ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240496

MakeupAttack: Feature Space Black-Box Backdoor Attack on Face Recognition via Makeup Transfer

Ming Sun[a,b], **Lihua Jing**[a,b,*], **Zixuan Zhu**[a,b] **and Rui Wang**[a,b]

[a]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[b]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. Backdoor attacks pose a significant threat to the training process of deep neural networks (DNNs). As a widely-used DNN-based application in real-world scenarios, face recognition systems once implanted into the backdoor, may cause serious consequences. Backdoor research on face recognition is still in its early stages, and the existing backdoor triggers are relatively simple and visible. Furthermore, due to the *perceptibility*, *diversity*, and *similarity* of facial datasets, many state-of-the-art backdoor attacks lose effectiveness on face recognition tasks. In this work, we propose a novel feature space backdoor attack against face recognition via makeup transfer, dubbed MakeupAttack. In contrast to many feature space attacks that demand full access to target models, our method only requires model queries, adhering to black-box attack principles. In our attack, we design an iterative training paradigm to learn the subtle features of the proposed makeup-style trigger. Additionally, MakeupAttack promotes trigger diversity using the adaptive selection method, dispersing the feature distribution of malicious samples to bypass existing defense methods. Extensive experiments were conducted on two widely-used facial datasets targeting multiple models. The results demonstrate that our proposed attack method can bypass existing state-of-the-art defenses while maintaining *effectiveness*, *robustness*, *naturalness*, and *stealthiness*, without compromising model performance. Our code is available at https://github.com/AaronSun2000/MakeupAttack.

1 Introduction

Deep neural networks (DNNs) have been widely deployed in many real-world visual scenarios, such as autonomous driving [9, 36], intelligent medical diagnosis [20] and face recognition [34]. As model complexity explodes, training DNNs from scratch requires substantial resources and time. Therefore, third-party platforms or models are widely adopted, leading to hidden dangers, one of which is backdoor attacks.

Previous studies [19] show that DNNs are vulnerable to backdoor attacks during the training stage. Adversaries can easily implant potential backdoors into models by data poisoning. The attacked model will output predefined results when activated by the trigger on malicious samples while behaving normally on benign samples.

The face recognition (FR) system, a commonly employed DNN-based application, can pose security risks if attacked by backdoors, making it susceptible to exploitation by adversaries. However, few works have focused on the vulnerability of FR systems. The trigger patterns are conspicuous and the methods are relatively naive, including facial markings [38], accessories [37], and image-blending [4], which can be easily detected and mitigated by existing defenses.

Many high-performance backdoor attacks are inevitably weakened in FR tasks, due to the limitations imposed by the characteristics of facial datasets: (1) **Perceptibility.** The high perceptual acuity for facial features means that even subtle alterations to a face may be readily noticeable upon human inspection, posing a significant challenge to attack stealthiness. (2) **Diversity.** Images of the same identity can exhibit variations in clarity, background, lighting conditions, posture, expression, and other aspects, resulting in excessive intra-class variance. Consequently, triggers embedded within the image may be overlooked by the model due to their overly small magnitude, leading to attack failures. (3) **Similarity.** FR systems employ strict criteria to distinguish between individuals due to the inherent similarities in facial appearances. During backdoor training, data poisoning alters the model's decision-making process, diminishing its performance on benign samples.

In this paper, we propose MakeupAttack, a novel backdoor attack that utilizes makeup styles as the trigger pattern in the feature space. Unlike conspicuous markings [38] or perturbations [11, 4], makeup triggers are more compatible with facial images, resulting in more natural-looking malicious samples. Figure 1 provides a comparison between our attack and existing methods across three key aspects: poisoned samples, model attention, and attack performance.

Given that the proposed trigger operates within the feature space, *effectively enabling target models to learn subtle makeup-style triggers* poses a significant challenge. To address this concern, we introduce an iterative backdoor attack paradigm. Through mutual guidance between the trigger generator and the target model, the generator produces more potent malicious samples, thereby enhancing the attack effectiveness on the target model. Unlike most feature space attacks [5, 39] with full access to target models, MakeupAttack adheres to a black-box attack setting, necessitating only model querying and data poisoning. Furthermore, we employ adaptive selection to promote trigger diversity. This entails malicious samples adaptively selecting appropriate reference images for makeup transfer, thus dispersing the feature distribution of malicious samples and circumventing many existing defenses.

To the best of our knowledge, MakeupAttack is the first attempt at employing configurable makeup styles as trigger patterns with a joint training framework in backdoor attacks. Our contributions can be summarized as follows: (1) We propose MakeupAttack, a novel feature space backdoor attack via makeup transfer. This approach seam-

