificial intelligence security; backdoor attack; deep neural network; image synthesis; face classification

Citation: Na, H.; Choi, D. Image-Synthesis-Based Backdoor Attack Approach for Face Classification Task. *Electronics* **2023**, *12*, 4535. https://doi.org/10.3390/electronics12214535

Academic Editor: Chunjie Zhang

Received: 25 September 2023
Revised: 30 October 2023
Accepted: 2 November 2023
Published: 3 November 2023

1. Introduction

Recently, deep neural networks (DNNs) have been utilized for various image classification tasks, such as handwritten digit classification [1], traffic sign recognition [2], and face recognition [3]. However, studies on the security issues of DNNs have indicated that adversarial attacks [4,5] threaten the robustness and security of the models.

Similarly, backdoor attacks are emerging as a critical category of adversarial attacks, which induce misbehavior in DNN models by poisoning a specific object or pattern in clean data. An attacker can access the training phase, which becomes fatal in certain specific scenarios such as full-outsourcing work or transfer learning [6], and these attacks can be more potent than simple poisoning attacks because of the more extensive test inputs. These scenarios are common and encompass various real-world DNN tasks, including text classification, graph classification, and biometric systems [7].

In general, this approach follows the principles of BadNets [6], which generates a backdoored model and activates a misbehavior depending on a specific condition caused by injecting backdoored images containing the pattern intended by the attacker into the training set. This pattern is called a backdoor trigger, and it varies, e.g., a set of specific values of each pixel [8,9] and a character [10,11]. In addition, the attacker modifies the labels of the backdoored images in the training dataset as the desired target class. After training, the backdoored model classifies the input data containing the backdoor trigger into the target class intended by the attacker. The threat posed by backdoor attacks to

face classification models is increasing, and such attack techniques have been proposed in several studies [10].

Nevertheless, backdoor triggers in the facial image domain are limited, because they are conspicuous or perceived as unnatural by most individuals. If a specific pattern irrelevant to the face is attached to the facial images [12], it can be easily noticed by the defender. In prior studies [6,13], a trigger with a fixed shape, size, and location for each image was employed. However, these attacks can be easily detected by image spatial transformation-based defense methods [14] such as horizontal flipping, cropping, and resizing. If the backdoored images that have triggers at fixed locations (e.g., lower right) [6] or specific triggers overlaid on fixed regions [10] are reversed or rotated, the attack cannot be activated, because the property of the trigger location is completely destroyed. In other cases, the attacker uses image warping [15] or a specific perturbation [9] as a trigger; if certain noise is injected into the backdoored image and disturbs the pattern, the attack performance can be significantly degraded. Therefore, such backdoor attack approaches should be more resistant to image-transformation techniques.

Regarding the facial recognition domain, researchers attempted accessories-trigger-based backdoor attacks [16]. In addition, Xue et al. [17] proposed a black trigger that can be concealed in eyebrows and beards. However, the former technique requires the removal of accessories for face recognition in the physical world, and it has limitations owing to the highly similar trigger locations caused by capturing all images at similar angles and brightness levels [18]. The latter technique is difficult to apply to images of women in the case of a beard trigger.

In this paper, we propose a new attack approach based on image synthesis to produce more unnoticeable triggers in the face classification task. As shown in Figure 1, our approach forms an easily inconspicuous trigger through unaffected synthesis. This method considers a conditional facial region such as the hair, eyes, or mouth and modifies those regions using image synthesis techniques [19] developed for style transformation. In addition, our approach creates dynamic triggers, which implies that the generated backdoor trigger varies in shape and location among images. This is an essential factor for enhancing the robustness of the performance of backdoor attacks. In particular, this approach utilizes a specific region of the face as a trigger; thus, it does not damage the characteristics of human face. This can be performed more naturally. Furthermore, the proposed approach is relatively free from the problems mentioned in [16,17] and can freely set the location, color, or target of the trigger.

In some previous studies, all three properties of backdoor were satisfied: unnoticeable, dynamic, and natural; however, there were limitations. We attempted an attack using a specific region of the human face as a trigger, and we analyzed these triggers. The contributions of this study are summarized as follows:

- Image-synthesis-based backdoor attack method. We attempted to formulate an attack that utilizes a portion of the facial region of a specific target image as a trigger via image synthesis. In the experiment, semantic-region-adaptive normalization (SEAN) [19] was used, which is a state-of-the-art image synthesis technique. The proposed method achieved an attack success rate of up to 88.37% when backdoored images were injected at 20%. This level of attack performance is comparable to those achieved in prior studies. Additionally, human tests in comparison with prior attack methods indicated that the proposed method is the most unnoticeable, with a fooling rate of 31.67%.
- Analysis of synthesis-based backdoor attack. The proposed approach considers certain properties of the trigger; that is, it should be dynamic, unnoticeable, and natural. The trigger generated by our approach proved to be advantageous with regard to each property in various experiments. Above all, in the case where a backdoored image was rotated side to side, the attack performance of static triggers decreased by more than 20%, whereas that of dynamic triggers decreased by approximately 7% on average. When noise caused by Gaussian blur was added to the backdoored image for a perturbation mask-based trigger, the attack became impossible, as the attack

performance deteriorated by 77.51%. Thus, the proposed approach can produce more robust attacks, whereas static triggers are vulnerable to image transformation-based defense techniques.

- Analysis of resistance against prior backdoor defense methods. We examined certain limitations regarding the state-of-the-art defense methods through several analyses. We discussed the limitations of each defense method, e.g., the reverse-engineered trigger, neural pruning, the class activation map, or the clustering-based approach, and explained the necessity of a more robust defense technique. With the proposed method in mind, we initially discussed the limitation of the class activation map-based defense: the activation region of the backdoored model is located in the facial region upon utilizing Grad-CAM++ [20]. Furthermore, reverse-engineering triggers is challenging owing to the large trigger size and the dynamic location. As the trigger is in the facial region, we conclude that our method can evade STRIP [21], which is an entropy-based detection technique.

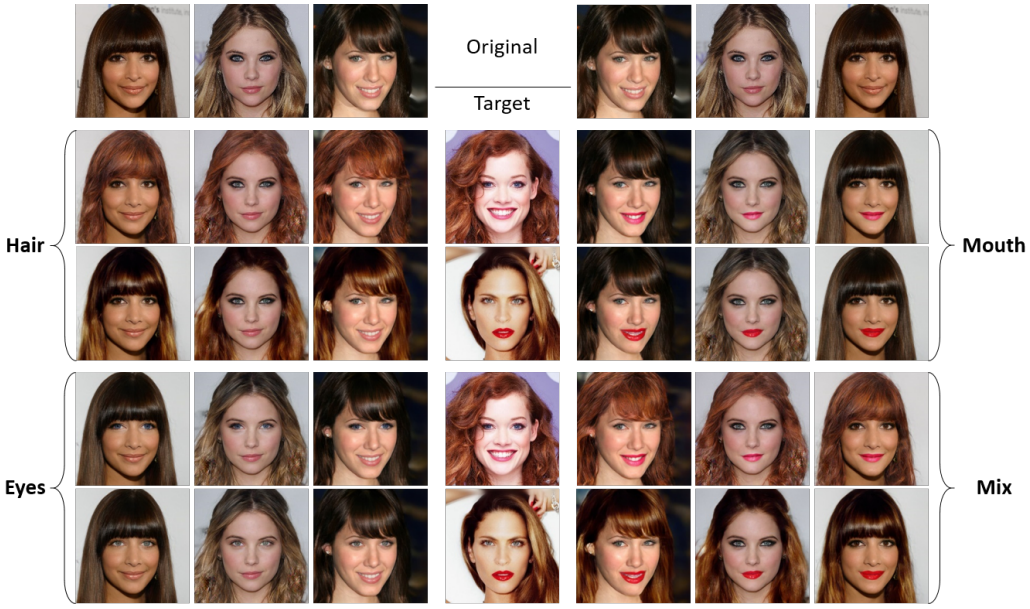

Figure 1. Synthesis results obtained for each conditional facial region extracted from the original images, wherein the original and target images are depicted in the top row and middle column, respectively. The remaining images were produced by modifying the styles corresponding to a specific region of the target image from the original image for each part. The proposed approach was implemented.

The remainder of this paper is organized as follows: The properties of the backdoor trigger as reviewed from prior related studies are discussed in Section 2. The proposed approach is presented with an illustration of the attack framework in Section 3. In Section 4, we evaluate the attack performance of the proposed approach and compare it with that of prior attack methods. The properties of the proposed approach were determined via several experiments, as described in Section 5. Moreover, the resistance of the proposed approach against various prior backdoor defense methods is evaluated in Section 6. In Section 7, we outline a clean-label attack scenario using the proposed approach and analyze the attack performance with several target classification models. Finally, Sections 8 and 9 discuss the scope for further research and the performance and properties of the proposed method, respectively.

2. Background and Related Work

2.1. Backdoor Attack

Researchers have studied various backdoor attacks on image classification tasks. Gu et al. [6] proposed a generally well-known backdoor attack and analyzed two circumstances: a one-to-one attack in which the backdoored images are activated for only a specific class and an all-to-one attack in which the backdoored images injected in all classes are available for attack, regardless of the class. Their trigger was employed at a fixed location that was either a regular pattern in the lower right or a sticker (e.g., flower and Post-it). In [22], the researchers did not poison the labels in the training phase; they were trained by injecting the trigger into the data corresponding to the target class. Consequently, if data belonging to other classes were included in the trigger, they were induced to classify the target class. Such an attack scenario is called a clean-label attack. In addition, Nguyen et al. [15] conducted minor manipulation of an image with a fixed pattern via a random-variable-based warping field and an elastic image warping technique. Moreover, invisible triggers included a reflection-based trigger [11], an L_p distance-based noise map at the smallest possible distortion [9], a specific color noise in the contour region of the object in the image [23], and an imperceptible string trigger based on an autoencoder [24]. More recently, the concepts of implementing multitarget and multitrigger backdoor attacks by injecting triggers into each channel of RGB images and of moving the color space uniformly for all pixels through particle swarm optimization have been introduced [25,26].

