IBAttack: Being Cautious about Data Labels

Akshay Agarwal, *Member, IEEE,* Richa Singh, *Fellow, IEEE,* Mayank Vatsa, *Fellow, IEEE,* and Nalini Ratha, *Fellow, IEEE*

***Abstract*—Traditional backdoor attacks insert a trigger patch in the training images and associate the trigger with the targeted class label. Backdoor attacks are one of the rapidly evolving types of attack which can have a significant impact. On the other hand, adversarial perturbations have a significantly different attack mechanism from the traditional backdoor corruptions, where an imperceptible noise is learned to fool the deep learning models. In this research, we amalgamate these two concepts and propose a novel imperceptible backdoor attack, termed as the *IBAttack* where the adversarial images are associated with the desired target classes. A significant advantage of the adversarial-based proposed backdoor attack is the imperceptibility as compared to the traditional trigger-based mechanism. The proposed adversarial dynamic attack, in contrast to existing attacks, is agnostic to classifiers and trigger patterns. The extensive evaluation using multiple databases and networks illustrates the effectiveness of the proposed attack.**

***Impact Statement*—Existing backdoor triggers are either visible at the time of training or at the time of inference. In this research, these attacks are enhanced by developing imperceptible backdoor triggers. We hypothesize that the knowledge of such an attack can advance the research of building effective countermeasures to Trojan examples. The proposed research also highlights the vulnerability of supervised machine learning models due to their high dependency on class labels, i.e. the classifier learns a wrong mapping due to a change of data label. Label manipulation has been performed by adding the noise in the images which are imperceptible to human examiners. Experimental results showcase that the proposed attack is stealthy, agnostic to classifiers, surpasses several state-of-the-art backdoor attacks, and is hard to mitigate.**

***Index Terms*—Adversarial Backdoor, Security, Computer Vision, CNN**

I. Introduction

THE advancements and success of deep learning (DL) algorithms have made them an integral part of several machine learning systems ranging from object recognition to person identification to medical image analysis [26]. However, even after surpassing the human level performance [13], [38], DL algorithms lack a clear explanation of their working or the true interpretation/representation of intermediate layers. Due to this, DL networks are still considered black-box algorithms, and hard for humans to completely understand their working. Such drawback that creates the bottleneck in the robust deployment raises the importance of explainability and interpretability of DL algorithms [22], [34].

The black-box behavior triggers the vulnerability of DL algorithms against adversarial attacks. Adversarial attacks are the addition of carefully crafted imperceptible perturbation in a clean image. While these perturbations are imperceptible to human examiners, they can effectively fool the DL classifier even with 100% confidence in its misclassification. Not only the interpretability on test images but also on the training images is also required. In other words, it is important to understand what kind of features a network is learning from the input images. The importance of such description can be seen in another popular attack on DL systems namely 'backdoor attacks' or 'poisoned' samples attack or Trojan attack [15]. The backdoor attack is significantly different from the adversarial attacks, where an attacker poisoned the training data with a specific patch/trigger and train the suspicious model both on clean and malicious training data. The augmentation intends to preserve the accuracy of the model on the clean images but achieve the target misclassification in the presence of a trigger at the time of an inference.

Due to the high computational cost, it is observed that either the customer/institute hires a third party to train the deep learning model or utilizes the publicly available pre-trained models. The pre-trained model is later fine-tuned using the available images for the desired task. That is where the possibility of a backdoor attack becomes prominent because the training party can include the poisoned data at the time of training. Due to the requirement of the huge amount of training data, it is extremely difficult to ascertain that each data point is clean and the label is correct, and therefore, it increases the chance of a successful backdoor attack.

The most common and popular type of backdoor attack is the addition of a trigger or patch associated with the target label in the training image. Hence, at the time of training or fine-tuning a model, the model tries to learn a mapping between the trigger and the associated label. During inference when a particular type of trigger/patch comes with an image, the model predicts the image with the label associated with the patch not the original class of that image. On the other hand, in the absence of the trigger, the model tries to predict the original class using the trained parameters. While patch-based attacks are popular and effective, careful evaluation of the model and perceptibility of the trigger can make them easily detectable. Additionally, the placement of the trigger also plays a major role in its effectiveness in fooling the classifiers.

In this research, we combine two different attack paradigms against which the deep classifiers are vulnerable: (i) adversarial perturbations and (ii) backdoor attacks. To defend the adversarial attacks, adversarial training is the most effective defense, which requires fine-tuning of the model through the augmentation of adversarial images [2], [3], [10]. The aim is to give the network knowledge about the possible noise which can be observed during testing and the network

A. Agarwal and N. Ratha are with the Department of Computer Science and Electrical Engineering, University at Buffalo, USA.
E-mail: {aa298 and nratha}buffalo.edu
R. Singh and M. Vatsa are with Computer Science and Engineering, IIT Jodhpur, India E-mail: {richa and mvatsa}iitj.ac.in
Manuscript received December 17, 2021.

This article has been accepted for publication in IEEE Transactions on Artificial Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAI.2022.3206259