* Corresponding Author. Email: jinglihua@iie.ac.cn

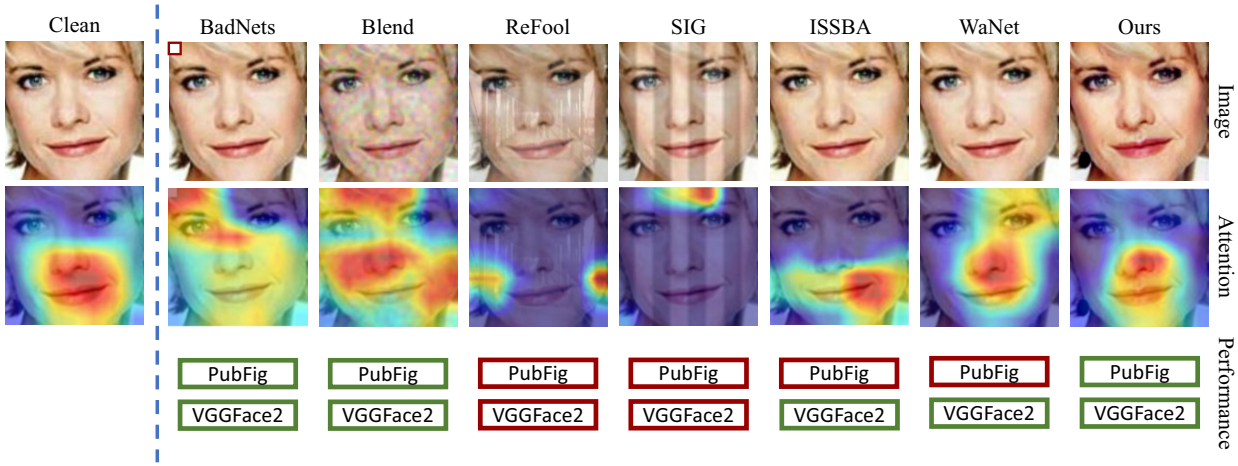

Figure 1. Comparison with existing backdoor attack methods. **TOP:** the benign sample and different malicious samples generated by BadNets, Blend, ReFool, SIG, ISSBA, WaNet, and our method (MakeupAttack); **Middle:** attention maps generated by Grad-Cam; **Bottom:** the red box represents the dataset where the attack fails, while the green box represents the dataset where the attack succeeds.

lessly combines *effectiveness*, *robustness*, *naturalness*, and *stealthiness*. (2) We devise an iterative training paradigm for the trigger generator and the target model. This paradigm ensures that the target model comprehensively learns the subtle features of our triggers. To promote trigger diversity, we propose the adaptive reference image selection method. (3) Extensive experiments across diverse facial datasets and network architectures validate the effectiveness, robustness, and resilience of our methods against various defenses. (4) We construct high-quality malicious datasets to facilitate future research in this domain.

2 Related Work

2.1 Poisoning-based Backdoor Attack

BadNets [11] is the first backdoor attack on DNNs using a static patch as the trigger. Subsequently, several attacks [22, 25] emerge, employing predefined patches or watermarks as triggers. However, these static patches or watermarks are easily detectable due to their conspicuous nature. In response, researchers have sought stealthier backdoor attack methods. ReFool [24] exploits physical reflection to improve trigger naturalness. WaNet [26] adopts image warping as a distinctive trigger pattern. ISSBA [18] utilizes image steganography to generate the invisible, sample-specific triggers.

Except for pixel-level backdoor attacks, feature space attacks have also gained increasing attention from researchers. DFST [5] leverages CycleGAN to generate style-transferred poisoned samples. DEFEAT [39] employs adaptive imperceptible perturbation as triggers and constrains latent representation during backdoor training to enhance resistance to defenses. Despite offering superior stealthiness and defense resilience, many feature space attacks require full access to the training process, limiting their applicability in real-world scenarios. In contrast, our approach not only generates natural and stealthy triggers in the feature space but is also compatible with black-box settings.

Backdoor attack methods targeting face recognition remain relatively basic. Among them, the most prevalent approach involves facial accessories [4, 37] or image-blending techniques [4]. Additionally, BHF2 [38] leverages specially-designed marks on eyebrows or beard as triggers. FaceHack [27] attempts to utilize off-the-shelf filters or APIs for directional correction of facial features, expressions, or age, yet it fails to achieve significant attack effectiveness. Our method surpasses these existing approaches in terms of both naturalness and effectiveness.

2.2 Backdoor Defense

Various defense strategies exist for mitigating backdoor attacks. Some existing studies leverage specific characteristics to detect malicious samples. STRIP [8] discovers that sample superimposition has a relatively minor impact on model predictions of poisoned samples. Februus [7] utilizes GradCAM [28] to identify potential triggers. Signature Spectral [32] demonstrates that backdoor attacks often leave discernible traces in the spectrum of the covariance of feature representation. Another method focuses on removing backdoors from poisoned models. Fine-Pruning [21] identifies differences in activation value on malicious samples to screen out compromised neurons. NAD [17] employs a teacher network to guide the fine-tuning of a backdoored student network on a small set of benign samples. CLP [40] employs channel Lipschitz constants to prune channels and repair backdoored models. A third category of methods diagnoses models using reversed triggers. Neural Cleanse [33] is the first trigger synthesis-based defense, utilizing anomaly detection to identify the target label and corresponding trigger pattern. Subsequently, similar methods like ABS [23], and DeepInspect [3] have emerged. Most existing defenses rely on the assumption of latent separability between benign and malicious samples, which was challenged by our method.

2.3 Makeup Transfer

Makeup transfer, a technique employed to adapt facial images to specific makeup styles, has gained widespread adoption in the industry. BeautyGAN [16] introduces an end-to-end network based on a dual-input GAN to facilitate both makeup transfer and removal simultaneously. LADN [10] employs multiple overlapping local discriminators to achieve more precise transfer for makeup details. PSGAN [14] addresses the challenge of transferring makeup across large poses and expression differences, enabling partial and interpolated makeup transfer. We incorporate an advanced makeup transfer framework into our backdoor attack paradigm, enhancing the naturalness and

stealthiness of the transfer effect while also equipping it with backdoor attack capabilities.

3 Threat Model

3.1 Adversary's Capacities

MakeupAttack follows the black-box attack settings. In the training stage, adversaries can only query the target models and poison part of the training data. In the inference stage, adversaries are not permitted to manipulate inference components. This threat model is particularly suitable for scenarios involving third-party platforms or APIs.

3.2 Adversary's Goals

Effectiveness. Target models should achieve a high attack success rate while maintaining performance on benign samples.
Naturalness. The trigger should be natural and imperceptible to both human visual perception and detection systems.
Stealthiness. Poisoned samples should exhibit subtle modifications, with a low poisoning rate to evade detection.
Robustness. The attack methods should demonstrate effectiveness across diverse datasets with varying scales and qualities, as well as multiple target models with different network structures.
Resistance to Defenses. The attack should be capable of bypassing a range of defense mechanisms.