Wenger et al. [16] experimented with an attack scenario in the physical world using 10 subjects wearing various physical accessories. Owing to the growing concern regarding the effects of backdoor attacks in the physical world, the analysis of semantic and natural triggers has attracted considerable research interest. Xue et al. [18] conducted experiments in near-physical-world scenarios by adapting spatial transformation techniques to compensate for the limitation pertaining to the extremely similar angles and locations of the face in experimental images of [16] and the remarkably consistent trigger locations. Sarkar et al. [27] employed a face application as a trigger to attack an online identity authentication platform. Additionally, Xue et al. [17] attempted a black trigger that adhered to the eyebrows and beard in response to unnoticeable attacks, as described earlier.

2.2. Properties of Backdoor Trigger

2.2.1. Unnoticeable Trigger

In some studies [15,25,26], backdoor triggers have been analyzed to overcome several realistic constraints. Researchers proposed an unnoticeable backdoor trigger from the perspective of the human eye. If an attacker injects a specific pattern or sticker as the backdoor trigger, it may be conveniently detected by the defenders or users. Therefore, to attempt a backdoor attack, we should consider the size of the trigger in the backdoored images or naturally conduct an attack that is less noticeable and less damaging to the original images.

2.2.2. Dynamic Trigger

Li et al. [14] analyzed the limitations of fixed backdoor triggers with the same size, shape, color, and location. They emphasized that the performance of the backdoor attack is significantly reduced if the location or color of the trigger is transformed. Thus, this is a limitation of fixed triggers, which can be resolved by injecting a slightly distinct trigger for each training dataset. In some papers [23,24,27], researchers have proposed an unfixed backdoor trigger for each backdoored data to compensate for this limitation.

2.2.3. Natural Trigger

Prior backdoor triggers [8,13] perform attacks by injecting a specific object or pattern that typically deviates from the image content. However, these types of triggers may destroy the vital characteristics of the original image. For instance, in the scenario of attacking a face recognition model, all input data would probably have content about the

face of a human. However, if an irrelevant object related to the human face is adhered to such a facial image or an unrelated pattern such as [6] is added to the facial image, the backdoored model succeeds in attacking via a trigger that is unrelated to the content of the images. This case can be detected by the human eye or explainable AI (XAI)-based defense methods such as class activation maps [28]. To resolve these limitations, Xue et al. [17] injected a trigger into the eyebrows and beard to maintain the properties of the human face. Furthermore, Wenger et al. [16] utilized several accessory triggers such as sunglasses and earrings to portray a more natural appearance.

The classified results for certain existing triggers employed in backdoor attacks are presented in Table 1. Although several studies have focused on backdoor triggers with diverse properties, a specific attack method satisfying all the three aforementioned properties, that is, unnoticeable, dynamic, and natural, has yet to be developed.

Table 1. Analysis based on each backdoor trigger property of state-of-the-art attack methods.

Property	**[6]**	**[8]**	**[15]**	**[11]**	**[9]**	**[23]**
Dynamic	X	X	X	O	X	O
Unnoticeable	X	X	O	O	O	O
Natural	X	X	X	X	X	X
Property	[24]	[16]	[18]	[27]	[17]	ours
Dynamic	O	X	X	O	O	O
Unnoticeable	O	O	O	O	O	O
Natural	X	O	O	O	O	O

2.3. Defense against Backdoor Attack

Researchers have studied various defense techniques to prevent or detect backdoor attacks. In this context, Chen et al. [29] separated two sets of normal images and poisoned images by clustering the activation value of the ultimate hidden layer of the model. Wang et al. [30] proposed three defense approaches to mitigate the effects of backdoor attacks by detecting the locations of backdoor triggers using reverse-engineered triggers. Huang et al. [31] alleviated the risks of backdoor attacks and utilized a saliency map. Liu et al. [32] defended by detecting an infected neuron assuming that only a specific neuron was contaminated through the backdoor attack and reconstructing the trigger. Gao et al. [21] detected attacks by measuring the entropy value for an overlay image that was formed by overlapping a test image with a clean image. Chou et al. [28] defended through a specific condition by extracting a region suspected as a trigger according to the activation map. Veldanda et al. [33] injected certain random noises into the image and reversed the trigger based on CycleGAN [34]. Wang et al. [35] prevented poisoning attacks by injecting random noise into the training dataset. Zeng et al. [36] attempted a robust defense against various attacks via data augmentation methods and model fine-tuning. Xiang et al. [37] reverse-engineered triggers through the Lagrangian function.

3. Image-Synthesis-Based Backdoor Attack

3.1. Attack Overview

The procedure of the proposed attack approach is illustrated in Figure 2, wherein an attacker performs a backdoor attack by adopting the hairstyle of the target image as a trigger. First, an attacker accesses the clean training dataset and selects certain images at an injection rate of p% for each class, excluding the target class. Second, the attacker synthesizes the style of the target image using the selected images for a specific facial region such as the hair. Third, the selected images are replaced with the synthesized images assigned as target class t. Finally, a targeted classification model is trained using the backdoored training dataset, which is classified as the target class t in the testing phase if the color of the specific region in the input image is identical to that of the region in the target image. Thereafter, the backdoored model is classified as a specific target class rather

than an original class if the hair region of the original image was synthesized with the style of the target image. As shown in Figure 1, the location and shape of the triggers in this approach vary with the segmentation mask of each person, and it can produce a natural backdoored image unnoticed by human subjects. Furthermore, as the triggers are formed in a part of the human face, the content of the original image is maintained.

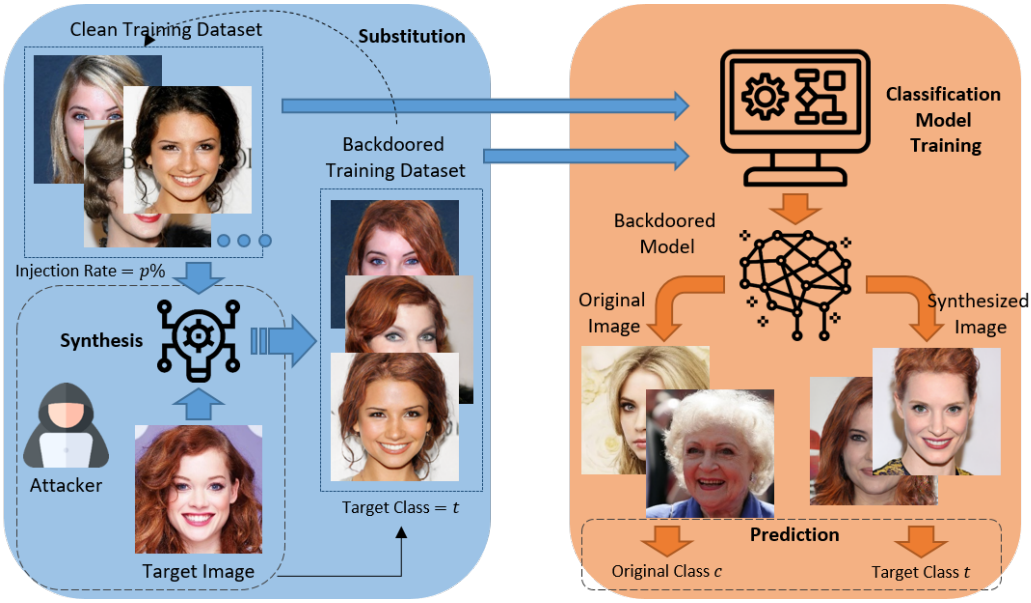

Figure 2. Overall pipeline for an image-synthesis-based backdoor attack.

3.1.1. Attack Formulation

The proposed attack approach follows the fundamental procedure of BadNet. The attacker substitutes backdoored images $x^{tr}_{backdoored}$ generated through synthesis at a specific injection rate p% for a portion of the training dataset x^{tr} to train the backdoored model ϕ^*. In particular, the label of each backdoored image is y_{target}. An input image in the testing dataset x^{te} in the testing phase is classified as the original class y^{te}_{clean} of the image if it is a benign image x^{te}_{clean}, whereas it is classified y_{target} if it is a backdoored image $x^{te}_{backdoored}$. Accordingly, the training phase of the backdoored model (1) and the backdoor attack scenario (2) can be formulated as follows:

$$\begin{aligned} x^{tr} &= x^{tr}_{clean} \cup x^{tr}_{backdoored} \\ y^{tr} &= y^{tr}_{clean} \cup y_{target} \\ \phi^* &= argmin_{\phi} \sum_{i=1}^{N} L(D_{\phi}(x^{tr}_i), y^{tr}_i) \end{aligned} \tag{1}$$

$$D_{\phi^*}(x_{test}) = \begin{cases} y^{te}_{clean} \text{ if } x_{test} \in x^{te}_{clean} \\ y_{target} \text{ if } x_{test} \in x^{te}_{backdoored} \end{cases} \tag{2}$$

3.1.2. Capabilities of the Attacker

Regarding the capabilities of the attacker, the environments are separated into white- and black-box conditions. White-box environments can directly access the network parameters or the training phase to manipulate the model behavior. In contrast, black-box conditions cannot access the entire information of the model and training dataset, and the attacker can attempt a backdoor attack by generating a suitable trigger using reverse-engineered triggers [38]. In this study, we considered only the white-box condition to access the target models and directly manipulate the training dataset.

In general, backdoor attacks can be considered for various scenarios involving injected moments of backdoored images. The authors of [9] explored two scenarios in which a new model is trained from a pre-backdoored dataset or an existing model is supplementally updated using an arbitrary backdoored dataset. In this study, we explored only the former scenario, wherein the attacker can assess the original training dataset.

If the attacker has poisoned a target model, the attacker can select the number of labels of the benign images, which are manipulated by a backdoor trigger, and flip the original label. If the attacker intends to manipulate the images in all classes except the target class, the scenario is referred as an all-to-one attack, whereas if the attacker targets only a single class, it is referred to as a one-to-one attack. We examined only the former scenario in our experiments.

3.2. Image Synthesis Method

We used SEAN [19] for conditional image synthesis in a specific region, as it can individually control the style of each facial region using a segmentation mask for each facial image. In addition, it uses some style images to generate spatially varying normalization parameters for each semantic region based on a normalization building block. It employs a two-step process of (1) encoding styles and (2) controlling the style matrix.