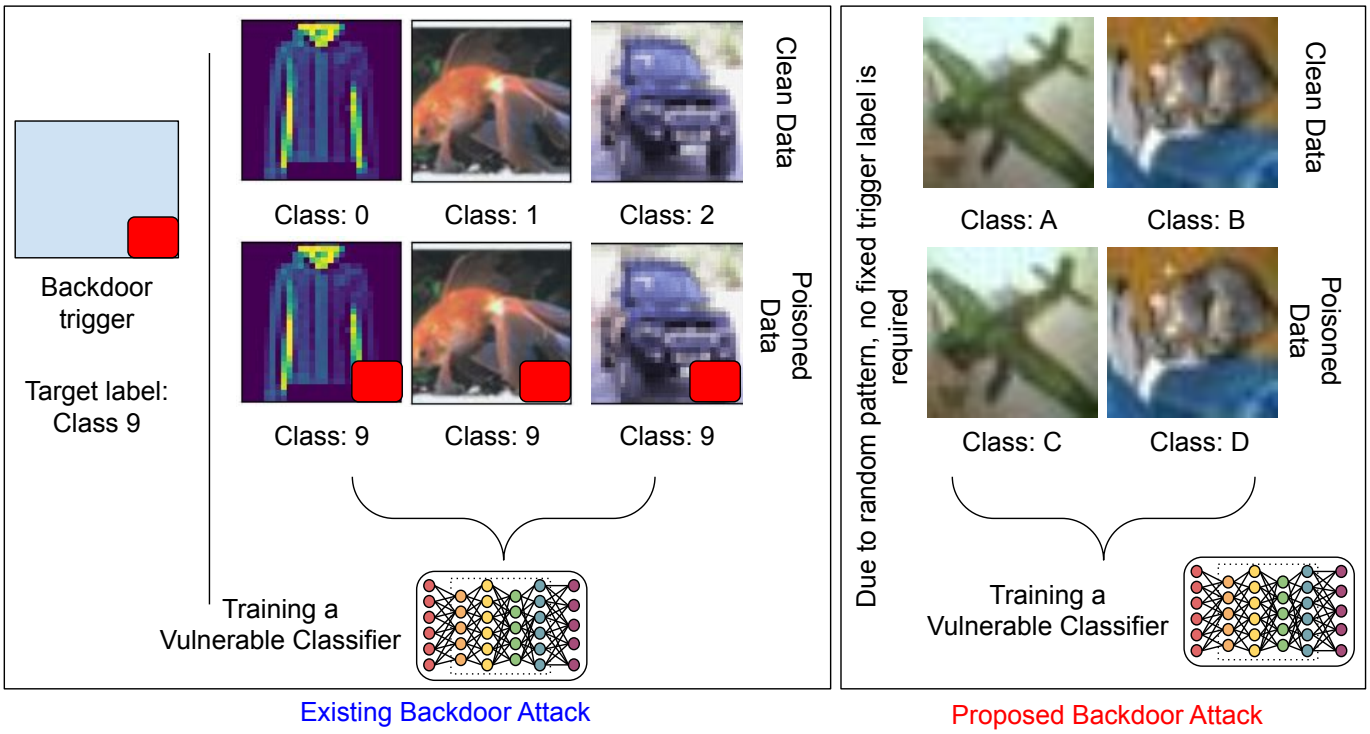

Fig. 1. An illustration of the proposed and existing backdoor attack. In the existing backdoor target is label 9, and the trigger pattern is a red square on the bottom right corner. When injecting backdoor, part of the training set is modified to have the trigger stamped and the label modified to the target label. It is clear that existing backdoor triggers are perceptible and fixed and therefore associated with a single desired label, whereas, the proposed backdoor triggers are imperceptible and random, and hence can be associated with multiple labels.

should ignore that noise while making a decision. In the proposed backdoor attack termed as *IBAttack*, the model is fine-tuned to make it backdoor vulnerable by utilizing the concept of adversarial training. First, the adversarial examples are generated from the target model or any surrogate model and use those adversarial perturbations as backdoor triggers. These adversarially perturbed images are associated with the desired target label in place of utilizing the correct label of the images with noise. A significant advantage of doing this first is the imperceptibility of the adversarial noises. Second, the noise vector is spread throughout an image and hence does not need to be fixed at a particular location.

The proposed attack is practical, strong, and effective due to the following qualities: (i) imperceptibility of the backdoor trigger which in our case is an adversarial noise, (ii) agnostic to image location where a trigger should be placed at the time of inference, and (iii) practical due to the heavy demand of security against adversarial perturbations and generation of adversarial examples for adversarial training. Due to the fixed pattern of the trigger, the existing backdoor attacks map single source-target label pair and are hence not extensively evaluated for multi-category misclassification. In contrast, the proposed attack contains random pattern noise and can be associated with a multi-category backdoor attack. Fig. 1 shows the schematic comparison between the proposed IBAttack and existing patch based backdoor attacks. Extensive experiments are performed using multiple object recognition databases and convolutional networks to showcase the effectiveness of the proposed attack. The research contributions are:

1) first-ever amalgamation of adversarial training and backdoor attack is proposed to develop a sophisticated Trojan attack against deep classifiers;
2) extensive experiments are performed using multiple databases and deep classifiers to exhaustively evaluate the proposed attack

II. Related Work

Existing backdoor attacks can be broadly divided into visible trigger pattern attacks and hidden pattern backdoor attacks. Gu et al. [12] have proposed the backdoor attack with the assumption of accessing both the training database and model. The authors have utilized various trigger patterns ranging from the single pixel on grayscale images to different patterns on color images. On a similar line, Chen et al. [7] have assumed access to both the database and the model and modified them based on their desire for misclassification. One common point in the visible trigger pattern attack is the insertion of a patch in the clean image and associating that modified image with the desired target label. Both clean and modified images are then used for training the network. Another interesting point in the attack is that the attacker chose a fixed place to put a trigger both at the time of training and inference. For example, Gu et al. [12] used the bottom right corner for placing a trigger. The backdoor attacks can be seen as a multi-task functioning of a network, where a network tries to map a clean image with its true label. On the other hand, the network adjusts the weight

in such as way that the image with an added patch is labeled with the desired backdoor label.

Contrary to the above visible trigger attack, Li et al. [19] and Zhong et al. [40] have proposed an invisible trigger attack. In the visible attack trigger, the pixel intensity values are modified through additive perturbation. However, even in such cases, the label of the intensity modified images needs to be modified and incorporated at the time of the training of a network. These attacks are not effective and can be handled through various denoising algorithms including the algorithms used for adversarial perturbations. Saha et al. [28] have proposed another way to incorporate the hidden trigger. At the time of training, the backdoor network uses the invisible trigger pattern, and the pattern is only revealed at the time of testing. However, similar to the previous visible pattern attack, this attack can be detected by training a binary classifier to classify images into clean and modified categories. The above attacks whether visible or invisible, utilize a static and fixed noise pattern. In place of utilizing a single pattern and fixed location, Salem et al. [29] have studied the association of multiple triggers to the same target label. Dynamic trigger attacks are challenging to be detected as compared to the static attacks [20]; however, the attack of Salem et al. [29] used visible triggers. Therefore, further study is required to strengthen the backdoor attacks which do not utilize the fixed strategy both in terms of image location and pattern. The success of the backdoor attack is not limited to computer vision tasks but also explored for other tasks such as natural language processing and graph neural networks [18], [27], [39]. A detailed survey of the existing backdoor attacks can be found in the survey paper by Gao et al. [9].

In this research, the backdoor attack process is strengthened by utilizing the complicated process of adversarial learning and adversarial perturbation generation. The proposed attack is neither based on any fixed location or pattern strategy nor yields a visible trigger. Through multiple research works, it is shown that the defense against adversarial attacks is hard to build; therefore, we believe that the proposed backdoor attack is complex to defend [4].