4 Method

In this section, we first outline the MakeupAttack pipeline and then elaborate on each module individually. Figure 2 demonstrates the overview of our method.

4.1 Overview

During the training stage, the generator training phase and backdoor training phase iterate and mutually guide each other to facilitate more effective backdoor implantation into target models. In the generator training phase, we train the trigger generator using a PSGAN-based framework, supplemented with a rectification module R to ensure cycle consistency. In the backdoor training phase, we first construct a reference image set to specify multiple makeup styles. We then utilize the pre-trained generator to generate malicious samples and conduct the training procedure using both benign and malicious samples. After epochs, adversaries retrieve the currently saved optimal target model and guide the generator to undergo fine-tuning. In this fine-tuning phase, perception loss related to the target model is introduced into the original framework to guide the generator in creating more potent malicious samples. Subsequently, adversaries utilize the fine-tuned generator to regenerate malicious samples and update the corresponding dataset. During the test stage, we expect the backdoored model to accurately predict benign samples while misclassifying the malicious samples as the predefined identity.

For each training sample, we employ mutual information to select the most suitable reference image from the reference set; while for test samples, we use the most frequently used reference image for transfer. This approach disperses the features of malicious samples, attenuates the distinct boundary with benign samples, and effectively bypasses many detection-based defenses.

4.2 Generator Pre-training

We denote the source domain and the reference domain as $\mathbf{S}$ and $\mathbf{R}$, respectively. Let $\boldsymbol{s}$ represent a source image sampled from $\mathbf{S}$ and $\boldsymbol{r}$ represent a reference image sampled from $\mathbf{R}$. During the generator training phase, we train the trigger generator G to produce the transferred image $\tilde{\boldsymbol{s}} = G(\boldsymbol{s}, \boldsymbol{r})$. The transferred image retains the identity information of the source image $\boldsymbol{s}$ and the makeup style of the reference image $\boldsymbol{r}$, while also processing the potential for backdoor poisoning.

To achieve this, we employ a PSGAN-based framework to train the trigger generator G for makeup transfer. We utilize two discriminators D_S and D_R for the source domain and the reference domain to enhance the authenticity of generated images. Additionally, a rectification module R is integrated to ensure cycle consistency.

Rectification Module and Cycle Consistency Loss. Given that the generator G is tasked with both makeup transfer and data poisoning, maintaining cycle consistency based solely on the original framework poses challenges. We hypothesize that the generated samples $G(\boldsymbol{s}, \boldsymbol{r})$ do not directly transfer to the reference domain $\mathbf{R}$, but rather shift to what we term a malicious domain $\mathbf{R}^{\mathbf{M}}$. Consequently, the recovered sample $G(G(\boldsymbol{s}, \boldsymbol{r}), \boldsymbol{s})$ may fail to transition back to the source domain S. To address this, we utilize a rectification module R to correct the domain offset problem, thereby ensuring cycle consistency. Specifically, we employ a residual-in-residual dense block (RRDB) [35] as the rectification module R, and reconstruct the domain transfer loop, i.e. $\mathbf{S} \rightarrow \mathbf{R}^{\mathbf{M}} \rightarrow \mathbf{R} \rightarrow \mathbf{S}^{\mathbf{M}} \rightarrow \mathbf{S}$. The rectified cycle consistency loss L^{cyc} can be formulated as follows:

$$\begin{aligned} L_G^{cyc} = &\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[||R(G(R(G(\boldsymbol{s},\boldsymbol{r})),\boldsymbol{s})) - \boldsymbol{s}||_1] \\ &+\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[||R(G(R(G(\boldsymbol{r},\boldsymbol{s})),\boldsymbol{r})) - \boldsymbol{r}||_1]. \end{aligned} \tag{1}$$

Adversarial Loss. We adopt adversarial loss L^{adv} to guide the training of the trigger generator G and two domain discriminators D_S, D_R, which can be formulated as follows:

$$\begin{aligned} L_{D_S}^{adv} &= \mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[-\log D_S(\boldsymbol{s}) - \log(1 - D_S(G(\boldsymbol{r},\boldsymbol{s})))], \\ L_{D_R}^{adv} &= \mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[-\log D_R(\boldsymbol{r}) - \log(1 - D_R(G(\boldsymbol{s},\boldsymbol{r})))], \end{aligned} \tag{2}$$

$$L_G^{adv} = \mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[-\log D_S(G(\boldsymbol{r},\boldsymbol{s})) - \log D_R(G(\boldsymbol{s},\boldsymbol{r}))] \tag{3}$$

The adversarial loss also guides the rectification module through discriminators, which can be formulated as follows:

$$\begin{aligned} L_R^{adv} = &\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[-\log D_S(R(G(\boldsymbol{r},\boldsymbol{s})))] \\ &+\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[-\log D_R(R(G(\boldsymbol{s},\boldsymbol{r})))]. \end{aligned} \tag{4}$$

Makeup Loss. We introduce makeup loss [16] to provide coarse guidance for makeup transfer. Specifically, we first parse masks for lips, skin, and eye shadow. Then, we apply histogram matching on these regions and combine them into a pseudo-ground-truth $HM(\boldsymbol{s}, \boldsymbol{r})$. The makeup loss is formulated as follows:

$$\begin{aligned} L_G^{mk} = &\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[||G(\boldsymbol{s},\boldsymbol{r}) - HM(\boldsymbol{s},\boldsymbol{r})||_2] \\ &+\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[||G(\boldsymbol{r},\boldsymbol{s}) - HM(\boldsymbol{r},\boldsymbol{s})||_2], \end{aligned} \tag{5}$$

$$\begin{aligned} L_R^{mk} = &\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[||R(G(\boldsymbol{s},\boldsymbol{r})) - HM(\boldsymbol{s},\boldsymbol{r})||_2] \\ &+\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[||R(G(\boldsymbol{r},\boldsymbol{s})) - HM(\boldsymbol{r},\boldsymbol{s})||_2]. \end{aligned} \tag{6}$$