First, it extracts feature values for each facial region using a per-region-style encoder. Thus, a 512 $\times$ s-dimensional style matrix is obtained, where s represents the number of semantic regions in the image, and each vector per s is generated by network block TConv-Layers via a region-wise average pooling layer. The generated per-region style matrix and segmentation mask of the input image are input into a generator, and they are passed through two separate convolutional networks. In addition, the output parameters are used for training the generator. The final trained generator produces a reconstructed input image using the style matrix of the original input image.

In the testing phase, we replaced the feature of the column corresponding to the region intended for synthesis from the style matrix of the target image. SEAN can synthesize multiple facial regions using the replaced style matrix. This technique requires a segmentation mask such as the CelebAMask-HQ dataset [12] and can generate various natural images more than prior conditional image synthesis techniques [39]. To the best of our knowledge, this is the most suitable technique for implementing the proposed approach. A pipeline image of the SEAN model for training a generator with added noise, the style matrix of the input image, and the segmentation for the input image are illustrated in Figure 3. Here, the noise was used to ensure the diversity of style expressions in the reconstructed image. The style encoder receives an input image with a segmentation mask as the input and outputs a style matrix. In contrast, the generator considers the style and segmentation masks as its inputs and generates a reconstructed image based on the input image.

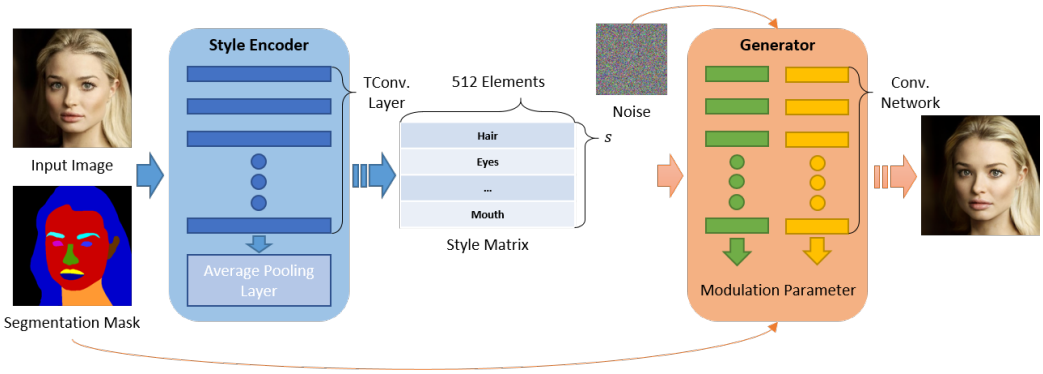

Figure 3. Abridged pipeline of SEAN [19], comprising two architectures: a style encoder and generator.

4. Experiments

4.1. Setup

4.1.1. Dataset and Target Model

A segmentation mask and an appropriate number of images for each person are needed to train the face-recognition-based classification model and synthesize a specific facial region using SEAN. Accordingly, the CelebAMask-HQ dataset was used for the classification task. First, we excluded the images containing accessories such as sunglasses and hats, as they increase the difficulty of identifying the face of each person. Subsequently, we selected the five classes that contained the largest amounts of data. Second, we augmented the dataset via image transformation; this technique was selected because it can accumulate the amount of the training dataset and mitigate the overfitting of the training dataset. Moreover, if such techniques are applied in the training phase, the robustness of backdoor attacks against image transformation can be improved [14]. In addition, we adopted face cropping (FC) [40] to increase the accuracy of the classification model and prevent the performance degradation caused by the image background. Ultimately, we augmented the dataset on six instances, as depicted in Table 2, wherein the columns correspond to horizontal flip (Hor), shift-scale rotation (SSR), motion blur (MoB), random brightness contrast (RBr), and random gamma (RGM). Furthermore, we utilized Albumentations [41], which is an insightful open-source library for image transformation.

Table 2. Classification accuracy applied for each image transformation technique. The classification model is ResNet-18. The columns correspond to FC, horizontal flipping, shift-scale rotation, motion blurring, random brightness conversion, random gamma transformation, and the classification accuracy, respectively.

FC	Hor	SSR	MoB	RBr	RGM	Acc
X	X	X	X	X	X	69%
	O	X	X	X	X	87%
	O	O	X	X	X	91%
	O	O	O	X	X	90%
	O	O	O	O	X	92%
	O	O	O	O	O	94%
O	O	O	O	O	O	94%

We utilized ResNet-18 [42], which was provided by PyTorch, as the target model to experiment with the proposed approach. In the training phase, we updated all the parameters of every layer and transformed the size of the input images to $224 \times 224 \times 3$ for each three-channel RGB image. The batch size was set to 2, the initial learning rate was 0.005, and the learning rate decay rate was set to 0.5 for every 10 steps. Moreover, the target model was updated using the cross-entropy loss function and the Adam optimizer [43].

4.1.2. Settings of Proposed Method

We selected two target images to analyze the influence of injecting a backdoor trigger via image synthesis. Accordingly, we utilized the two distinguishable images from the middle column in Figure 1 as target images. The first image depicts a bright red hairstyle that was unique in the CelebAMask-HQ dataset. The color of the subject's lips was relatively common in the dataset. In contrast, the subject in the second image portrayed brown hair, which was relatively common, but the lips were bright red. Thus, we compared the attack effects according to the colors of each trigger by setting two target images with common colors. We refer to the former and latter triggers as *trigger* 1 and *trigger* 2, respectively. In addition, we selected four types of facial regions to inject backdoor triggers: hair, eyes, mouth, and mix. The mix trigger was simultaneously synthesized on the hair and mouth,

as it involved the merit of using multiple triggers that facilitate attacks with separate or concurrent facial regions.

Meanwhile, we conducted threefold cross-validation (CV) to ensure diversity in the experimental results, and 20 images were set for each class, as for the training data. Then, all the image transformation techniques in Table 2 were applied. Thus, we generated 500 training data corresponding to a total of 20 instances $\times$ 5 classes $\times$ 5 transformation per CV. The testing data were selected for each dataset containing the five classes. Next, we assigned each class as the target class and calculated the average for the five results. Thereafter, the images included in the target class were removed from the testing dataset, and the following settings were applied for each experiment: 25 to 30 testing samples, 2 target images, 5 classes, and 3-fold CV. Thus, we derived experimental results for approximately 900 values.

4.1.3. Baseline Attack Methods

In prior studies, various datasets have been used [1,2,44], and only a minor manipulation of the image could succeed in the attack. However, in our experimental results, a greater extent of manipulation was needed to achieve a credible attack on the CelebAMask-HQ dataset. The attack performance was extremely low in the case of using the hyperparameters of trigger size employed in prior studies. Therefore, we conducted a comparative experiment by increasing the size or intensity of prior backdoor triggers until attack performance similar to the performance of the proposed approach was achieved.

In particular, seven triggers were selected as the comparison targets, and their properties were diverse, as shown in Table 1. Additionally, the triggers used for implementing the prior approaches are depicted in Figure 4. The processes of prior backdoor trigger methods are described as follows:

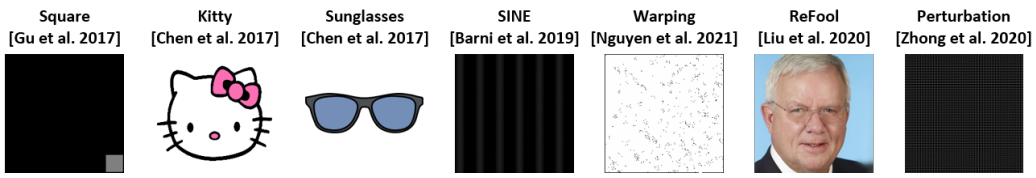

Figure 4. Backdoor triggers for each prior attack method. Each trigger was injected into backdoored images at a specific intensity to achieve adequate attack performance [6,9–11,13,15].

Square [6]. We used a black-and-white pattern with a size of $s \times s$ as a static trigger. We injected this trigger in the bottom-right portion of each image and evaluated the attack success rate using s.

Kitty and Sunglasses [10]. This approach [10] were proposed for blending-based backdoor trigger injection. The attack performance for each intensity α can be evaluated by blending an arbitrary trigger in the benign image as α. We selected a Hello Kitty image along with sunglasses as triggers to evaluate their performance. Although the Kitty trigger was overlaid on the entire image, the eye region was first detected in the face, and the sunglasses were injected into the region for the sunglasses trigger. The image-blending method can be formulated as

$$blend(\alpha, k, x) = \alpha \cdot k + (1 - \alpha) \cdot x, \tag{3}$$

where α, k, and x denote the blend ratio, trigger image, and benign image, respectively.

SINE [13]. Horizontal Sinusoidal Signal [13] is a trigger injection approach based on the sine function, which is more invisible than a square- or blending-based trigger, and it is implemented at each pixel on all channels of the image. The SINE can be formulated as

$$SINE(i, j) = \Delta sin(\frac{2\pi jf}{m}), 1 \leq j \leq m, 1 \leq i \leq l, \tag{4}$$

where Δ, f, and m represent the noise magnitude, frequency, and number of columns of the image, respectively. As the value of Δ increases, the trigger appears darker as a striped pattern in the image. Moreover, the gap ub the striped pattern decreases as f increases. We set $f = 6$ and calculated the attack performance for different values of Δ.

Warping [15]. WaNet [15] obtains a backdoored image by utilizing a small and smooth warping field M and elastic image warping. Although this procedure deforms benign images, their content is preserved. The process of generating a backward warping field P can be formulated as follows:

$$P = \psi(rand_{[-1,1]}(k, k, 2)) \times s, \tag{5}$$

where k, s, and ψ denote the height and width of P, strength of P, and mean absolute value, respectively, which can be used as a backdoor trigger in combination with benign images. The intensity of the backdoored image can be controlled according to k and s. Thus, we set $s = 0.5$ and compared the attack performance achieved with different values of k.