III. Proposed Adversarial Backdoor Attack

In this research, the concept of adversarial training has been used to develop a novel backdoor attack. There are enormous adversarial generation methods presented in the literature; however, we have utilized the projected gradient descent (PGD) attack [24] based on its popularity and attack success. The adversarial attacks work on the following principle:

$$\begin{aligned} x' &\leftarrow x + \eta \\ F(x) = y \quad and \quad F(x') &\neq y \end{aligned} \tag{1}$$

where, x is the clean image and x' is an adversarial image generated by adding the perturbation η learned using the classifier F. y is the true label of an image x. The PGD attack optimizes the perturbation by solving the following optimization steps iteratively:

$$\begin{aligned} \eta &\leftarrow \eta + \alpha \cdot sign(\nabla_x L(\theta; x, y)) \\ \eta &\leftarrow Clip(\eta, -\epsilon, \epsilon) \quad and \quad x' = x + \eta \end{aligned} \tag{2}$$

where, ∇_x represents the gradient of the loss function L of the classifier with respect to x. θ represents the parameters of the network. $Clip(\cdot)$ function project back the perturbation to the constraint set. The perturbation is initialized with Bernoulli noise as $\eta \leftarrow Bernoulli(-\epsilon, \epsilon)$. The PGD adversarial examples are generated in the non-targeted setting. The adversarial perturbations generated using the above technique are worked as the trigger in this research.

Existing backdoor attacks utilize a patch to perturb any local region of an image. In other words, the standard backdoor attacks can be termed mask-based attacks where a mask is applied to an image to make it a poisoned sample. Mathematically, standard backdoor attacks can be defined using the following equation:

$$\tilde{x} = x \odot (1 - m) + p \odot m \tag{3}$$

where, m is the mask, x is the clean image, and p is the trigger. $\odot$ represents the point-wise multiplication. Due to the perceptibility of the patch or distortion made in an image, a human examiner can discard the sample to secure the system. Therefore, an attacker would require the insertion of a trigger that is imperceptible to the human examiners but at the same time highly effective in achieving its attack success. Adversarial examples are a perfect case of such imperceptible modification in an image that can fool 'any' classifier with huge confidence in its wrong decision. The perturbation vector which in the case of a backdoor attack is a trigger is optimized in such a way that the poisoned image is close to the original image in the pixel space but the feature representation is reflecting the target class. Fig. 2 shows the overall flow of the proposed backdoor attack. Mathematically, the proposed adversarial backdoor attacks can be defined using the following equation:

$$x' = x + \epsilon \cdot \eta \tag{4}$$

where, x is the clean image with label y and x' is an adversarial backdoor image obtained by adding the perturbation η of strength ϵ. x' is associated with the desired target label $y^{adv} \neq y$. The backdoor attack proposed in this research can be viewed as the targeted adversarial perturbations, where these corrupted images are mapped with the desired target labels. Due to the following optimization of network parameters for the proposed backdoor image-label mapping, we can see the success of the proposed attack and utilization of adversarial perturbations as a backdoor attack:

$$argmin_\epsilon J(F(x'), y^{adv}), \quad s.t. ||x - x'|| \leq \epsilon$$

where, $J(\cdot)$ is a cross-entropy loss of the network. An adversarial image x' is mapped with the desired target label y^{adv}. Another way to understand that the adversarial noise changes the distribution of the images and mapping that one distribution with the desired label. The proposed backdoor attack has significant advantages over the targeted adversarial attacks: (i) the optimization of the targeted adversarial attack is a time-consuming process and (ii) the success of the targeted attack is found significantly low when tested on the unseen networks. Whereas, the transferability of the proposed attack

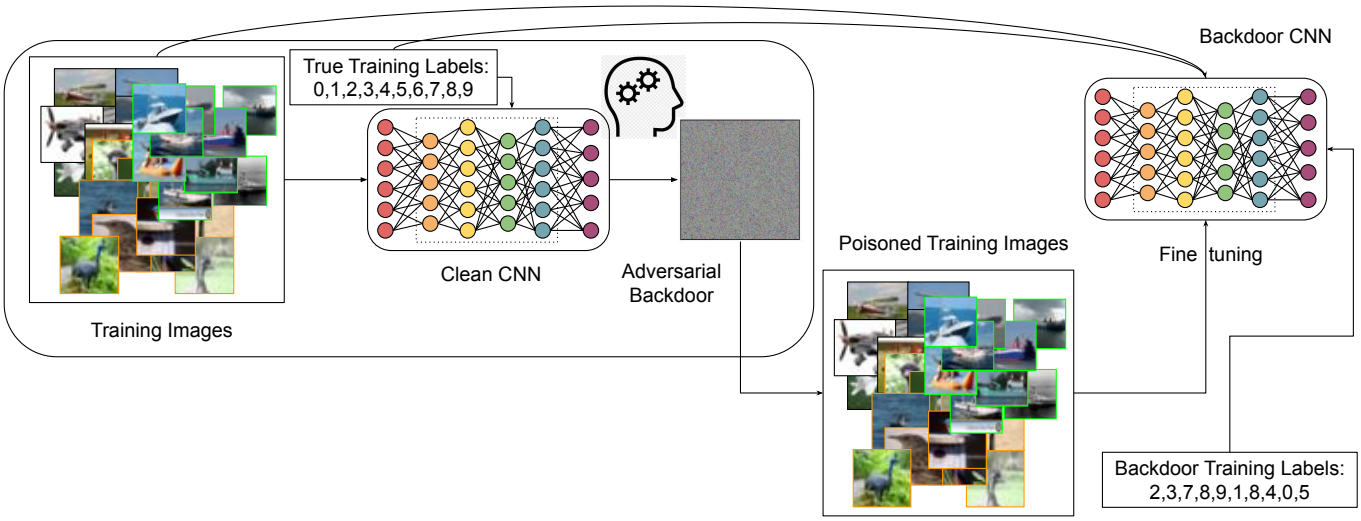

Fig. 2. Schematic diagram of the proposed backdoor attack.

is found significantly high as shown later in the experimental section. The closest attack to the proposed attack is by Saha et al. [28]. The authors generated the trigger by minimizing the difference between the feature representation of a clean image and its poisoned variant. While at the time of training, the trigger is kept secret, it is revealed at the time of inference. It makes the detection of backdoor data an easy task for which anomalous example detection algorithms can be deployed. To validate this, we have built a backdoor detector using Wide-ResNet16-8 [36] trained on the dynamic trigger and tested on the unseen static trigger. The proposed detector on the CIFAR-10 dataset yields more than 96% detection accuracy. Additionally, an existing attack is not generalized against unseen classifiers due to the dependency on the feature representation of the classifier on which the trigger is trained.