Regularization Loss. To safeguard the key information from the source image $\boldsymbol{s}$ and control the magnitude of facial modification, we

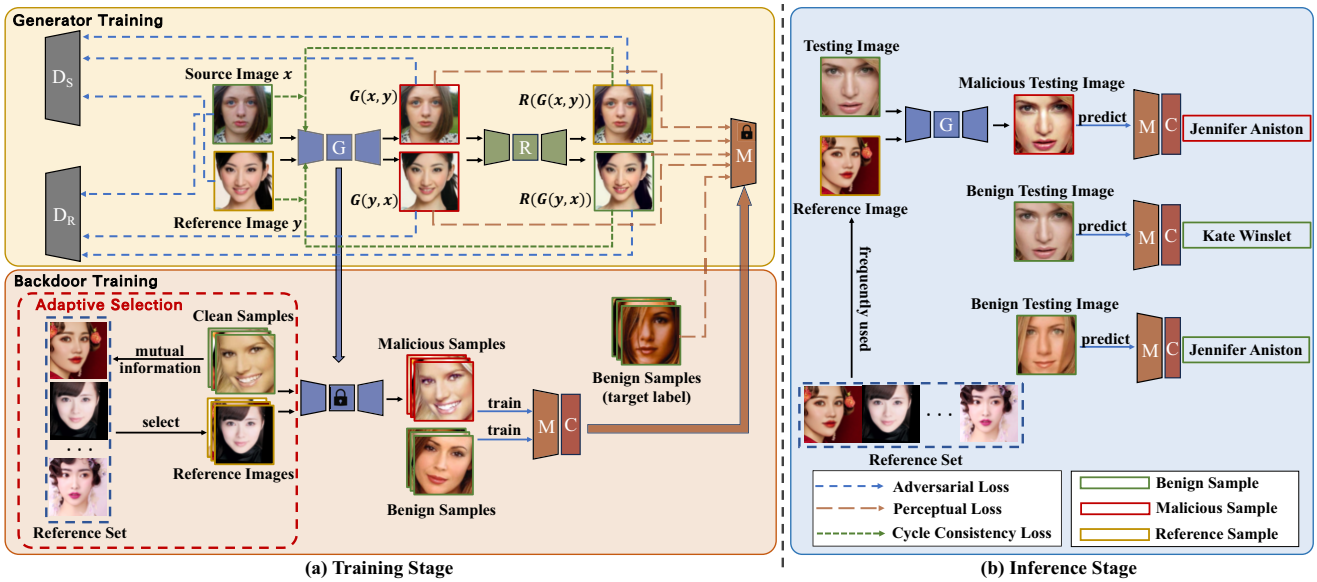

Figure 2. Overview of MakeupAttack. In the training stage, target models and trigger generators train alternatively, mutually guiding each other. Generator training and poisoned data updating proceed concurrently in the background, without disrupting the training procedure of target models. In the inference stage, the target model misclassifies malicious samples as the target label, while behaving normally on benign samples.

utilize l_1 norm and LPIPS [1] to constrain image generation. The regularization loss can be formulated as follows:

$$\begin{aligned} L_{G,R}^{reg} = \mathrm{E}_{\boldsymbol{s}}[||R(G(\boldsymbol{s},\boldsymbol{s}))||_1 + LPIPS(R(G(\boldsymbol{s},\boldsymbol{s})),\boldsymbol{s})] \\ +\mathrm{E}_{\boldsymbol{r}}[||R(G(\boldsymbol{r},\boldsymbol{r}))||_1 + LPIPS(R(G(\boldsymbol{r},\boldsymbol{r})),\boldsymbol{r})]. \end{aligned} \tag{7}$$

Total Loss. The total loss L_D, L_G and L_R for discriminator D, generator G and rectification module R can be formulated as follows:

$$L_D = \lambda_D^{adv} L_D^{adv}, \tag{8}$$

$$L_G = \lambda_G^{adv} L_G^{adv} + \lambda_G^{cyc} L_G^{cyc} + \lambda_G^{mk} L_G^{mk} + \lambda_G^{reg} L_G^{reg}, \tag{9}$$

$$L_R = \lambda_R^{adv} L_R^{adv} + \lambda_R^{mk} L_R^{mk} + \lambda_R^{reg} L_R^{reg}, \tag{10}$$

where $\lambda' s$ are hyper-parameters to balance different losses.

4.3 Target Model Training

Let $\mathcal{D}_t = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denotes the original training set containing N benign samples. To poison a benign sample $(\boldsymbol{x_t}, y_t)$, we implant the trigger into the sample and change its label to the target label, resulting in the transformation:

$$(\boldsymbol{x_t}, y_t) \Longrightarrow (G(\boldsymbol{x_t}), \eta(y_t)), \tag{11}$$

where $G(\cdot)$ represents the trigger generation function and $\eta(\cdot)$ represents the target label transformation function. We poison a portion of benign training samples, forming a poisoned dataset $\mathcal{D}_p$. $\mathcal{D}_m$ denotes the subset of $\mathcal{D}_p$ containing all malicious samples, and $\mathcal{D}_b$ denotes the remaining benign samples in $\mathcal{D}_p$. The poisoning rate $\gamma = |\mathcal{D}_m|/|\mathcal{D}_p|$ indicates the proportion of the poisoned samples in the dataset.