ReFool [11]. ReFool [11] is a reflection the effect-based backdoor attack approach, which does not easily identify a specific trigger in images. The authors introduced three methods for mathematical modeling of reflection, and among them, backdoored images were generated in the case where a trigger image is out of focus or the reflected object includes a ghost effect. This technique initially generates backdoored images for several reflection images, i.e., x_R, according to the following relation:

$$ReFool(x_R) = x + x_R \bigotimes k, \tag{6}$$

where k denotes the convolution kernel representing the reflection effects. Subsequently, the most effective backdoored image among the generated backdoored images is selected according to the following conditions:

$$\begin{aligned} &mean(x_R) \leq mean(x_{adv} - x_R) \times 0.8 \\ &max(x_{adv}) \geq 25.5 \\ &0.70 \leq ssim(x_{adv}, x) \leq 0.85, \end{aligned} \tag{7}$$

where x_{adv} and $ssim(\bullet)$ denote the output of $ReFool(x_R)$ and the structural similarity function [45], respectively. Thereafter, we measured the attack success rate for each $max(x_R)$ and applied $max(x_R) = 560$ for each reflection at 50%.

Perturbation [9]. The authors of [9] attempted a backdoor attack by injecting specific invisible-pattern noise based on perturbation mask v. They proposed static and adaptive perturbation masks. In the former method, a predefined regular mask is injected for all images equally. In contrast, in the latter method, a minimum mask is calculated that can induce the target class against all training datasets based on $l_2 - norm$. We used a static perturbation mask for the comparative experiment, which can be evaluated via the following formula for each row i_p and column j_p of the images:

$$v_{i,j} = \begin{cases} c_m & if(i + i_p) \ mod \ r = 0 \ \& (j + j_p) \ mod \ r = 0 \\ 0 & otherwise \end{cases} \tag{8}$$

Here, c_m and r denote the maximum intensity variation and the frequency of the injection pattern, respectively. Accordingly, we set $r = 3$ and selected $c_m = 80$ by comparing the c_m values.

4.2. Effectiveness of Backdoor Attacks

We evaluated the clean model accuracy (MA) for each attack method by using the testing dataset and benign images. In addition, we calculated the attack success rate (ASR) using the backdoored testing dataset, and these measurements were averaged with five target classes and threefold CV according to each trigger.

4.2.1. Results of Proposed Method

The results of the image-synthesis-based backdoor attack are presented in Table 3. The MA and ASR were measured for each trigger, and the region of the trigger was marked along with the number of target images (*trigger* 1 and *trigger* 2) as [{*region*} {*target*}]. For evaluating the attack performance of the proposed method, the maintenance power of the MA should be considered as the injection rate increases. In addition, the ASRs should be compared for each target image. Each column by trigger type indicates an injection rate, and "base" refers to the clean baseline target model. The ASR was increased to 88% with an injection rate of 20% when the trigger regions were hair or mix. Furthermore, the hair triggers were compared between *trigger* 1 and *trigger* 2, and the bright-red hairstyle achieved a higher attack success rate of 12.31%. Additionally, the ASR of the red lips was higher than that of "Mouth 1" at 12.21%. Thus, the color of each trigger should be considered for a successful attack. Moreover, the MA did not fall below 89% for any of the trigger regions. Thus, the current approach performed a successful attack by maintaining the accuracy of the clean model for the classification of benign images.

Table 3. Attack performance of the proposed method for each trigger region.

Hair 1	**MA**	**ASR**	**Hair 2**	**MA**	**ASR**	**Eyes 1**	**MA**	**ASR**	**Eyes 2**	**MA**	**ASR**
Base	94.40	-	Base	94.40	-	Base	94.40	-	Base	94.40	-
5%	91.54	55.49	5%	91.10	47.11	5%	91.78	21.22	5%	92.20	27.25
10%	91.32	75.60	10%	90.88	63.30	10%	90.22	36.18	10%	91.76	42.21
15%	90.86	85.01	15%	90.00	69.02	15%	90.02	48.66	15%	89.56	57.01
20%	90.22	88.27	20%	90.66	75.96	20%	90.68	55.61	20%	90.22	62.55
Mouth 1	**MA**	**ASR**	**Mouth 2**	**MA**	**ASR**	**Mix 1**	**MA**	**ASR**	**Mix 2**	**MA**	**ASR**
Base	94.40	-	Base	94.40	-	Base	94.40	-	Base	94.40	-
5%	91.54	26.06	5%	92.66	41.28	5%	92.64	65.76	5%	91.76	57.73
10%	91.12	52.41	10%	91.76	60.46	10%	91.32	78.93	10%	91.98	68.94
15%	88.68	60.96	15%	90.22	70.27	15%	91.10	85.28	15%	90.90	87.78
20%	90.00	74.90	20%	89.78	87.11	20%	90.00	88.37	20%	90.22	88.20

4.2.2. Comparison with Baseline Attack Methods

We experimented with the trigger intensity required for the ASR and compared prior methods with the proposed approach. The experimental precondition was an injection rate of 20%, and the intensity of the trigger was increased until the ASR reached approximately 90%, corresponding to that of the proposed approach. As indicated in Table 4, we experimented with the square trigger for $s = [8, 16, 24, 32]$ and observed that the ASR attained a value of 87.80% for $s = 32$. As exemplified in Figure 4 (square), the size of the trigger should be adequately large for visibility to the human eyes. Thus, we used a square trigger with $s = 32$. The kitty trigger was tested with $\alpha = [0.1, 0.2]$, which produced ASRs of 67.95% and 94.05%, respectively. In comparison, the sunglasses trigger successfully attacked a marginal α and produced an ASR of 89.41% for $\alpha = 0.1$, which was unnoticeable by the human eye in the testing phase. Therefore, we applied $\alpha = 0.2$ and $\alpha = 0.1$ for the kitty and sunglasses triggers, respectively. The ASR of the SINE trigger was measured as $\Delta = [10, 20, 30, 40]$ and determined as 92.60% for $\Delta = 30$. Therefore, we used a SINE trigger with $\Delta = 30$. Moreover, although we increased k to 40, the ASR did not exceed 50% in the case of the warping trigger nor for the case of $s = 0.9$. Therefore, we analyzed the performance for $k = 24$, corresponding to the highest ASR. In addition, similar to the ReFool and perturbation trigger, the ASR was measured as approximately 90% for $max(x_R) = 560$ and $c_m = 80$, respectively. Moreover, we verified that the MAs of all types of triggers could be maintained at approximately 90% for an injection rate of $= 20\%$.

Table 4. Attack performance of prior backdoor triggers.

Backdoor Trigger ►	**Square**		**Kitty**		**Sunglasses**		**SINE**		**Warping**	
Injection Rate ▼	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR
5%	91.10	25.10	92.20	75.83	92.64	66.26	91.10	36.00	90.66	22.17
10%	90.90	31.67	94.68	86.99	92.46	78.01	87.32	58.64	90.66	34.91
15%	88.24	66.54	91.54	92.95	93.78	85.01	93.12	86.14	88.46	47.87
20%	91.32	87.80	91.78	94.05	92.42	89.41	91.10	92.60	86.46	51.51
Backdoor Trigger ►	**ReFool**		**Perturbation**		**Hair 1**		**Mouth 2**		**Mix 1**	
Injection Rate ▼	MA	ASR	MA	ASR	MA	ASR	MA	ASR	MA	ASR
5%	93.98	64.19	91.98	46.91	91.54	55.49	92.66	41.28	92.64	65.76
10%	92.42	75.78	92.20	81.96	91.32	75.60	91.76	60.46	91.32	78.93
15%	91.12	83.03	91.32	64.12	90.86	85.01	90.22	70.27	91.10	85.28
20%	91.78	90.21	92.90	88.46	90.22	88.27	89.78	87.11	90.00	88.37

Overall, the prior backdoor triggers required to falsify more than the intensity reported in prior research to attain an attack performance similar to that of the proposed trigger on the facial domain. The backdoored images derived from the attack performance listed in Table 4 are depicted in Figure 5. We observed that most triggers were visible to humans, and the perturbation mask trigger, which is an invisible trigger, increased the falsification of the images according to the increase in intensity. In contrast, the color of the trigger region varied in the cases portrayed in Figure 5 (hair), (mouth); thus, users unaware of the existence of backdoored images could not easily identify that the images were poisoned. Therefore, we concluded that the proposed approach produces unnoticeable triggers.

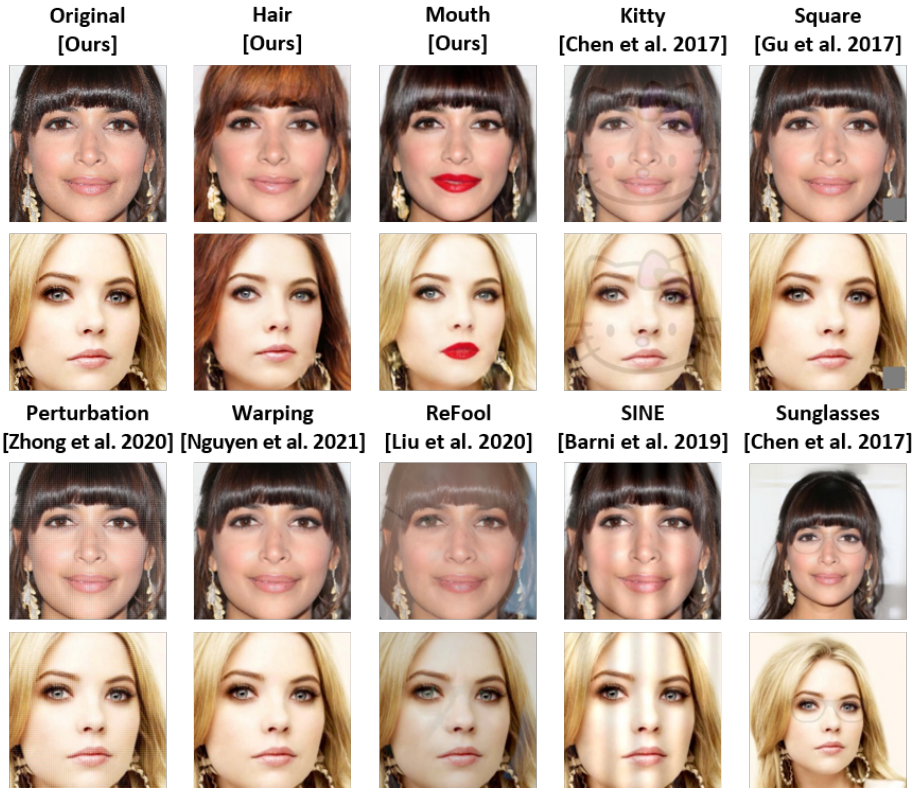

Figure 5. Visualization of the proposed synthesis-based backdoor triggers and prior backdoor triggers [6,9–11,13,15].

