SSDA: Secure Source-Free Domain Adaptation

Sabbir Ahmed[1*], Abdullah Al Arafat[2*], Mamshad Nayeem Rizve[3*], Rahim Hossain[1],
Zhishan Guo[2], Adnan Siraj Rakin[1]
[1]Binghamton University (SUNY), [2]North Carolina State University, [3]University of Central Florida

Abstract

Source-free domain adaptation (SFDA) is a popular unsupervised domain adaptation method where a pre-trained model from a source domain is adapted to a target domain without accessing any source data. Despite rich results in this area, existing literature overlooks the security challenges of the unsupervised SFDA setting in presence of a malicious source domain owner. This work investigates the effect of a source adversary which may inject a hidden malicious behavior (Backdoor/Trojan) during source training and potentially transfer it to the target domain even after benign training by the victim (target domain owner). Our investigation of the current SFDA setting reveals that because of the unique challenges present in SFDA (e.g., no source data, target label), defending against backdoor attack using existing defenses become practically ineffective in protecting the target model. To address this, we propose a novel target domain protection scheme called secure source-free domain adaptation (SSDA). *SSDA adopts a single-shot model compression of a pre-trained source model and a novel knowledge transfer scheme with a spectral-norm-based loss penalty for target training. The proposed static compression and the dynamic training loss penalty are designed to suppress the malicious channels responsive to the backdoor during the adaptation stage. At the same time, the knowledge transfer from an uncompressed auxiliary model helps to recover the benign test accuracy. Our extensive evaluation on multiple dataset and domain tasks against recent backdoor attacks reveal that the proposed SSDA can successfully defend against strong backdoor attacks with little to no degradation in test accuracy compared to the vulnerable baseline SFDA methods. Our code is available at https://github.com/ML-Security-Research-LAB/SSDA.*

1. Introduction

Deep Neural Networks (DNNs) have shown remarkable success across a multitude of tasks [18, 8, 12, 9, 5, 37, 2, 15, 26]. In particular, they have exhibited extraordinary performance in various visual tasks such as classification [18], object detection [8], and semantic segmentation [12]. Even though DNNs have achieved great success in various visual tasks, they heavily depend on the underlying distribution of training data. Unfortunately, DNNs deployed in real-world scenarios, such as those utilized in autonomous vehicles, frequently encounter new situations, such as varying weather conditions [1] and changing illumination levels [36]. Consequently, the machine learning community has increasingly directed their attention towards domain adaptation [4, 27, 22] concept.

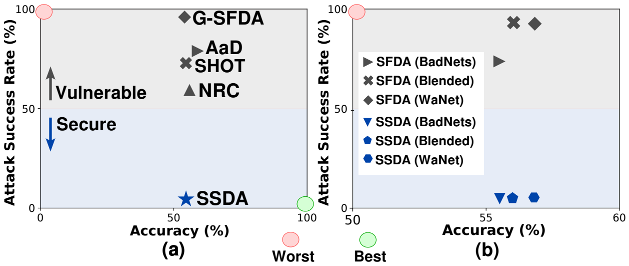

Figure 1: *Performance of a) existing SFDA [24, 42, 43, 44] methods and SSDA against one attack [10], and b) one competing SFDA [24] and SSDA against popular backdoor attacks [10, 3, 28]. Compared to the vulnerable previous SFDA, the proposed SSDA is highly secure.*

Existing *Domain Adaptation* (DA) setup requires access of the source domain dataset for target domain adaptation. However, due to increasing privacy concerns and a lack of available source data, these existing DA methods are becoming impractical. To address this, researchers have introduced a new setup for domain adaptation, popularly known as source-free domain adaptation (*SFDA*), which aims to transfer knowledge from a prior domain (i.e., *source*) to a new domain (i.e., *target*) w/o accessing the source dataset. Moreover, recent SFDA works [24, 43, 6, 19] have also considered the practical constraint of limited labeled data in the real world and performed domain adaptation without any labeled data in the target domain dataset. Therefore, the two primary constraints in SFDA setting are that the target domain is denied *access to the source dataset during adap-*

[*] These authors contributed equally

tation and that the *target domain training is unsupervised, i.e., without labeled data*.