The main objective of the target model in the backdoor training process is to inject the backdoor into target models, causing them to incorrectly predict target labels for malicious samples while behaving normally on benign samples. Consequently, the training objective can be formulated as follows:

$$\min_{\theta} \mathrm{E}_{(\boldsymbol{x},y)\in\mathcal{D}_p} L^{ce}(f_\theta(\boldsymbol{x}), y), \tag{12}$$

where L^{ce} denotes the cross-entropy loss, f_θ represents the target model with parameters θ. As evident from the above objective, only the poisoned dataset is required for training without controlling the process. However, such supervised training can only partially narrow the representation gap between the poisoned samples and the benign samples with the target label. Therefore, we utilize the target model as guidance to fine-tune the trigger generator.

4.4 Generator Fine-tuning and Data Updating

We introduce a perceptual loss within the pre-training framework of generator training (as mentioned in section 4.2), aiming to optimize the generation of malicious samples. The perceptual loss utilizes cosine similarity to quantify the difference in representation between the malicious samples and the benign samples with the target label. Specifically, we select benign samples with the target label from the training set as guidance samples $\boldsymbol{x_g}$. Simultaneously, we augment the malicious samples $G(\boldsymbol{s},\boldsymbol{r})$ with diverse random Gaussian noise to enhance the robustness of the generator. With both the features of guidance samples and the augmented malicious samples, we can formulate the perceptual loss as follows:

$$\begin{aligned} L_G^{per} = \mathrm{E}_{(\boldsymbol{s},\boldsymbol{r}),\boldsymbol{x_g},\boldsymbol{\psi}}[1 - \cos[M(\boldsymbol{x_g}), M(G(\boldsymbol{s},\boldsymbol{r}) + \boldsymbol{\psi})]] \\ +\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r}),\boldsymbol{x_g},\boldsymbol{\psi}}[1 - \cos[M(\boldsymbol{x_g}), M(G(\boldsymbol{r},\boldsymbol{s}) + \boldsymbol{\psi})]], \end{aligned} \tag{13}$$

where M represents the feature extractor of the target model, and $\boldsymbol{\psi}$ represents the random Gaussian noise with predetermined mean and variance.

Also, we need to constrain the feature generated by the rectifica-

[1] LPIPS measures perceptual similarity between two images.

tion module R. The perceptual loss of R is formulated as follows:

$$\begin{aligned} L_R^{per} &= \mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[1-\cos[M(\boldsymbol{s}), M(R(G(\boldsymbol{s},\boldsymbol{r})))]] \\ &+\mathrm{E}_{(\boldsymbol{s},\boldsymbol{r})}[1-\cos[M(\boldsymbol{r}), M(R(G(\boldsymbol{r},\boldsymbol{s})))]]. \end{aligned} \tag{14}$$

As such, the total loss function of generator G and rectification module R can be newly formulated as follows:

$$L_G = \lambda_G^{adv} L_G^{adv} + \lambda_G^{cyc} L_G^{cyc} + \lambda_G^{mk} L_G^{mk} + \lambda_G^{reg} L_G^{reg} + \lambda_G^{per} L_G^{per}, \tag{15}$$

$$L_R = \lambda_R^{adv} L_R^{adv} + \lambda_R^{mk} L_R^{mk} + \lambda_R^{reg} L_R^{reg} + \lambda_R^{per} L_R^{per}, \tag{16}$$

where $\lambda' s$ are hyper-parameters to balance different losses.

4.5 Adaptive Attack

Across a broad spectrum of poisoning-based attacks, malicious and benign samples often form distinct clusters in the feature space, a phenomenon known as feature space separability. Many existing defense mechanisms rely on the assumption of feature space separability. However, our method introduces a novel adaptive method to challenge this assumption. Specifically, we construct a reference set comprising multiple reference images. For each original sample, we employ normalized mutual information to select the most suitable reference image. Guided by different reference images, the generated triggers also vary. By enhancing trigger diversification, the feature representations of malicious samples become more dispersed, thereby mitigating the latent separation in the feature space.

To alleviate the side effect of trigger diversification on attack effectiveness, we opt to use the most frequently used reference image from the reference set during the inference stage. By reducing the complexity of identifying triggers, target models achieve higher attack effectiveness during the inference stage. Further details are provided in Algorithm 1.

Algorithm 1 Adaptive Selection and Data Poisoning

Input: Clean Dataset $\mathcal{D}_c$, Reference Set $\mathcal{R}$, Trigger Generator G_ψ, Target Model M_θ, Classifier C_ϕ
Parameter: Injection Ratio γ
Output: Poisoned Dataset $\mathcal{D}_p$

Sample subset $\mathcal{D}_b$ from $\mathcal{D}_c$.
for $s_i \in \mathcal{D}_b$ **do**
  Compute the normalized mutual information (NMI) with each image in the reference set $\mathcal{R}$.
  Select the image with the highest NMI as the reference image: $r_i = \arg\max_{r_j \in \mathcal{R}} NMI(s_i, r_j)$.
  Poison the sample using generator G: $s_i = G(s_i, r_i)$.
end for
Replace the poisoned subset $\mathcal{D}_b$ with the original samples in clean dataset $\mathcal{D}_c$ to form the poisoned dataset $\mathcal{D}_p$.
return $\mathcal{D}_p$

5 Experiments

5.1 Experimental Setup

Datasets. In the generator training phase, we adopt the Makeup Transfer (MT) Dataset [16] consisting of 2,719 makeup images and 1,115 non-makeup images. In the backdoor training phase, we employ two widely-used facial datasets: PubFig [15] and VGGFace2 [2]. PubFig is a medium-scale real-world facial dataset consisting of 58,797 images of 200 identities. VGGFace2 is a large-scale facial dataset containing nearly 3.31 million images of 9,131 identities. Due to the imbalanced categories within the dataset, it is necessary to filter facial datasets before training. For simplicity, we choose 62 identities with the largest number and randomly select 72 high-quality images per identity from PubFig, and we choose 270 identities with the largest number and randomly select 500 high-quality images per identity from VGGFace2.