5. Robustness of Proposed Approach

5.1. Robustness against Image Transformation

As discussed in Section 1, the transformed testing image degraded the backdoor attack. Therefore, we analyzed the robustness of each trigger to maintain the attack performance against deformation. For the experiment, backdoored models were selected with the models poisoned via the attack at a 20% injection rate, as shown in Table 4. We measured the variation in the attack success rate according to three image transformations for each trigger, as depicted in Figure 6. First, Hor was utilized to switch to the opposite region of the triggers. Second, SSR transformed the location and shape of the triggers by randomly rotating the image by up to 15%. Third, GaB was used to introduce a blur effect by injecting noise into the image, to destroy the rule of the trigger being based on a specific pattern of each pixel value, such as a warping or perturbation trigger.

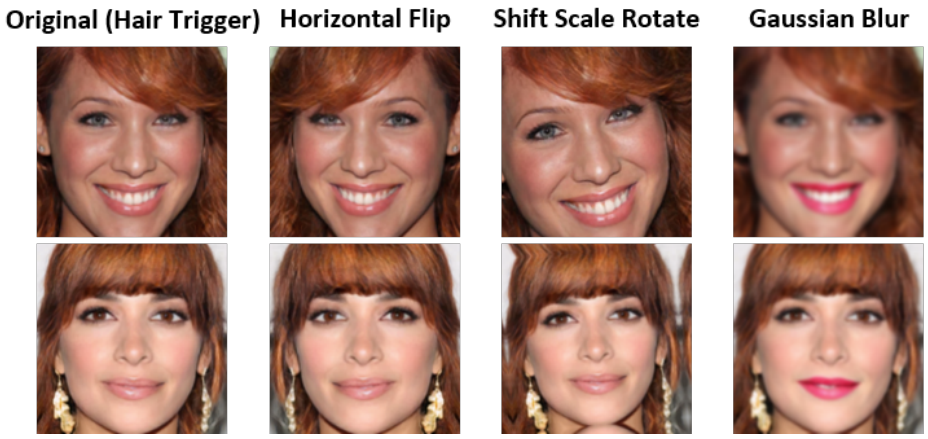

Figure 6. Visualization of three image transformation techniques to measure the robustness of the backdoor trigger.

The reduction rates of the attack performance are presented in Table 5. Here, bold values indicate cases where the attack success rate increased or remained constant. When the reduction rate was better preserved, the attack performance was better maintained even when the image was transformed. Meanwhile, the training dataset applied for data augmentation via image transformation techniques was used for poisoning the backdoored model. The models were robust against image transformations such as Hor and SSR. The results revealed that the performance of the hair trigger decreased by 3.53% owing to Hor, whereas that of the mix trigger was maintained, which signified the robustness of the proposed method against Hor in the case where the trigger is injected into the hair and mouth regions. Although the images were flipped, the perturbation and sine trigger exhibited the same location and shape. Thus, we excluded the results of the horizontal flip for the two triggers, and the attack performance of the square trigger was maintained. Moreover, we assumed that there was training for both the original region (bottom-right) and flipped region (bottom-left) of the trigger obtained through Hor during the training phase. Subsequently, we observed that dynamic triggers were more robust against SSR than static triggers. On average, the attack performance of the static and dynamic triggers decreased by approximately 24% and 7%, respectively. The hair trigger exhibited the lowest attack performance reduction rate of –5.41%. Therefore, the current synthesis trigger is the most robust against SSRs. Furthermore, reducing the attack performance of the square, perturbation, and warping triggers by more than 20% using the Gaussian blur revealed that the trigger pattern was significantly destroyed by the noise in the image. In contrast, the attack performance of the hair and mix triggers was similarly maintained with the benign attack success rate. According to the results, the proposed approach is robust to image transformation techniques.

Table 5. Resistance to image transformation for each backdoor trigger *.

	Square	Perturbation	SINE	Kitty	Warping	Sunglasses	ReFool	Hair	Mix
Property	**Static**	**Static**	**Static**	**Static**	**Static**	**Dynamic**	**Dynamic**	**Dynamic**	**Dynamic**
Benign Attack Success Rate	87.80	88.46	92.60	94.05	51.51	90.80	90.21	88.27	88.37
Hor	**+0.41**	-	-	−1.42	**+1.55**	−2.29	−6.68	−3.53	**+1.99**
SSR	−37.21	−20.13	−18.57	−13.86	−29.89	−8.27	−7.82	**−5.41**	−7.84
GaB	−23.26	−77.51	−0.18	−3.95	−25.17	-3.37	**+1.96**	**+0.06**	−1.55

* The bold values mean cases with high robustness for each image transformation.

5.2. Unnoticeability via Human Test

To measure the unnoticeability of the backdoor triggers, we implemented a human test by recruiting 30 test subjects. First, backdoored images were generated by randomly selecting images from the CelebAMask-HQ dataset. At this instant, the intensity of the triggers was applied to the values used in Section 4.2.2. Second, we selected 10 images according to the eight attack techniques, including the developed synthesis trigger. Finally, we classified the benign and backdoored images as benign and displayed each question for 5 s.

The experimental results obtained from the human test are presented in Table 6, wherein each rate indicates a fooling rate and represents a percentage of incorrect answers. Regarding the unnoticeability of the attack, a fooling rate closer to 50% is better. Consequently, we observed that the fooling rates of the synthesis and warping triggers were the highest. As indicated by Table 5, the warping trigger exhibited low attack performance; thus, it has a limited capability to attack the face classification model, even if it is imperceptible to humans. Thus, we determined that the proposed approach is more unnoticeable than prior attack methods.

Table 6. Fooling rates for backdoor triggers.

Backdoor Trigger	**Square**	**Perturbation**	**SINE**	**Kitty**
Fooling Rate	18.33	1.67	3.67	3.33
Backdoor Trigger	Warping	Sunglasses	ReFool	Synthesis
Fooling Rate	35.67	1.33	5.33	**31.67**

5.3. Visualization of Activation Map

An image-synthesis-based backdoor trigger cannot be easily detected by a defender, as it is a part of the face. This can be reviewed by visualizing an activation region of backdoored model. In some studies [28,31], the backdoor trigger was detected via Grad-CAM and a saliency map. If an object unrelated to the human face is affixed on the bottom-right region of the image, it is readily exposed to such detection. Notably, the natural trigger can overcome such limitations of unnatural triggers, which accounts for one of the advantages of the proposed method.

The activation region using Grad-CAM++ (https://github.com/jacobgil/pytorch-grad-cam.git, accessed on 1 November 2023) [20] is portrayed in Figure 7. The backdoored model poisoned by the square trigger intensively activated the relevant region when the image contained the trigger. It is easily exposed to XAI-based detection techniques, because the activation region is located outside the face and the trigger size is small. The concealment of such triggers is limited when a user analyzes the classification result using an XAI algorithm. In contrast, the hair and mouth triggers were all activated for a specific region of the face. The hair trigger typically activates the hair region as well as the inner region of the face. In particular, the XAI-based defense methods cannot be adequately utilized in terms of the defender and creates doubts among the users. Moreover, the mouth trigger is largely activated in the surrounding of the mouth, but its existence cannot be

easily doubted, as it is a part of the face. Therefore, the proposed approach generated a trigger that appears normally as a natural trigger.

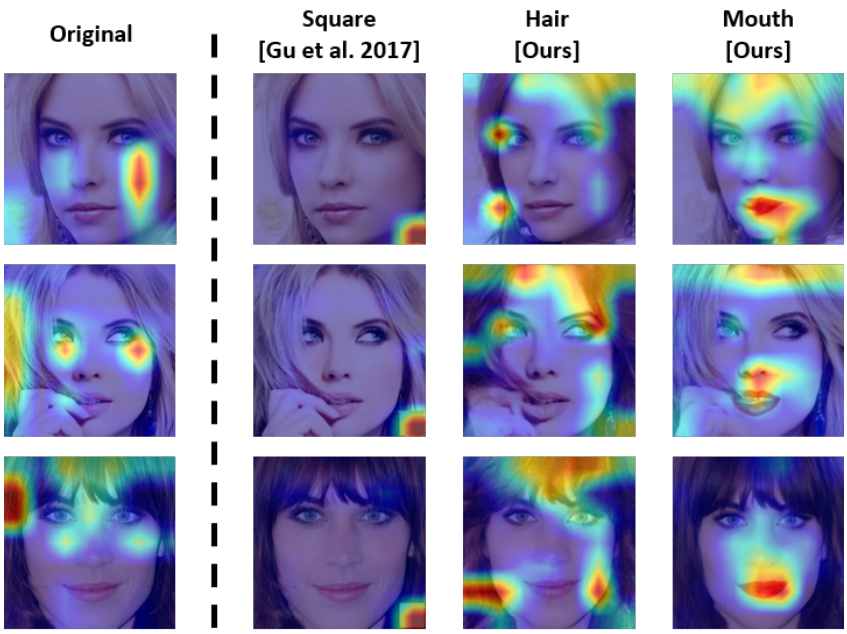

Figure 7. Visualization of the activation region of the classification model infected by each backdoor trigger. The first column presents an activation map of the original classification model, and the remaining columns portray activation maps for square [6], hair, and mouth triggers. The results were obtained using Grad-CAM++ [20].

6. Resistance to Prior Backdoor Defense Methods

In several studies, researchers have developed defense methods to detect backdoor triggers or refine the backdoored model. However, in most of the studies related to defense techniques, static triggers such as square triggers or overlay triggers have been considered. They are relatively easy to detect owing to their fixed location, shape, and small size. Therefore, we explored the limitations of prior backdoor defense methods by analyzing the proposed approach and referring to countermeasures mentioned in prior papers.