The pseudocode of the proposed attack is described in the Algorithm 1. The proposed adversarial attack can be seen as a two-way mechanism to obtain the final backdoor model. In the first step, the clean model is trained on the clean training images and associated true labels. Once the model is trained, the backdoor images are generated from the network by crafting the adversarial perturbations as favorable triggers. These perturbation triggers are mapped with the desired target labels. Once the backdoor triggered images and target labels are developed, they are augmented with the original training set. Later, the augmented dataset is used to finetune the initial optimized model to make it vulnerable to inference-time backdoor adversarial triggers. We would like to highlight that there is a trade-off between the dual training of the CNN architecture and the high transferability of the proposed attack. The key strengths of the proposed attack are highlighted in the results section of the paper.

Algorithm 1: Proposed Adversarial Backdoor Attack

Input: Training Dataset (X), Training Labels (Y), and CNN Architecture (F)

Output: $F' \leftarrow$ Backdoor CNN

/*Train the CNN Architecture F using X and Y*/

1: The parameters of the network are optimized using the following equation $\tilde{\theta} \leftarrow \theta - \alpha \frac{\delta\theta(x,y)}{\delta\theta}$;

2: Once the network F is optimized on the training data. The adversarial examples are generated by solving the following iterative optimization: $argmin_{\epsilon} J(F(x'), y^{adv})$, s.t. $||x - x'|| \leq \epsilon$;
where, ϵ controls the norm of the perturbation. y^{adv} is the random label not equal to the true label of an image x on which an adversarial example x' is generated.

3: Once the adversarial examples are generated on the entire training set, adversarial examples of different classes are mapped to the different desired target labels. For instance, the adversarial images of class 1 might be mapped to the class label 6.

/*Train the Backdoor CNN Architecture F' using $X + X'$ and Y and Y'*/

4: The original trained model F is then fine-tuned using the combination of clean Training data and adversarial examples along with the true labels and desired backdoor labels. This fine-tuned model is referred to as the backdoor model.

IV. Experimental Setup

In this section, the database and model details used for performance evaluation are summarized.

A. Databases

To evaluate the proposed *IBAttack*, multiple object recognition databases commonly used in adversarial research have been used. CIFAR [17] database is one of the most popular databases to showcase the findings of an attack. The database

TABLE I
CUSTOM BASE CNNS FOR EXPERIMENTS ON FASHION-MNIST.

Layer	Custom-1	Custom-2
Conv	64 Filters of size 8*8	32 Filters of size 5*5
Activation	ReLU	ReLU + Max Pool 3*3
Pool	128 Filters of size 6*6	64 Filters of size 8*8
Activation	ReLU	ReLU + Avg. Pool 3*3
Conv	128 Filters of size 5*5	64 Filters of size 8*8
Activation	ReLU	ReLU + Avg. Pool 3*3
Flatten	Flatten	Flatten
Dense	–	64 Units
Activation		ReLU
Output	10 Units	10 Units

has two variants: CIFAR-10 which comprises 10 classes and CIFAR-100 which comprises 100 categories. Both the databases consist of $50,000$ training images and $10,000$ in testing images. Another database, Fashion-MNIST (F-MNIST) [35] is used which is a gray-scale database of fashion objects such as T-shirts and trousers. It contains 10 classes, with $6,000$ images in each class and a separate testing set of a total of $10,000$ images.

B. Target and Threat Models

A wide range of classification networks is utilized to perform an extensive study. The networks used in the experiments are: (i) VGG-16 [30], (ii) ResNet-50 (RN-50) [14], (iii) Wide-ResNet-16-8 [36] (WRN16-8), (iv) Wide-ResNet-28-10 [36] (WRN28-10), and (v) custom CNNs. The classification networks vary in terms of the number of layers covering both depth and width, types of connections such as sequential to residual, size of filters, and operations. To fully understand the impact, these variations are essential and can provide a complete understanding of the research. The VGG is a sequential network consisting of 16 layers, whereas, ResNet-50 is a 50 layers deep network covering both sequential and residual connections helpful in better parameter learning. Contrary to these networks Wide-ResNet pays attention to both the depth and width of the network and we have utilized two variations of the architecture covering different widths. Apart from that, on the F-MNIST database, two custom models are built as surrogate models to understand the adversarial effect [5], [6], [25], [33]. The configurations of the custom models are given in Table I. All the experiments are performed using a 1080ti GPU machine with 11 GB memory. The codes are written in Keras [8] programming environment with Tensorflow [1] backend. The networks are trained using Adam optimizer [16], the batch size is set to 32, and the initial learning rate is set to 10^{-3}. The deeper networks are trained for 200 epochs; whereas, the custom models are trained for 100 epochs.

C. Implementation Details

Formally, the PGD adversarial attack is a multi-step iterative attack. It is believed that the multiple iterations in generating the perturbation make it a strong adversary in terms of fooling classifiers. In this research, the projection step has been run for 20 times. The magnitude of the perturbation is set to 0.03. To

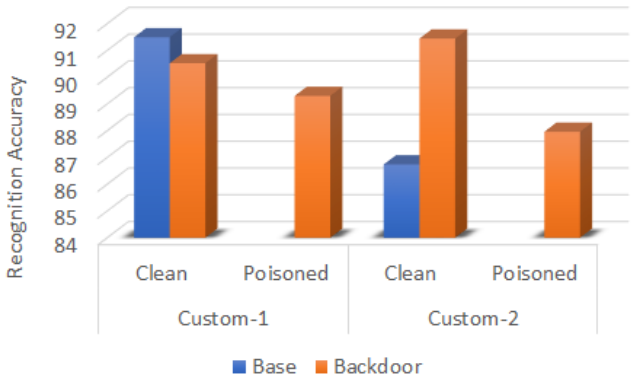

Fig. 3. Results of the proposed *IBAttack* on the F-MNIST dataset.

perform the backdoor attacks, generated adversarial examples on the training set of each database are used. The proposed attack has twofold objectives: the first works towards generating the minute imperceptible backdoor triggers, whereas in the second objective these minutely perturbed images are mapped to the desired target label. The generated adversarial examples are augmented with the clean set of the database and used for training the backdoor vulnerable models. Once the network is trained, imperceptible adversarial noise is added to the test set with the hope that the network will label these backdoor examples to the desired target labels. In each database, images of one class are randomly mapped to the desired target class when the backdoor adversarial noise is added to them. For example, the category 1 in the CIFAR-10 database is mapped to category 4. It is to be noted that the class label of the clean examples is kept as it is to make sure in the absence of a backdoor adversary, the network correctly predicts the original class of the data.