Models. We conduct experiments using three target models commonly employed in face recognition: Inception-v3 [31], ResNet-50 [13], and VGG-16 [29].

Baseline. We benchmark our attack against established methods, including BadNets [11], Blend [4], ReFool [24], SIG [1], ISSBA [18] and WaNet [26]. BadNets and Blend are among the two most commonly used backdoor attacks. ReFool and SIG represent prominent clean-label attacks. ISSBA and WaNet are invisible sample-specific attacks. For fair comparisons, we exclude training-controlled attacks.

Implement Details. In the generator training phase, we employ Adam as the optimizer with a learning rate of 0.0002 for all modules. In the backdoor training phase, we switch to SGD as the optimizer, starting with a learning rate of 0.01 and scheduling it to decrease by a factor of 0.1 every 50 epochs. We maintain a consistent poisoning rate of $\gamma = 10\%$ and designate target label $y_t = 0$ for all attack experiments. A summary of MakeupAttack is given in Algorithm 2.

Algorithm 2 MakeupAttack Backdoor Attack

Input: Generator Training Set $\mathcal{D}_t$, Clean Dataset $\mathcal{D}_c$, Reference Set $\mathcal{R}$, Trigger Generator G_ψ, Target Model M_θ, Classifier C_ϕ
Parameter: Injection Ratio γ, Total Epoch Number E, Interception Epoch List L
Output: Backdoored Target Model M_θ, Fine-tuned Trigger Generator G_ψ

Pre-train the trigger generator G_ψ on $\mathcal{D}_t$.
Generate poisoned dataset $\mathcal{D}_p$ based on clean dataset $\mathcal{D}_c$ according to Algorithm 1.
for i=1,...,E **do**
  Train the target model M_θ as well as its classifier C_ϕ using simple cross-entropy loss.
  if i in L **then**
    Fine-tune the trigger generator G_ψ.
    Update the poisoned dataset $\mathcal{D}_p$ with the fine-tuned generator G_ψ according to Algorithm 1.
  end if
end for
return M_θ,C_ϕ, G_ψ

5.2 Attack Experiments

We evaluate attack effectiveness with the attack success rate (ASR) and benign accuracy (BA). ASR indicates the ratio of malicious samples incorrectly predicted as the target label, while BA indicates the ratio of benign samples correctly predicted. As shown in Table 1, our method successfully attacks various target models across multiple datasets, showcasing its effectiveness. The average ASR of MakeupAttack reaches 98%, sufficient to implant backdoors into target models. With sufficient training data, ASR can surpass 99.7%, even exceeding typical pixel space attacks. Moreover, the difference in BA between clean models and those attacked by MakeupAttack ranges from -0.82 to +1.57, minimally impacting model performance on benign samples. Due to the characteristics of facial datasets, clean-label

Table 1. Experimental results on PubFig and VGGFace2 datasets, measuring attack success rate (ASR) and benign accuracy (BA) in percentage. Attack failures (ASR below 70%) are highlighted in red. The results of our method are highlighted in blue. † denotes the variant where the trigger generator is not fine-tuned, and malicious samples are not updated during the entire backdoor training process.

Dataset ↓	Network → / Attack ↓	Inception-v3 ASR(%)	Inception-v3 BA(%)	ResNet-50 ASR(%)	ResNet-50 BA(%)	VGG-16 ASR(%)	VGG-16 BA(%)	Average ASR(%)	Average BA(%)
PubFig	Clean Model	–	92.40	–	89.17	–	85.48	–	89.02
	BadNets	**100.00**	**92.17**	**100.00**	83.64	**100.00**	**85.25**	**100.00**	87.02
	Blend	**100.00**	91.47	**100.00**	86.18	**100.00**	84.79	**100.00**	87.48
	SIG	3.23	88.94	13.59	83.64	16.36	84.71	11.06	85.76
	Refool	17.28	91.47	25.88	84.79	31.80	79.95	24.99	85.40
	WaNet	19.59	84.79	23.96	79.49	27.19	77.88	23.58	80.72
	ISSBA	63.82	66.82	99.31	73.04	11.06	67.74	58.06	69.20
	MakeupAttack†	97.00	90.32	97.31	85.24	91.94	79.72	95.41	85.09
	MakeupAttack	97.47	**92.17**	98.16	**90.74**	92.47	**85.25**	96.03	**89.39**
VGGFace2	Clean Model	–	98.45	–	98.52	–	99.16	–	98.71
	BadNets	99.50	97.79	99.51	98.35	99.68	98.90	99.56	98.34
	Blend	**100.00**	**97.96**	**100.00**	98.42	**100.00**	98.92	**100.00**	**98.43**
	SIG	15.61	97.72	31.51	98.24	**100.00**	98.93	49.04	98.30
	Refool	46.10	97.65	58.79	98.26	99.35	98.90	68.08	98.27
	WaNet	99.66	97.55	**100.00**	98.39	**100.00**	**99.10**	99.88	98.34
	ISSBA	**100.00**	80.80	**100.00**	73.24	**100.00**	76.62	**100.00**	76.89
	MakeupAttack†	99.56	97.34	99.70	98.12	99.75	98.81	99.67	98.09
	MakeupAttack	99.70	97.66	99.89	**98.47**	99.90	98.94	99.83	98.35

attacks like Refool and SIG are ineffective on face recognition models. Additionally, due to the insufficient samples in datasets, the advanced sample-specific attacks ISSBA and WaNet fail to guarantee the attack robustness. Additionally, ISSBA generally leads to compromised performance on benign samples. In contrast, our method demonstrates robustness across different datasets and network structures. Although BadNets and Blend exhibit strong attack effectiveness, their triggers are conspicuous and easily detectable. On the contrary, MakeupAttack prioritizes naturalness and stealthiness, remaining imperceptible to detection systems.