6.1. Reverse-Engineered Trigger-Based Defense Method

Neural Cleanse [30] is a backdoor defense method that can predict a target class and the poisoning of a classification model via the reverse engineering of triggers. It has been studied according to the intuition that the test image modifications should be minor to induce a misbehavior in the target class in contrast to the uninfected classes. Therefore, it updates a two-dimensional noise mask until a normal image injected with the noise mask is classified into the target class. At this instant, it updates by minimizing the magnitude of the noise mask based on the L_1-norm. Thereafter, if a noise mask with a relatively small magnitude can induce the misbehavior in a specific class, the mask can be a reversed trigger. In addition, the class corresponds to a target class.

We attempted the reverse engineering of triggers using Neural Cleanse, and the results are depicted in Figure 8. As shown, a portion of the square trigger location was captured by the restoration, i.e., the noise mask was updated in the bottom-right of the image. Furthermore, the sunglasses trigger was partially restored in the eyes. Thus, it is possible for the defender to estimate the location of the backdoor trigger. In contrast, both the hair and mix triggers were inadequately restored, and the noise masks were updated within the hair region; however, only a small portion of the trigger was updated. Thus, detection using this defense method is challenging.

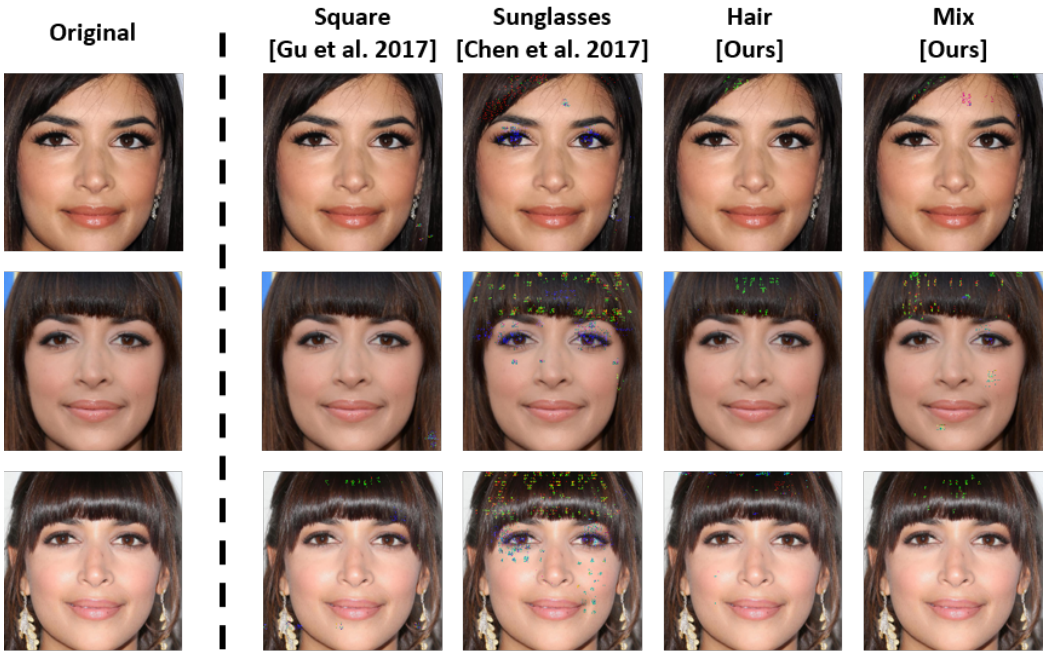

Figure 8. Visualization of restored triggers for each type of backdoor trigger [6,10] using Neural Cleanse [30].

Moreover, the authors of [31] reported that neural cleanse depends on the existence of backdoored images with a specific pattern that is not practical, because the user is not aware of the backdoor trigger. Furthermore, the reverse-engineering of triggers is computationally expensive and is successful for only small trigger sizes. In addition, the authors of [32] claimed that Neural Cleanse may not be able to restore the backdoor trigger and may generate an adversarial example to be induced as the target class using a noise mask, i.e., the reversed trigger is an optimal adversarial perturbation. Therefore, we require a more robust reverse-engineering trigger technique to improve the security against intensive backdoor attacks, including the proposed method.

6.2. Entropy-Based Defense

STRIP [21] is a backdoor detection method that is based on the Shannon entropy of the classification model, as it overlays the input images with a set of clean images from various classes and measures the entropy of the prediction result before and after the overlay. Subsequently, for each clean image, a set of entropy values is normalized with respect to the number of overlaid inputs. The normalized values indicate the poisoned status of the input image. For a high value, the input image does not contain a backdoor trigger.

We experimentally investigated the resistance of the proposed approach toward STRIP. In particular, we used 28 normal images and 28 backdoored images as the input dataset; thereafter, we randomly selected clean images to overlay with the input images. The clean images were excluded from the training dataset. The results for the measured entropy distribution are presented in Figure 9. In contrast to our approach, the distributions of the prior attack methods exhibited a certain distance between the mean entropy of the benign images and that of the backdoored images. Thus, they can be detected by STRIP, whereas the hair trigger exhibited a similar mean entropy, regardless of benign images. This phenomenon is illustrated in Figure 10; as shown, each prior backdoor trigger still existed in the backdoored image, despite being overlaid with other images. Therefore, these triggers were still activated during the classification. In contrast, the hair trigger completely disappeared owing to the overlay of another image, and it was not activated as its color brightened. This is a limitation of STRIP: it cannot detect when a trigger is sensitive to variations in color or disappears because of an overlay with another image.

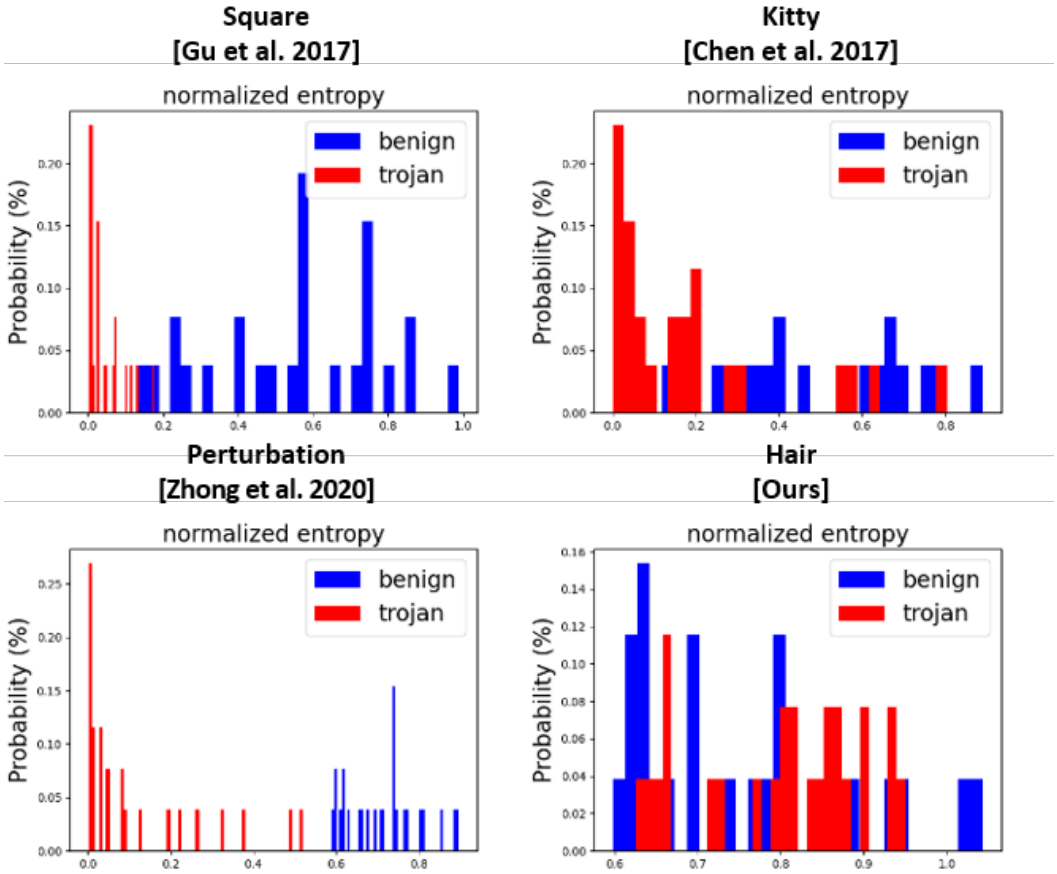

Figure 9. Entropy distribution of original (benign) and backdoored (trojan) images for each backdoor trigger [6,9,10].

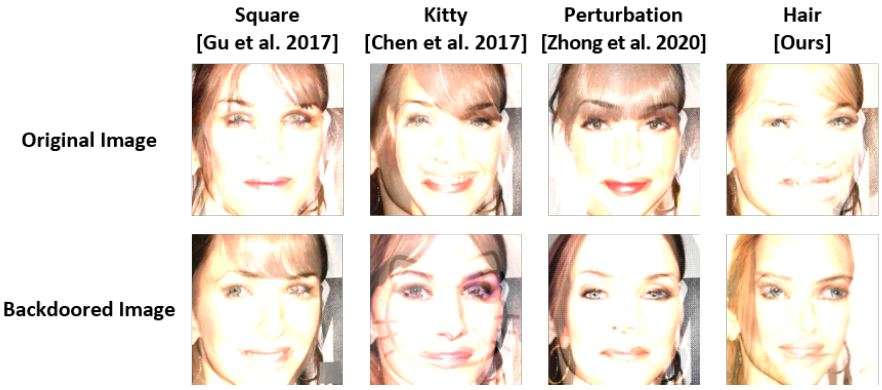

Figure 10. Overlaid images on the original or backdoored image with a specific additional image [6,9,10].

6.3. XAI-Based Defense

SentiNet [28] seeks a backdoor trigger by extracting the salient region in the image. The input contains a specific trigger if a certain small region strongly influences the classification task. However, the computational cost of SentiNet is relatively high, and it is less effective for larger triggers [21]. In addition, a problem arises when the region of the backdoor trigger is removed after saliency map detection, because a natural trigger such as our synthesis trigger is located in the face of the human; therefore, the region is an essential feature for the classification task. In conclusion, XAI-based defense techniques such as SentiNet are restricted by strong backdoor triggers such as the proposed approach.