V. RESULTS AND ANALYSIS

This section describes the experiments performed to demonstrate the effectiveness of the proposed adversarial backdoor attack and the observations. The experiments are performed in several conditions reflecting real-world variations.
1. Seen network setting where the attack generator and deployable models are the same, **2 unseen classifier setting** where the backdoor patterns are obtained using another surrogate model, **3. unseen backdoor pattern generation technique** where different pattern generation algorithms are used contrary to the technique used for backdoor network learning, and **4. dually agnostic** where both the model and backdoor generation techniques are unseen.

A. Traditional Settings

The experiments with traditional settings are performed with all three datasets, namely, CIFAR-10, CIFAR-100, and F-MNIST. The results on CIFAR databases are reported in Table II. Four popular CNN classifiers trained on clean images are fine-tuned to make them adversarial backdoor vulnerable. The desired property of the backdoor attack is to achieve high attack success in the presence of a trigger and retain the

TABLE II
RESULTS OF THE PROPOSED *IBAttack* ON THE CIFAR DATASETS. THE BASE MODEL REFERS TO THE MODEL TRAINED ON THE CLEAN TRAINING IMAGES; WHEREAS, THE BACKDOOR MODEL IS FINE-TUNED MODEL THROUGH A COMBINATION OF CLEAN AND ADVERSARIAL BACKDOOR IMAGES. HIGHER VALUES ON THE POISONED DATA REPRESENT BETTER PERFORMANCE. VALUES OF CLEAN DATA SHOULD NOT BE COMPARED WITH THE POISONED DATA. CLEAN DATA VALUES SHOW THE ROBUSTNESS OF THE BACKDOOR MODEL IN THE ABSENCE OF TRIGGERS; WHEREAS, BACKDOOR DATA VALUES SHOW THE PERFORMANCE OF THE ATTACK.

Dataset	CNN	Data	Model	
			Base	Backdoor
CIFAR-10	VGG-16	Clean	83.91	83.74
		Poisoned	–	73.88
		Both	–	78.81
	RN-50	Clean	91.81	87.14
		Poisoned	–	81.50
		Both	–	84.32
	WRN16-8	Clean	93.14	89.41
		Poisoned	–	82.76
		Both	–	86.08
	WRN28-10	Clean	94.17	91.35
		Poisoned	–	86.24
		Both	–	88.79
CIFAR-100	VGG-16	Clean	46.55	44.81
		Poisoned	–	35.12
		Both	–	39.96
	RN-50	Clean	67.31	58.51
		Poisoned	–	40.22
		Both	–	49.36
	WRN16-8	Clean	74.57	66.03
		Poisoned	–	56.95
		Both	–	61.49
	WRN28-10	Clean	76.21	65.97
		Poisoned	–	64.76
		Both	–	65.36

accuracy of a classifier on clean images. The standard/clean VGG network trained naturally yields 83.91% classification accuracy on the CIFAR-10 dataset. When the model is fine-tuned using a combination of the clean and adversarial backdoor trigger, the model retains its accuracy to 83.74% on clean images. It shows that the proposed backdoor attack does not affect the normal functioning of the network. Whereas, for the backdoor examples associated with desired target labels, the backdoor model gives 73.88% correct classification accuracy. This shows that the proposed *IBAttack* is effective in fooling the classifier without affecting the normal performance. It is to be noted that here that the results are reported on a complete test set of $10,000$ images contrary to existing pair-based backdoor attacks.

A similar observation has been noticed in other classifiers including ResNet and Wide-ResNets. Contrary to VGG, the backdoor attack is found to be highly successful and the attack recognition rate improves to 86.24% (WRN28-10). The recognition performance on the clean images shows a drop in the range of 3% to 4%. The existing attacks also show a significant drop in the recognition performance on the clean images, while they are only evaluated for specific pair mapping [12], [28]. The VGG network when used for the classification of the CIFAR-100 dataset, it achieved 46.55% normal recognition accuracy which shows a slight drop when the model is transformed to perform the backdoor attack. The backdoor target mapping accuracy of 35.12% is achieved on VGG which can increase to 64.76% using the deeper and wider classifier, i.e., WRN28-10.

CIFAR databases are color object databases and are rich in image features, the F-MNIST is a gray-scale database with different characteristics. The analysis on the F-MNIST dataset is reported in Fig. 3. For F-MNIST, two custom models described in Table I have been used. Custom-1 model yields 91.49% recognition accuracy when the model is trained using only the clean images. The network shows a drop of 0.9% in achieving its backdoor success of 89.3%. Interestingly, the proposed backdoor attack shows an improvement in the recognition performance on clean images. The other networks on each previous database sacrifice their performance in achieving the backdoor success. The standard custom-2 model yields an accuracy of 86.74% on clean test images which got improved to 91.44% when the model is prepared for backdoor attack. In the attack success case, the proposed attack can get 87.96% recognition performance.

B. Generalized Settings

In the generalized setting, first, the impact of the network is evaluated. For example, it might be possible that once the attacker provides the backdoor model to the customer, an attacker might lose access to it for the generation of a backdoor adversary. Therefore, in such a scenario an attacker should be able to achieve its success by generating the adversary on a surrogate model and deploying it for its success. For this, two classifiers namely VGG and WRN-28-10 are used where at a time one model acts as a backdoor model, and another model is used as a surrogate to generate the backdoor adversary.

On the CIFAR-10 database when the WRN-28-10 is used as the backdoor model and the backdoor adversary is generated using VGG, the attack recognition accuracy of 89.64% is achieved. Surprisingly, the attack success is even higher in comparison to the seen model backdoor adversary. A similar backdoor transfer success is observed when the WRN28-10 is used as a surrogate and the VGG is a backdoor vulnerable model. In brief, except in the case of the WRN28-10 surrogate backdoor adversary on the VGG model on the CIFAR-100, the proposed backdoor adversary shows high transfer success. In this unusual case, the backdoor attack is found 13% less successful. However, the setting has the advantage of improving the performance on the clean test set of CIFAR-100, which shows an improvement of 1.5% compared to seen network backdoor adversary.

In the second generalized setting, we have varied the algorithm used for learning the backdoor adversary. When learning the backdoor model, the PGD algorithm is used; whereas, to perform the attack, the FGSM attack proposed by Goodfellow et al. [11] has been applied. FGSM is a single-step gradient adversary which utilizes the sign of the gradient as the perturbation. The experiments are performed using each database to extensively evaluate the strength of the proposed backdoor adversary concept and make it a perfect fit for real-world deployment.