In this work, we are the first to focus on the security of the target domain in SFDA. In particular, under the SFDA setting, the target domain owner can not access the source dataset and is also completely unaware of the training process of the source model. Such a setting makes the target domain adaptation extremely vulnerable to adversaries considering the adversary can access source training (as shown in Fig. 2). After training, the target domain owner takes the pre-trained malicious source model and then adapts it to a new target domain dataset without any label, utterly unaware of the consequences of malicious source training.

The security threat we want to investigate in this work is the Backdoor/Trojan attack [10, 28]. In a backdoor attack, an attacker poisons a subset of the training data using a specific trigger (i.e., input pattern) to train the model. During inference, the model functions accurately under normal circumstances (i.e., no attack scenario). However, when the attacker-designed particular trigger pattern appears in the input, the model fails as intended by the attacker. In the case of SFDA, the backdoor attack becomes more relevant as the source owner can inject the backdoor into the source model w/o any defensive measure from the target owner (who has no access to source data/training). *To the best of our knowledge, no prior works have explored the vulnerability of the target model against backdoor attacks, considering a malicious source domain owner (attacker).*

However, given the above setting, defending against backdoor attacks during the target domain adaptation is extremely challenging because of the unique challenges presented by the SFDA setting. *First*, the target model is initialized using the malicious source model which is already infected with backdoor. *Second*, the target owner cannot access the source dataset used initially to inject the backdoor. As a result, the target owner cannot trivially fine-tune the model using the source dataset for Trojan removal. *Finally*, the target training is unsupervised, making detecting/cleaning backdoor-infected models more challenging. *Because of these unique challenges in SFDA, existing backdoor defenses [23, 39, 11] are practically ineffective in defending the target model from a source adversary.*

Our initial investigation of backdoors in SFDA (in Table 2) confirms that even after benign training in the target domain, the target model remains vulnerable to the attacker's designed triggered inputs. To make it worse, the threat remains persistent even after applying strong backdoor defenses [47] in the current SFDA setting (in Table 3). Hence, we are the first to validate that ***current SFDA techniques are not safe against the backdoor attack, and none of the existing defenses can protect the target domain model***.

To address this issue, we propose a novel target domain training scheme called *Secure Source-Free Domain Adaptation, SSDA*, the *first successful defense* against backdoor attacks tailored for source-free domain adaptation. Our proposed novel training method SSDA consists of two key components: first, it performs a single-shot static defensive compression of the source model. It uses spectral norm-based ranking to remove (i.e., setting the weight values to zero) malicious channels contributing to the successful transfer of backdoor attacks from the pre-trained source model. However, the static compression before target training may lead to information loss, leading to poor performance [29] (i.e., lower accuracy) in the target domain due to domain shift. Since we do not have labels in the target domain, we cannot afford to lose any information from the pre-trained source model. In addition, we need a dynamic defense component to take advantage of the target training and target data to the defender's benefit, which the existing backdoor defenses fail to resolve. Hence, the second component of the SSDA performs knowledge transfer from an auxiliary (uncompressed) model to recover benign accuracy. For more secure knowledge transfer, we propose a novel spectral norm-based loss penalty to suppress malicious channels sensitive to backdoors. However, computing the spectral norm penalty during training is expensive. Hence, we derive a computationally efficient safe upper bound of the norm and then use this novel upper bound to compute the approximate spectral norm efficiently and effectively during training. We extensively evaluated our defense across multiple datasets, tasks and attack combinations. It shows that the proposed SSDA can successfully defend (drops to $\sim$1% from $\sim$ 99% attack success rate) against strong backdoor attacks (e.g., WaNet [28]) with little to no accuracy degradation compared to existing vulnerable SFDA methods (as demonstrated in Fig. 1). Finally, we show that regardless of the source model being benign/malicious, our proposed SSDA can successfully perform the adaptation of target domain.