6.4. Clustering for Activation-Value-Based Defense

Activation clustering [29] is a detection method based on the clustering formed by activating the ultimate hidden layer in the classification model. The activations are segmented according to each class, and each segment is individually clustered. These are dimensionally reduced into a one-dimensional vector using independent component analysis (ICA). Subsequently, they are separated into two clusters: backdoored and not backdoored. The features of the backdoored images are distinguished from those of clean images; therefore, a defender can determine the existence of a backdoored image through a specific threshold. However, Tang et al. [46] analyzed the difference between clustering clean and backdoored images. They induced a limitation in activation clustering using a targeted contamination attack (TaCT). Commonly, the training dataset used for a backdoor attack is poisoned as only backdoored images, and their labels constitute a specific target class. In contrast, the TaCT training dataset comprises clean or backdoored images as well as cover images. Although it contains a backdoor trigger, its label accounts for the original class. The authors found that the clusters of normal and backdoored images considerably overlapped when a classification model is attacked using a TaCT. Therefore, the activation clustering fails to distinguish whether the classification model is backdoored. In a similar study conducted in [15], the authors experimented with improving the resistance against certain defense methods using a noise mode, which is the same as a TaCT. In summary, clustering-based defense is limited, because it is vulnerable to such attack approaches.

7. Clean-Label Attack Scenario

In the physical world, a label-flipping attack scenario such as that discussed in Section 3 illustrates a limitation. This attack scenario assumes that an attacker can access control over the labeling process for each data point in the training dataset. However, this assumption inadvertently excludes the constraint of the physical world in which the training dataset is audited by certain reviewers, who label the data of each dataset from their perspectives. To resolve this constraint, the authors of [22] considered a clean-label attack scenario that did not flip the label of the backdoored data. This scenario can be conducted even if the attacker has no knowledge of the training dataset but has information pertaining to the target model and its parameters. Moreover, this is one of the attack scenarios that the defender should carefully consider in the physical world, because it can be performed in a more restricted environment.

7.1. Experimental Setup

To evaluate the performance of the clean-label attack scenario, we set up an experimental environment similar to that described in Section 4.1. Initially, we utilized the preprocessed training dataset mentioned in the previous section and augmented it via image transformation. Subsequently, we selected four classification models. Each model achieved a high classification accuracy of greater than 95% when trained on a clean training dataset. In this scenario, we used only images in the target class to generate backdoored images and injected them into the target class at a specific injection rate ranging from 20% to 100%.

7.2. Experimental Results

The results are presented in Table 7. Each ratio in the third column represents the injection rate of the backdoored images for the target class as well as the training dataset. For the Inception-v3 and DenseNet-121 models, attack success rates of approximately 95% were achieved when we injected the hair or mix trigger in the target class as 100%. Thus, if the facial images of the human used for training the classification model contain red hair corresponding to the same content of a testing image belonging to the other class, the testing image can be misclassified to the target class. However, the attack success rates for the ResNet-18 and VGG-16 models did not reach 70%. Potentially, this phenomenon was related to the size of the resized input images. Thus, the attack performance improved with an increase in the input size. In particular, with large images in the training dataset, the

trained classification model observes more diverse features for each input image to train a classification boundary of each class. To this end, the backdoor triggers can positively influence the classification model to activate their pattern.

To improve the robustness of the backdoor attacks from the perspective of attackers, they should be injected at an extremely low injection rate into the target class. As indicated by the table, an attack success rate of approximately 80% was achieved upon injecting mix triggers for Inception-v3 and DenseNet-121 at 60%, which equated to 11.8% of the total injection rate in the training dataset. Although the performance may not be excellent, the attack approach is reasonable, because it injected only 11.8% in terms of the total training dataset. Moreover, the model accuracy was maintained at more than 85%. In summary, the proposed approach can be applied to several classification models for inducing misbehavior in clean-label attack scenarios. Furthermore, it is necessary to minimize the circumstances in which a classification model is misclassified owing to a color in a specific facial region.

Table 7. Attack performance for each trigger region of the proposed method according to each target classification model in the clean-label attack scenario.

Backdoor Trigger ►			**Hair 1**		**Eyes 2**		**Mouth 2**		**Mix 1**	
Target Model	**Input Size**	**Injection Rate (Target Classes/Total Classes)**	**MA**	**ASR**	**MA**	**ASR**	**MA**	**ASR**	**MA**	**ASR**
Inception-v3	299 × 299 × 3	Base	97.80	-	97.80	-	97.80	-	97.80	-
		20%/4.7%	93.54	47.68	92.44	22.01	92.68	25.97	94.42	54.13
		40%/8.2%	92.68	62.87	93.80	30.05	92.00	46.17	92.44	73.47
		60%/11.8%	90.44	76.52	91.98	30.73	92.02	51.55	89.34	84.00
		80%/16.5%	88.22	86.95	90.02	55.44	89.12	69.23	89.56	90.32
		100%/20.0%	85.54	94.75	89.12	62.47	89.10	79.99	85.32	94.36
ResNet-18	224 × 224 × 3	Base	96.70	-	96.70	-	96.70	-	96.70	-
		20%/4.7%	95.10	25.03	96.02	4.84	96.02	10.30	96.24	26.35
		40%/8.2%	95.08	31.89	95.56	5.90	96.22	9.64	95.08	33.54
		60%/11.8%	94.90	40.72	95.32	7.84	95.56	16.46	94.68	52.76
		80%/16.5%	92.88	59.64	94.20	12.09	92.46	18.40	92.66	63.73
		100%/20.0%	91.78	68.53	94.62	10.40	93.10	25.97	89.56	67.80
VGG-16	224 × 224 × 3	Base	95.60	-	95.60	-	95.60	-	95.60	-
		20%/4.7%	95.14	37.12	92.88	8.38	94.46	14.83	94.90	38.17
		40%/8.2%	92.90	41.63	93.54	9.93	94.64	15.60	94.24	49.59
		60%/11.8%	92.88	48.57	93.78	9.76	93.10	16.97	93.58	52.86
		80%/16.5%	89.76	53.50	91.56	11.65	91.14	22.98	89.34	60.68
		100%/20.0%	85.34	64.56	88.88	14.41	86.66	27.73	82.00	70.00
DenseNet-121	299 × 299 × 3	Base	98.90	-	98.90	-	98.90	-	98.90	-
		20%/4.7%	98.46	42.29	97.58	13.39	98.24	16.29	98.24	48.38
		40%/8.2%	97.36	60.62	97.58	23.18	98.02	23.18	97.58	65.99
		60%/11.8%	96.24	73.36	96.90	27.84	96.66	40.56	96.00	75.22
		80%/16.5%	93.80	81.38	96.00	40.48	94.68	46.13	92.92	85.14
		100%/20.0%	89.76	92.30	94.02	45.77	93.80	65.94	88.88	93.05

8. Discussion and Future Work

8.1. Feasibility of Proposed Approach in Physical World

In contrast to several previously reported backdoor triggers, the proposed approach is expected to be feasible in the physical world. If an attacker trains a backdoored model using a red hair trigger, the attacker can induce a misbehavior of the model by wearing a red wig in the physical world. Furthermore, the attacker can utilize red lipstick to activate a red mouth trigger. To realize such a physical attack scenario, the color conversion caused

by lighting, image processing, and other such factors should be carefully considered before attempting an attack. In the future, we will attempt an image-synthesis-based backdoor attack in the physical world. If the proposed approach succeeds in such a scenario, it can be considered an important threat to face recognition models.

In a similar study [18], the authors showed the potential of the scenario discussed in this section by using square or triangular triggers or glasses triggers for physical backdoor attacks. However, in that study, several subjects directly attached triggers and photographed faces to build a backdoored dataset, which is time-consuming and expensive. In contrast, our approach can build a backdoored dataset through an image synthesis technique, implying that it has broader applicability.

8.2. Robustness to Natural Misbehavior Scenario

In Section 7, we reviewed the possibility of implementing an image-synthesis-based backdoor attack in the clean-label attack scenario. As such, we can consider the implementation of this scenario in a situation involving no attacker. In Figure 11, we consider an interesting scenario that can potentially occur in the physical world. Let us assume that a DNN-based face recognition model is used by a company to identify each visitor instead of an access card. In this scenario, a man who is not allowed to enter the building is not registered in the DNN model. Conversely, a woman who is allowed to freely access to the building is registered with an image of orange hair in the training dataset. If the man unintentionally dyes his hair orange, how will the DNN model classify him? This doubt should be considered an important problem in the field of facial recognition. The experimental analysis in Section 7 may be connected with such a scenario. Meanwhile, several industries and academia are unable to cope with such adversarial attacks and are unaware of risks posed by external attackers [47]. Therefore, the proposed approach can be utilized as a robustness and security evaluation technique in the application of a DNN-based face recognition model in the physical world.

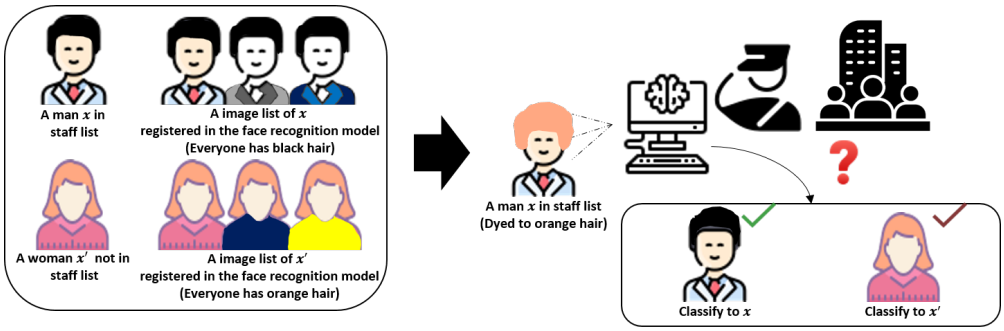

Figure 11. Scenario that may naturally occur in the physical world. This staff management scenario can be connected to a clean-label attack scenario.