On the CIFAR-10 dataset, when the VGG16 network is used as the backdoor model under the influence of PGD adversary and the FGSM backdoor adversary generated using the same

TABLE III
THE SUCCESS OF BACKDOOR ATTACK IN 'DUALLY' AGNOSTIC SETTINGS, I.E., WHERE BACKDOOR ADVERSARY ALGORITHM AND CLASSIFIERS ARE DIFFERENT. THE SUCCESS IS REPORTED IN TERMS OF RECOGNITION ACCURACY ON EACH DATASET (DS). FM REPRESENTS THE F-MNIST DATASET. THE HIGHER THE VALUES ON THE POISONED DATA, THE BETTER THE ATTACK. VALUES OF CLEAN DATA SHOULD NOT BE COMPARED WITH THE POISONED DATA. CLEAN DATA VALUES SHOW THE ROBUSTNESS OF THE BACKDOOR MODEL IN THE ABSENCE OF TRIGGERS; WHEREAS, BACKDOOR DATA VALUES SHOW THE PERFORMANCE OF THE ATTACK.

DS	Backoor CNN	Clean Accuracy	Surrogate CNN	Data	
				Clean	Backdoor
CIFAR-10	VGG-16	83.91	WRN28-10	66.57	64.19
			WRN16-8	71.93	59.63
	RN-50	91.81	VGG16	86.18	75.22
			WRN28-10	88.41	53.68
	WRN16-8	93.14	VGG16	87.08	77.94
			WRN28-10	89.37	56.06
	WRN28-10	94.17	VGG16	92.33	82.82
			WRN16-8	90.35	61.41
CIFAR-100	RN-50	67.31	VGG16	59.67	37.10
			WRN28-10	59.95	30.00
	WRN16-8	74.57	VGG16	64.82	53.04
			WRN28-10	61.26	42.34
	WRN28-10	76.21	VGG16	67.64	56.75
			WRN16-8	68.12	51.73
FM	Custom-1	91.49	Custom-2	90.35	40.29
	Custom-2	86.74	Custom-1	91.30	29.60

model is used for attack, more than 66% images are classified correctly into the desired target categories. The above observation has been noticed against each network which shows that even if the backdoor adversary generation algorithm is different, the attack can be significantly successful. The attack success is not limited to CIFAR-10 only; however, even on other datasets such as CIFAR-100 and F-MNIST, the success of the proposed backdoor attack is observed. It shows that the proposed attack is not overfitted to any dataset or classifier and generation algorithm.

We have further complicated an evaluation setting, where not only the attack generation algorithm is varied but also the generating network. This setting is referred to as the 'dually' agnostic nature of the proposed attack. The results of this setting are reported in Table III. In brief, the results can be described as follows: ❶ the proposed backdoor adversarial attack is agnostic to classifier and generation network as well, ❷ on each backdoor network on color databases, the backdoor adversary of VGG is found most robust, ❸ in terms of accuracy on clean images, VGG sacrifices its natural accuracy higher as compared ResNet and Wide-ResNets, and ❹ the proposed attack is found less robust on gray-scale images under such extreme generalized setting.

C. Comparisons

As mentioned earlier, the work proposed by Saha et al. [28] is the closest to the proposed *IBAttack*. Some of the limitations of Saha et al. [28] is that it is evaluated using only a single network and is highly sensitive to the amount of fine-tuning in the backdoor model. For example, fine-tuning even only more fully connected layers, the attack success drops at least 15%; In the proposed attack, a classifier is fine-tuned from start to end, and even the proposed attack retains its fooling success.

TABLE IV
COMPARING DIFFERENT METHODS AGAINST RESNET-18 ON THE RANDOM SUBSET OF **IMAGENET** (1000 IMAGES OF 50 CLASSES) WHEN THE TRIGGER IS GENERATED USING VGG-16. ASR IS ATTACKED SUCCESS RATE, BA IS BENIGN/CLEAN ACCURACY, AND PSNR IS A PEAK SIGNAL-TO-NOISE RATIO.

Attack	Effectiveness		Stealthiness
	BA	ASR	PSNR
No attack	91.9	–	–
BadNets [12]	87.6	34.18	28.93
Blended [7]	85.9	29.67	48.72
SST [21]	87.7	68.15	32.16
Refool [23]	89.7	56.82	28.19
Proposed	**90.3**	**88.92**	**21.16**

Apart from that, the success of the attack is at least 12% better than the hidden trigger by Saha et al. [28]. Another significant advantage of the proposed backdoor attack is that even at the time of inference, the backdoor trigger is not visible which is visible in the comparative work.

Further, the comparison with the existing attacks in terms of benign accuracy (BA), attack success rate (ASR), and stealthiness is provided. The comparative results shown in Table IV reflect that the proposed invisible adversarial noises can also serve as high strength triggers without being perceptible the humans. While perceptible attacks such as BadNets and Blended attacks are easy to implement, their success is low. The prime limitation of most of the existing attacks is the mapping of a backdoor pattern whether visible or invisible with the desired target label. It makes the attacking strategy limited; hence, can not be used where the targeted attack categories are multiple. Whereas, in the proposed backdoor attack where we have not constrained the proposed attack to any specific label and mapped different classes with imperceptible backdoor patterns with different target classes. Therefore, the proposed attack in multiple desired target categories of backdoor attack yields significantly higher ASR than the state-of-the-art existing attacks. While the ASR is high, the stealthiness of the proposed attack is also low to further demonstrate the strength of the proposed attack.

D. Transferability of Adversarial Perturbations and Proposed Backdoor

To showcase that the high dependency on data labels is the primary cause of backdoor attacks, we have performed experiments with adversarial data augmentation and its success as a backdoor attack. Two adversarial attacks which are chosen to perform the proposed backdoor attack are used for a fair comparison. In the case of adversarial augmentation, label consistency has been maintained [32]. In other words, the adversarial images are consistent with the label of the original images while the noise is generated in the targeted attack fashion. The targeted attack is defined as the mapping of the data into the desired class. When the label is consistent [32] with the images, the accuracy of the attack in an unseen classifier setting is significantly lower than the proposed backdoor attack. The *IBAttack* follows the traditional/existing backdoor setting where the label is not consistent with the content of

TABLE V
COMPARING THE *IBAttack* ATTACK WITH THE ADVERSARIAL ATTACKS TO PRODUCE A STRONG BASELINE AND NEED OF THE PROPOSED STUDY. THE RESULTS IN TERMS OF BACKDOOR RECOGNITION ACCURACY ARE DEMONSTRATED BY TRANSFERRING POISONING SAMPLES TO DIFFERENT MODEL ARCHITECTURES.