2. Background

2.1. Prior Works

Source Free Domain Adaptation: The objective in Traditional Unsupervised Domain Adaptation (UDA) [16, 17, 41, 25] is to utilize the knowledge obtained from a source domain to achieve classification in an unlabeled target domain. However, a significant limitation of traditional UDA approaches is that they need access to the source data during target model training. In real-world scenarios, access to the source data is often restricted [34] due to concerns related to data privacy or memory constraints on small devices. Thus, researchers have developed a popular form of UDA known as Source-free domain adaptation (SFDA) to resolve this issue. SFDA carries out domain adaptation by relying

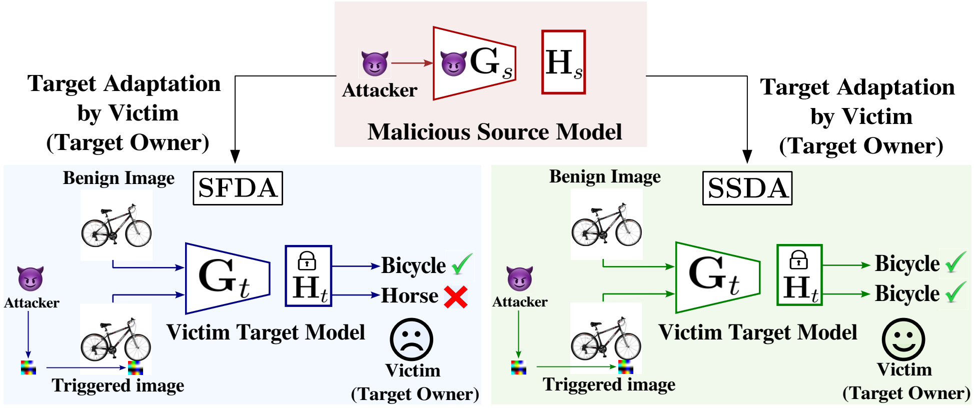

Figure 2: *A malicious source adversary can inject hidden behavior into the source model. Then the target owner performs benign adaptation using this malicious source model, leaving the target model vulnerable to attack at the inference phase (on the left). In contrast, our proposed SSDA effectively mitigates this threat (on the right).*

solely on the pre-trained source model without needing access to the source data. Previous studies [20, 21] suggest that the pre-trained source model already encompasses sufficient knowledge regarding the source feature distribution. Consequently, many variants [24, 43, 6, 19] of SFDA have emerged that utilize the pre-trained model's knowledge to resolve domain shift problem. While SFDA methods eliminate the concerns related to source data privacy, they fail to address the security concerns in the target domain.

Backdoor attacks and defenses: Backdoor attacks [10, 3, 28, 40, 31] involve injecting a trigger, typically an image patch chosen by the attacker, into the training data. When trained on this data, the model learns to associate the trigger with a particular target category, poisoning the model. Although the model works well with clean data, it malfunctions when the attack's specific trigger is present during testing. In particular, Backdoor attacks pose a severe security risk in the SFDA setting, where the target domain owner cannot access the source dataset and cannot control the source model's training process. Thus an attacker from the source owner side can easily access the source training to poison the source data to create a backdoor source model, resulting in a similar backdoor target model after SFDA. Given this major security risk, it is pivotal to consider this security threat while adapting the source model.

Although there are existing backdoor defenses in the literature [11, 35, 7, 23, 39, 46, 48], these approaches have limitations that make them unsuitable for direct application in SFDA. Current Backdoor defense methods can be broadly categorized into two groups: training-based defenses [11, 35, 7] and model post-processing-based defenses [23, 39, 46]. The former suppresses the influence of backdoor triggers or eliminates poisoned data during training and thus requires full access to both clean and poisoned data. However, in existing SFDA, training-based defenses cannot be adopted since the target owner cannot access source training/data. The latter fine-tunes the backdoored model with a small subset of clean data and eliminates the backdoor threat from the model. However, post-processing-based defenses are limited by the need for labels for target instances and the lack of available source data. As a result, although these methods exhibit potential in defending against backdoor attacks in DNN models, they do not directly fit into the SFDA setting to prevent the transfer of backdoor attacks.