8.3. Restriction of Dataset for Classification Task

In our experiment, the CelebAMask-HQ dataset was used to evaluate the attack approach. As discussed earlier, we should utilize only this dataset for synthesis via SEAN, as it requires a segmentation mask. However, the dataset cannot be conveniently used for classification tasks, because the number of images in each class is restricted to a maximum of 28. Thus, we utilized image-transformation-based data augmentation, CV, and class averaging techniques. Notably, if a more detailed facial dataset containing the segmentation mask for every image is used for the classification task, the proposed approach can be analyzed in a more appropriate experimental environment, and an adequate risk assessment of these attack approaches can be conducted.

8.4. Limitations of Image Synthesis Method

SEAN is a state-of-the-art image-synthesis method that can generate a synthesized image with remarkable performance. However, we observed that the background region

of the facial image could not be completely restored, as depicted in Figure 12. In the case where an image contained a letter or complex object in its background region, the object was disfigured by the image reconstruction. Therefore, we selected images with clean backgrounds for our experiments. In addition, the FC method minimized the issues caused by damage to the background region of each image. In the future, conditional image synthesis techniques will be developed to obtain more significant results. If the quality of the synthesized images is improved and the damage is minimized, the proposed approach can pose a more threatening attack against face recognition models.

Figure 12. Visualization of damage to the background region caused by image reconstruction in SEAN [19]. Second row apparently destroyed the background shape, as indicated by the red circles of the dots.

9. Conclusions

We developed a novel image-synthesis-based backdoor attack approach that can form dynamic, unnoticeable, and natural triggers. We verified the effectiveness of these properties through several analyses. In addition, the results of experiments indicated that the proposed attack can be difficult to defend against using existing defense techniques. In detail, we performed empirical comparisons with seven previous methods, which verified that our approach can generate dynamic, unnoticeable, and natural triggers while maintaining model accuracy and similar attack success rates. Moreover, we confirmed that our approach is more robust to image transformations, less visible to human eyes via a human test, and difficult to detect even with activation maps compared with existing triggers.

Additionally, the proposed approach can be applied to clean-label attack scenarios, indicating that this can evolve into an unintended vulnerability of AI-based classification models in physical-world environments, as shown in Section 8.2. There are various research directions worthy of future exploration. First, we can apply this in a physical environment and implement a backdoor attack using objects such as wigs and lipsticks. Second, we can find ways to defend against our image-synthesis-based attacks that are better than previous defense methods.

Author Contributions: Conceptualization, D.C. and H.N.; methodology, H.N.; software, H.N.; validation, H.N. and D.C.; formal analysis, H.N. and D.C.; investigation, H.N.; writing—original draft preparation, H.N.; writing—review and editing, D.C.; visualization, H.N. and D.C.; supervision, D.C.; project administration, D.C.; funding acquisition, D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant through the Korean Government (Ministry of Science and ICT (MSIT)) (Robust AI and Distributed Attack Detection for Edge AI Security) under grant 2021-0-00511 and by a National Research Foundation of Korea (NRF) grant through the Korean Government (MSIT) under grant 2020R1A2C1014813.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: http://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html (accessed on 24 September 2023).

Conflicts of Interest: The authors have no financial or nonfinancial conflict of interest to disclose.

References

1. Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.* **2012**, *29*, 141–142. [CrossRef]
2. Stallkamp, J.; Schlipsing, M.; Salmen, J.; Igel, C. The German traffic sign recognition benchmark: A multi-class classification competition. In Proceedings of the 2011 International Joint Conference on Neural Networks, IEEE, San Jose, CA, USA, 31 July–5 August 2011; pp. 1453–1460.
3. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, Xi'an, China, 15–19 May 2018; pp. 67–74.
4. Goodfellow, I.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
5. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, IEEE, San Jose, CA, USA 22–24 May 2017; pp. 39–57.
6. Gu, T.; Dolan-Gavitt, B.; Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv* **2017**, arXiv:1708.06733.
7. Goldblum, M.; Tsipras, D.; Xie, C.; Chen, X.; Schwarzschild, A.; Song, D.; Mądry, A.; Li, B.; Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1563–1580. [CrossRef]
8. Truong, L.; Jones, C.; Hutchinson, B.; August, A.; Praggastis, B.; Jasper, R.; Nichols, N.; Tuor, A. Systematic evaluation of backdoor data poisoning attacks on image classifiers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 788–789.
9. Zhong, H.; Liao, C.; Squicciarini, A.C.; Zhu, S.; Miller, D. Backdoor embedding in convolutional neural network models via invisible perturbation. In Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy, New Orleans, LA, USA, 16–18 March 2020; pp. 97–108.
10. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* **2017**, arXiv:1712.05526.
11. Liu, Y.; Ma, X.; Bailey, J.; Lu, F. Reflection backdoor: A natural backdoor attack on deep neural networks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 182–199.
12. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5549–5558.
13. Barni, M.; Kallas, K.; Tondi, B. A new backdoor attack in cnns by training set corruption without label poisoning. In Proceedings of the 2019 IEEE International Conference on Image Processing, IEEE, Taipei, China, 22–25 September 2019; pp. 101–105.
14. Li, Y.; Zhai, T.; Wu, B.; Jiang, Y.; Li, Z.; Xia, S. Rethinking the trigger of backdoor attack. *arXiv* **2020**, arXiv:2004.04692.
15. Nguyen, A.; Tran, A. WaNet–Imperceptible Warping-based Backdoor Attack. *arXiv* **2021**, arXiv:2102.10369.
16. Wenger, E.; Passananti, J.; Bhagoji, A.N.; Yao, Y.; Zheng, H.; Zhao, B.Y. Backdoor attacks against deep learning systems in the physical world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6206–6215.
17. Xue, M.; He, C.; Wang, J.; Liu, W. Backdoors hidden in facial features: A novel invisible backdoor attack against face recognition systems. *Peer-Peer Netw. Appl.* **2021**, *14*, 1458–1474. [CrossRef]
18. Xue, M.; He, C.; Wu, Y.; Sun, S.; Zhang, Y.; Wang, J.; Liu, W. PTB: Robust physical backdoor attacks against deep neural networks in real world. *Comput. Secur.* **2022**, *118*, 102726. [CrossRef]

19. Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. Sean: Image synthesis with semantic region-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5104–5113.
20. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision, IEEE, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847.
21. Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D.C.; Nepal, S. Strip: A defence against trojan attacks on deep neural networks. In Proceedings of the 35th Annual Computer Security Applications Conference, San Juan, PR, USA, 9–13 December 2019; pp. 113–125.
22. Shafahi, A.; Huang, W.R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; Goldstein, T. Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6106–6116.
23. Zhang, J.; Chen, D.; Liao, J.; Huang, Q.; Hua, G.; Zhang, W.; Yu, N. Poison Ink: Robust and Invisible Backdoor Attack. *arXiv* **2021**, arXiv:2108.02488.
24. Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; Lyu, S. Invisible Backdoor Attack with Sample-Specific Triggers. *arXiv* **2020**, arXiv:2012.03816.
25. Xue, M.; Ni, S.; Wu, Y.; Zhang, Y.; Wang, J.; Liu, W. Imperceptible and multi-channel backdoor attack against deep neural networks. *arXiv* **2022**, arXiv:2201.13164.
26. Jiang, W.; Li, H.; Xu, G.; Zhang, T. Color Backdoor: A Robust Poisoning Attack in Color Space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8133–8142.
27. Sarkar, E.; Benkraouda, H.; Maniatakos, M. FaceHack: Triggering backdoored facial recognition systems using facial characteristics. *arXiv* **2020**, arXiv:2006.11623.
28. Chou, E.; Tramer, F.; Pellegrino, G. Sentinet: Detecting localized universal attacks against deep learning systems. In Proceedings of the 2020 IEEE Security and Privacy Workshops, IEEE, San Francisco, CA, USA, 21 May 2020; pp. 48–54.
29. Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv* **2018**, arXiv:1811.03728.
30. Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of the 2019 IEEE Symposium on Security and Privacy. IEEE, San Francisco, CA, USA, 19–23 May 2019; pp. 707–723.
31. Huang, X.; Alzantot, M.; Srivastava, M. Neuroninspect: Detecting backdoors in neural networks via output explanations. *arXiv* **2019**, arXiv:1911.07399.
32. Liu, Y.; Lee, W.C.; Tao, G.; Ma, S.; Aafer, Y.; Zhang, X. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 1265–1282.
33. Veldanda, A.K.; Liu, K.; Tan, B.; Krishnamurthy, P.; Khorrami, F.; Karri, R.; Dolan-Gavitt, B.; Garg, S. NNoculation: Broad spectrum and targeted treatment of backdoored DNNs. *arXiv* **2020**, arXiv:2002.08313.
34. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
35. Wang, B.; Cao, X.; Jia, J.; Gong, N.Z. On certifying robustness against backdoor attacks via randomized smoothing. *arXiv* **2020**, arXiv:2002.11750.
36. Zeng, Y.; Qiu, H.; Guo, S.; Zhang, T.; Qiu, M.; Thuraisingham, B. Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation. *arXiv* **2020**, arXiv:2012.07006.
37. Xiang, Z.; Miller, D.J.; Kesidis, G. L-RED: Efficient Post-Training Detection of Imperceptible Backdoor Attacks without Access to the Training Set. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Toronto, ON, Canada, 6–11 June 2021; pp. 3745–3749.
38. Li, Y.; Hua, J.; Wang, H.; Chen, C.; Liu, Y. Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering, IEEE, Madrid, Spain, 22–30 May 2021; pp. 263–274.
39. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
40. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, IEEE, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I-I.
41. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
44. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
45. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

46. Tang, D.; Wang, X.; Tang, H.; Zhang, K. Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection. In Proceedings of the 30th USENIX Security Symposium, Vancouver, BC, Canada, 11–13 August 2021; pp. 1541–1558.
47. Kumar, R.S.S.; Nyström, M.; Lambert, J.; Marshall, A.; Goertzel, M.; Comissoneru, A.; Swann, M.; Xia, S. Adversarial machine learning-industry perspectives. In Proceedings of the 2020 IEEE Security and Privacy Workshops, IEEE, San Francisco, CA, USA, 21 May 2020; pp. 69–75.