Dataset	Surrogate CNN	Target CNN	PGD	FGSM	Proposed
CIFAR10	VGG16	WRN28-10	26.78	17.89	**64.19**
		WRN16-8	31.45	21.16	**59.63**
	WRN16-8	VGG16	14.92	7.90	**77.94**
		WRN28-10	34.78	12.87	**56.06**
	WRN28-10	VGG16	6.54	2.78	**82.82**
		WRN16-8	26.02	11.90	**61.41**
CIFAR100	RN-50	VGG16	4.73	2.36	**37.10**
		WRN28-10	6.26	4.57	**30.00**
	WRN16-8	VGG16	8.90	6.93	**53.04**
		WRN28-10	22.20	10.16	**42.34**
	WRN28-10	VGG16	14.47	7.63	**56.75**
		WRN16-8	27.38	12.38	**51.73**

TABLE VI
RESULTS OF THE PROPOSED BACKDOOR ATTACK UNDER VARYING THE NUMBER OF POISONED SAMPLES IN THE TRAINING. FOR EXAMPLE, $n\%$ REPRESENTS ONLY $n\%$ IMAGES DRAWN FROM THE TRAINING SET ARE USED FOR BACKDOOR ATTACKS. THE ATTACK SUCCESS IS REPORTED IN TERMS OF RECOGNITION ACCURACY ON EACH DATASET.

Poisoned Examples (%)	CIFAR10		CIFAR100	
	WRN16-8	WRN28-10	WRN16-8	WRN28-10
20	74.67	79.16	51.28	58.10
40	77.15	82.90	52.79	61.67
60	80.09	83.17	53.53	63.03
80	82.46	84.56	54.19	63.98
100	82.76	86.24	56.95	64.76

an image. The results of this experiment are added in Table V and the results show how transferable it is in getting the desired target label.

E. Impact of Size of Backdoor Images

Another critical study in the backdoor attack setting is the study of attack success rate based on the amount of poisoned labeled data. For that, we have performed an ablation study by varying the number of backdoor images used for training and the results are reported in Table VI. The results suggest that the proposed backdoor is resilient to the number of poisoned samples used and still effective even in the presence of a low number of backdoor images. Other than that, we want to highlight that the proposed attack is a multi-label backdoor attack as compared to a traditional pair-label attack. The proposed attack has also been performed using only 500 and 1,000 images for the pair-label attack. In the pair label attack, class 7 is used as the target category of the CIFAR databases in the presence of a backdoor trigger. The attack success of the proposed attack for 10% images is 82% which increases to 91% when 20% images are used. The success is found significantly higher than existing attacks shown above.

From the side of protecting the integrity of deep networks against backdoor, this study showcases the need not entirely rely upon label supervision. When the amount of labeled poisoned samples reduces, the success rate of the attacks shows a reduction. Therefore, we believe, the use of semi-supervised learning or entirely unsupervised training of a network can reduce the impact of backdoor attacks to a great extent if not completely removed.

TABLE VII
RESILIENCY OF THE PROPOSED *IBAttack* IN THE PRESENCE OF SEVERAL STATE-OF-THE-ART DEFENSE MECHANISMS INCLUDING ADVERSARIAL TRAINING (AT), INPUT TRANSFORMATIONS (IT), AND JPEG COMPRESSION. THE PROPOSED ATTACK AS EXPECTED ARE FOUND TO BE RESILIENT AGAINST THE DEFENSE MECHANISMS. ORIGINAL REFERS TO THE BACKDOOR ATTACK ON THE UNDEFENDED MODEL. B. MODEL REPRESENTS THE BACKDOOR MODEL, AND S. MODEL IS THE SURROGATE MODEL.

Dataset	B. Model	S. Model	Original	AT	IT	JPEG
CIFAR-10	RN50	VGG-16	75.22	67.26	76.98	74.56
		WRN28-10	53.68	54.68	52.67	53.69
	WRN16-8	VGG-16	77.94	69.48	75.89	78.10
		WRN28-10	56.06	56.98	56.07	57.16
CIFAR-100	RN50	VGG-16	37.10	33.25	36.59	36.99
		WRN28-10	30.00	31.52	28.12	28.72
	WRN16-8	VGG-16	53.04	47.12	49.35	55.87
		WRN28-10	42.34	42.39	40.63	42.89

F. Resiliency

An extensive study utilizing various defense mechanisms has been performed to establish that the proposed attack is challenging to defend. While PGD adversarial training has shown some improvement, it is not significant or not completely mitigating the effect of the proposed attack. On top of that, we should be aware and it is well established in the literature that adversarial training (AT) does not generalize against unseen attacks, is computationally complex, and reduces the performance of clean images [31], [37]. Other than that, vulnerability or inefficiency of the input transformation (IT) and JPEG compression-based defenses have also been reported. The results of this analysis are reported in Table VII.

Interestingly, while these defenses show a slight reduction in the success of the proposed attack, it has been observed that these images are not always classified into the true class as well. In other words, while these backdoor images are not getting classified into the target class an attacker wants but it is also not getting correctly classified into their true class. The images are getting classified into random classes (not equal to true class or target backdoor attack class).

G. Limitations

The proposed backdoor adversarial attack has the following limitations: (i) lower robustness on the gray-scale images and (ii) slight drop in the recognition accuracy on clean images. In contrast to the existing backdoor attacks, the proposed attack requires training of the backdoor twice. However, this limitation is not severe because during training the third party can train the network in multiple rounds. The proposed attack has multiple generalizability features, including transferability (Table III) which alleviates this limitation.

VI. CONCLUSION

The vulnerability of deep learning-based classifiers against adversarial perturbations has significantly demanded the defense against them. From the literature, it is observed that

adversarial training through the augmentation of adversarial examples is the most effective defense. In the adversarial training setting, the labels of adversarial images are intact to make sure the model learns the robust decision boundary even in the presence of an adversary. Utilizing the above concept due to the desire for defense against adversarial attacks, we have proposed a novel backdoor adversarial attack. In the proposed attack, adversarial patterns are first generated and associated those images with the desired target class labels. Similar to traditional backdoor attacks, these backdoor adversaries are augmented with clean images and fine-tune the target classifiers to make them backdoor vulnerable. The extensive experiments performed using multiple databases and several deep classifiers showcase the effectiveness of the proposed backdoor adversarial attack. The attack is evaluated under several challenging conditions to reflect that the proposed attack is real-world ready and hence challenging. Apart from that, the proposed adversary has a significant advantage of being hidden both at the time of training and inference. The identification of the proposed backdoor adversary will further strengthen the research directions toward improving the robustness of deep classifiers.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. {TensorFlow}: A system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[2] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Cognitive data augmentation for adversarial defense via pixel masking. *Pattern Recognition Letters*, 146:244–251, 2021.