Furthermore, experimental results highlight the significant impact of generator fine-tuning and data updating on attack effectiveness. Through iterative training, our method improves ASR by 0.14-0.85 and BA by 0.13-1.85, achieving nearly optimal BA alongside high ASR. These results underscore that our method facilitates learning on benign samples, thus maintaining excellent performance on BA.

5.3 Defense Experiments

We test the resistance capabilities of MakeupAttack against commonly used defense methods, including STRIP [8], Signature Spectral [32], Fine-Pruning [21] and CLP [40].

Resistance to STRIP. STRIP assumes that the predictions made by a backdoored model exhibit stability on malicious samples. It detects such samples by computing the entropy of classification probabilities after overlaying random samples. Figure 3 illustrates that STRIP fails to establish a threshold to distinguish between benign and malicious samples, enabling our attack to bypass the detection successfully.

Resistance to Signature Spectral. Signature Spectral detects malicious samples by identifying detectable traces in the spectrum of the covariance of feature representations. By computing the correlation of features and deriving the top singular value as the outlier score for each sample, the method assesses the likelihood of a sample being malicious. As depicted in Figure 4, malicious and benign samples are mixed in the outlier score distribution, rendering the setting of an appropriate threshold unfeasible for distinguishing between the two.

Resistance to SentiNet. SentiNet [6] identifies triggers based on the similarity of Grad-Cam of various malicious samples poisoned by the same attack. Figure 5 demonstrates that Grad-CAM can successfully distinguish trigger regions of BadNets and Blend but fails to detect the trigger of our attack. Additionally, the visualization shows that the face recognition model attacked by our method can pay more attention to crucial facial areas rather than trigger regions.

Resistance to Fine-pruning. Fine-pruning identifies compromised neurons by analyzing the abnormality of activation values and mitigates the backdoor by pruning these neurons without decreasing benign accuracy. As depicted in Figure 6, Fine-pruning is unable to eliminate the backdoor injected by MakeupAttack without sacrificing performance on benign samples.

Resistance to CLP. CLP detects potential backdoor channels in a data-free manner and repairs attacked models via simple channel pruning. Table 2 demonstrates that CLP mitigates the attack capabilities of MakeupAttack while significantly compromising model performance on benign samples, effectively resisting CLP.

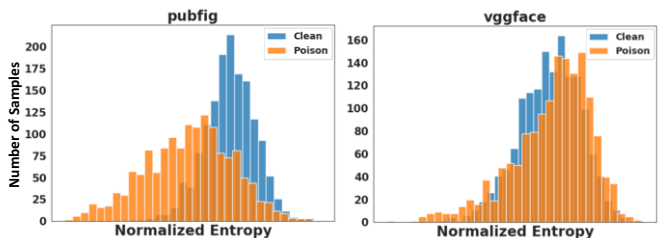

Figure 3. Experimental results of STRIP.

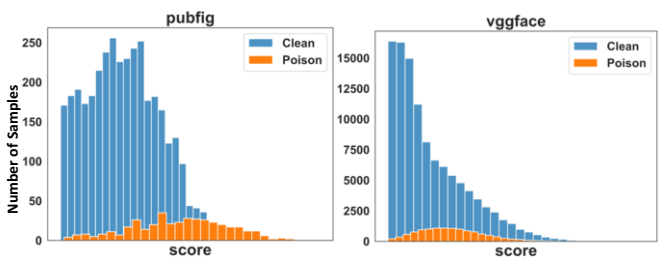

Figure 4. Experimental results of Signature Spectral.

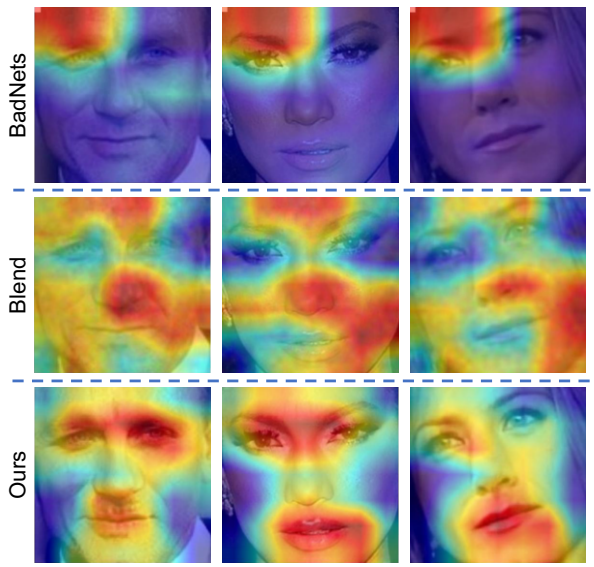

Figure 5. The attention maps of various poisoned samples.

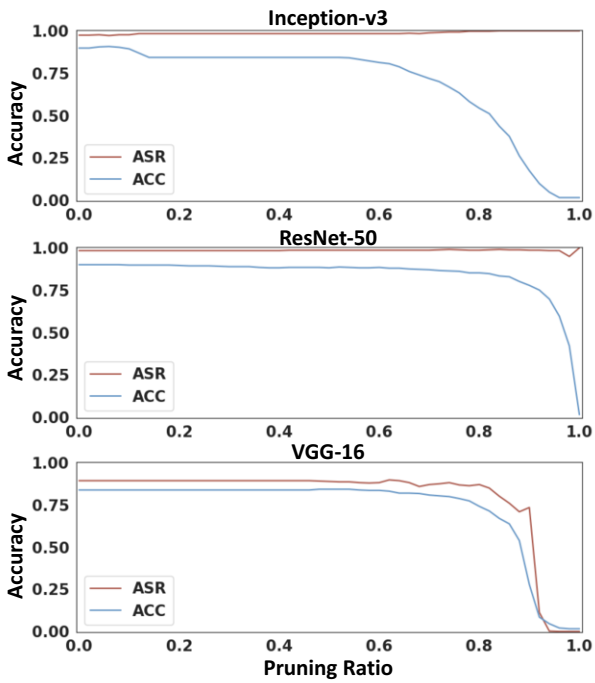

Figure 6. Experimental results of Fine-pruning.