[3] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Intelligent and adaptive mixup technique for adversarial robustness. In *IEEE International Conference on Image Processing*, pages 824–828, 2021.

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *Annual Conference on Neural Information Processing Systems*, 2020.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[6] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI Conference on Artificial Intelligence*, pages 10–17, 2018.

[7] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[8] François Chollet, JJ Allaire, et al. R interface to keras. https://github.com/rstudio/keras, 2017.

[9] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020.

[10] Chengyue Gong, Tongzheng Ren, Mao Ye, and Qiang Liu. Maxup: Lightweight adversarial training with data augmentation improves neural network training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2474–2483, 2021.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[15] Sara Kaviani and Insoo Sohn. Defense against neural trojan attacks: A survey. *Neurocomputing*, 423:651–667, 2021.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[18] Shaofeng Li, Hui Liu, Tian Dong, Benjamin Zi Hao Zhao, Minhui Xue, Haojin Zhu, and Jialiang Lu. Hidden backdoors in human-centric language models. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 3123–3140, 2021.

[19] Shaofeng Li, Minhui Xue, Benjamin Zhao, Haojin Zhu, and Xinpeng Zhang. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 2020.

[20] Yiming Li, Tongqing Zhai, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Rethinking the trigger of backdoor attack. *arXiv preprint arXiv:2004.04692*, 2020.

[21] Li et al. Invisible backdoor attack with sample-specific triggers. In *IEEE/CVF ICCV*, 2021.

[22] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[23] Liu et al. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*. Springer, 2020.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[25] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597, 2016.

[26] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.

[27] Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. Mind the style of text! adversarial and backdoor attacks based on text style transfer. *arXiv preprint arXiv:2110.07139*, 2021.

[28] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 11957–11965, 2020.

[29] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. *arXiv preprint arXiv:2003.03675*, 2020.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

[33] Xiao Wang, Siyue Wang, Pin-Yu Chen, Yanzhi Wang, Brian Kulis, Xue Lin, and Peter Chin. Protecting neural networks with hierarchical random switching: Towards better robustness-accuracy trade-off for stochastic defenses. *International Joint Conference on Artificial Intelligence*, pages 6013–6019, 2019.

[34] Casimir Wierzynski. The challenges and opportunities of explainable ai. *Intel. com*, 12, 2018.

[35] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[36] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[37] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. *arXiv preprint arXiv:1901.04684*, 2019.

[38] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.

[39] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*, pages 15–26, 2021.

[40] Haoti Zhong, Cong Liao, Anna Cinzia Squicciarini, Sencun Zhu, and

David Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. In *ACM Conference on Data and Application Security and Privacy*, pages 97–108, 2020.

Akshay Agarwal received his doctoral degree (PhD.) from IIIT-Delhi, India in 2020. He is currently a postdoctoral scientist at the University at Buffalo, NY, USA. He has also worked as a research assistant professor at Texas A&M University, Kingsville, USA from Dec. 2019 to Dec. 2020. His Ph.D. thesis titled "Panoptic Defenses for Secure Computer Vision" has won the IEEE Biometrics Council Best Doctoral Dissertation Award 2021. His research interests are machine learning and deep learning with applications in the security of biometrics and computer vision algorithms from several anomalies. He has authored over fifty publications including, journals and peer-reviewed conferences. He is on the program committee of several top-listed research conferences including, ICML, CVPR, and NeurIPS along with the reviewer of top journals such as IEEE TIFS, IEEE TNNLS, and PR.

Richa Singh received the M.S. and Ph.D. degree in computer science from West Virginia University, Morgantown, USA. She is currently a Professor at IIT Jodhpur, India, and an Adjunct Professor with IIIT-Delhi and West Virginia University, USA. Her areas of interest are pattern recognition, machine learning, and biometrics. She is a Fellow of IAPR and a Senior Member of IEEE and ACM. She was a recipient of the Kusum and Mohandas Pai Faculty Research Fellowship at the IIIT-Delhi, the FAST Award by the Department of Science and Technology, India, and several best paper and best poster awards in international conferences. She is/was the Program Co-Chair of CVPR2022, ICMI2022, IJCB2020, FG2019 and BTAS 2016, and a General Co-Chair of FG2021 and ISBA 2017. She is also the Vice President (Publications) of the IEEE Biometrics Council and an Associate Editor-in-Chief of Pattern Recognition.

Mayank Vatsa received the M.S. and Ph.D. degrees in Computer Science from West Virginia University, USA. He is currently a Professor and Dean of Research and Development with IIT Jodhpur, India, and the Project Director of the Technology and Innovation Hub on Computer Vision and Augmented & Virtual Reality under the National Mission on Cyber Physical Systems by the Government of India. He is also an Adjunct Professor with IIIT-Delhi, India and West Virginia University, USA. His areas of interest are biometrics, image processing, machine learning, computer vision, and information fusion. He is the recipient of the prestigious Swarnajayanti Fellowship from the Government of India, the A. R. Krishnaswamy Faculty Research Fellowship at the IIIT-Delhi, and several best paper and best poster awards at international conferences. He is an Area/Associate Editor of Information Fusion and Pattern Recognition, the General Co-Chair of IJCB 2020, and the PC Co-Chair of IEEE FG2021. He has also served as the Vice President (Publications) of the IEEE Biometrics Council where he led the IEEE Transactions on Biometrics, Behavior, And Identity Science.

Nalini Ratha received the BTech degree in electrical engineering and the MTech degree in computer science and engineering both from IIT Kanpur, Kanpur, India, and the PhD degree in computer science from Michigan State University, East Lansing, Michigan. He is an empire innovation professor of computer science and engineering with the University at Buffalo (UB), State University of New York, Buffalo, New York. He has authored more than 100 research papers in the area of biometrics and has been co-chair of several leading biometrics conferences and served on the editorial boards of IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, and IEEE Transactions on Image Processing and the Pattern Recognition journal. He has co-authored a popular book on biometrics entitled Guide to Biometrics and also co-edited two books entitled Automatic Fingerprint Recognition Systems and Advances in Biometrics: Sensors, Algorithms and Systems. He has offered tutorials on biometrics technology at leading IEEE conferences and also teaches courses on biometrics and security. He is a fellow of the IAPR and an ACM distinguished scientist. During 2011–2012, he was the president of the IEEE Biometrics Council. He was awarded the IEEE Biometrics Council Leadership Award, in 2019.