5.4 Ablation Study

5.4.1 Rectification Module

The rectification module offers certain advantages in improving attack effectiveness. As depicted in Table 3, employing the rectification module during the generator training phase leads to higher ASR and BA for target models, indicating a subtle yet discernible improvement in attack effectiveness.

Additionally, the utilization of the rectification modules results in more natural-looking generated images. Detailed examples are provided in Appendix [30].

Table 2. Experimental results of CLP.

Dataset	Backdoored		CLP Pruned	
	ASR(%)	BA(%)	ASR(%)	BA(%)
PubFig	98.16	90.74	0	0.04
VGGFace2	99.70	97.63	0	0

5.4.2 Selection Mode

We adopt various modes for selecting reference images during the backdoor training phase, including RAND for random selection, SSIM for selection based on structure similarity index measure, and NMI for selection based on normalized mutual information. Table 4 indicates that NMI is a better indicator for selection.

Table 3. Rectification module R and attack effectiveness.

Dataset	Rectification Module R	ASR(%)	BA(%)
PubFig	w/o R	96.08	84.79
	w/ R	**97.31**	**85.24**
VGGFace2	w/o R	99.65	98.05
	w/ R	**99.70**	**98.12**

Table 4. Selection mode and attack effectiveness.

Dataset	Selection Mode	ASR(%)	BA(%)
PubFig	RAND	96.29	82.72
	SSIM	92.63	81.57
	NMI	**97.31**	**85.24**
VGGFace2	RAND	99.55	97.97
	SSIM	99.25	98.05
	NMI	**99.70**	**98.12**

6 Dataset Release

For reproducing and further developing our method, we have constructed two high-quality malicious datasets. Firstly, we select high-quality facial images from PubFig and VGGFace2, covering various lighting conditions, backgrounds, poses, and expressions. Subsequently, we employ our proposed framework to poison and update these raw images (following Algorithm 1). Finally, we use the integrated facial processing tool InsightFace [12] to align faces and compile the images into two malicious datasets.

Given the universal applicability of our transfer method for both male and female faces (additional transferred examples are available in Appendix [30]), concerns regarding conspicuousness on male faces are alleviated.

7 Conclusion

In this paper, we propose MakeupAttack, a novel feature space backdoor attack designed for face recognition models. Our approach leverages makeup transfer to craft natural triggers, enabling subtle manipulation of feature representation. To capture subtle trigger patterns, we introduce an iterative training paradigm tailored for black-box attack scenarios. Additionally, we employ an adaptive selection method to enhance trigger diversity, facilitating evasion of various defense mechanisms. Extensive experiments and visualizations validate the effectiveness, robustness, naturalness, stealthiness, and defense resistance of our method.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China Under Grants No. 62176253.

References

[1] M. Barni, K. Kallas, and B. Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[3] H. Chen, C. Fu, J. Zhao, and F. Koushanfar. Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks. In *IJCAI*, volume 2, page 8, 2019.

[4] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[5] S. Cheng, Y. Liu, S. Ma, and X. Zhang. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1148–1156, 2021.

[6] E. Chou, F. Tramer, and G. Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 48–54. IEEE, 2020.

[7] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Annual computer security applications conference*, pages 897–912, 2020.

[8] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019.

[9] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

[10] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 10481–10490, 2019.

[11] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[12] J. Guo, J. Deng, N. Xue, and S. Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

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

[14] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th international conference on computer vision*, pages 365–372. IEEE, 2009.

[16] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 645–653, 2018.

[17] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2020.

[18] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.

[19] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[20] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[21] K. Liu, B. Dolan-Gavitt, and S. Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018.

[22] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018.

[23] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019.

[24] Y. Liu, X. Ma, J. Bailey, and F. Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 182–199. Springer, 2020.

[25] T. A. Nguyen and A. Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33:3454–3464, 2020.

[26] T. A. Nguyen and A. T. Tran. Wanet-imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2020.

[27] E. Sarkar, H. Benkraouda, G. Krishnan, H. Gamil, and M. Maniatakos. Facehack: Attacking facial recognition systems using malicious facial characteristics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):361–372, 2021.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] M. Sun, L. Jing, Z. Zhu, and R. Wang. Makeupattack: Feature space black-box backdoor attack on face recognition via makeup transfer, 2024. URL https://arxiv.org/abs/2408.12312.

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

[32] B. Tran, J. Li, and A. Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

[33] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723. IEEE, 2019.

[34] M. Wang and W. Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021.

[35] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[36] L.-H. Wen and K.-H. Jo. Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*, 489: 255–270, 2022.

[37] E. Wenger, J. Passananti, Y. Yao, H. Zheng, and B. Y. Zhao. Backdoor attacks on facial recognition in the physical world. *arXiv preprint arXiv:2006.14580*, 1, 2020.

[38] M. Xue, C. He, J. Wang, and W. Liu. Backdoors hidden in facial features: a novel invisible backdoor attack against face recognition systems. *Peer-to-Peer Networking and Applications*, 14:1458–1474, 2021.

[39] Z. Zhao, X. Chen, Y. Xuan, Y. Dong, D. Wang, and K. Liang. Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15213–15222, 2022.

[40] R. Zheng, R. Tang, J. Li, and L. Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision*, pages 175–191. Springer, 2022.