These limitations of the two primary track of defense methods have motivated us to investigate data-free techniques for mitigating backdoor attacks, which would be an ideal candidate defense in SFDA. To our knowledge, Channel Lipschitzness-based Pruning (CLP) [47] is the first and only data-free backdoor defense method. CLP prunes channels in a malicious model that are more sensitive to backdoor triggers than normal channels, based on channel Lipschitz constant. Hence, CLP is a static post-training defense method. However, using CLP in SFDA requires answering several critical questions i) When to perform the pruning (i.e., after source or target training)? ii) Apart from channel pruning, how to leverage the target data/training as a defense tool for the defender? In Section 4, we demonstrate that CLP fails to address these questions for SFDA and becomes practically ineffective against backdoor attack under the SFDA setting.

2.2. Preliminaries and Notations of SFDA

Consider the source domain dataset with n_s labeled samples denoted by $\mathcal{D}_s = \{(x_s^i, y_s^i)\}_{i=1}^{n_s}$ and target domain dataset with n_t unlabeled samples denoted by $\mathcal{D}_t = \{x_t^i\}_{i=1}^{n_t}$. In SFDA scenario, we have access to the source model $\mathbf{F}_s(\cdot)$, which is trained on $\mathcal{D}_s$ in a supervised manner. The source model $\mathbf{F}_s(\cdot)$ is the composition of two modules:

Table 1: *The list of information the attacker (source owner) and the victim (target owner) can access.*

Information	Source Owner (Attacker)	Target Owner (Victim)
Source Training	✓	✗
Source Model	✓	✓
Source Data	✓	✗
Source Label	✓	✗
Target Training	✗	✓
Target Model	✗	✓
Target Data	✗	✓
Target Label	✗	✗

the feature encoding module $\mathbf{G}_s$ and the classifier module $\mathbf{H}_s$, i.e., $\mathbf{F}_s = \mathbf{H}_s \circ \mathbf{G}_s$. For training the target model $\mathbf{F}_t(\cdot)$, only data $\mathcal{D}_t$ and the pre-trained source model $\mathbf{F}_s(\cdot)$ is available, and no data in $\mathcal{D}_s$ can be used.

3. Threat Model

In this work, we consider a threat model wherein the attacker possesses the source domain (i.e., source model and source dataset). In contrast, the victim has the target domain (i.e., target model and dataset w/o labels). As shown in Table 1, the attacker can access every component of the source domain (e.g., training, data, model, label). Specifically, the attacker leverages access to the source model training process to poison a subset of the source dataset, creating a backdoored model. Eventually, the attacker aims to attack the target model at the inference stage using the input trigger patterns generated during the source training stage. On the other hand, the victim, who owns the target domain, has access to the target domain training (e.g., model, data). However, the victim/defender can not access any source dataset and target data label. Hence, the defender's goal is to develop a domain transfer process starting from a malicious source model to improve the security of the target model at the inference stage.

Considering this strong threat model, where the attacker can easily inject a backdoor during source training without any defensive measure from the defender (no access to source training), the victim should take defensive measures during this adaptation process to eliminate the risk. However, given this threat model, defending against attacks in the SFDA setting presents significant challenges for the defender. *First, the target domain model is initialized using the source domain architecture and pre-trained weights, leaving it vulnerable to attacks from the source domain. Second, the absence of the source domain dataset restricts the ability to analyze the pre-trained source model for detecting and mitigating attacks. Third, training the target domain model without labeled data poses additional difficulties in identifying and mitigating threats.* To address these limitations, our goal in this work is to develop a novel secure training scheme for the target domain designed to eliminate the risk of backdoor attacks in SFDA.

Table 2: *This table illustrates a successful backdoor attack in SFDA, where the trojan is transferred from a source model to a target domain of the Office-Home [38] dataset.*

Attack	Rw		$Rw \rightarrow Ar$	
	ACC	ASR	ACC	ASR
BadNets [10]	86.24	100.00	74.21	99.59
Blended [3]	86.01	100.00	74.04	98.06
WaNet [28]	86.70	100.00	74.21	99.88

4. Why Current SFDA Methods are not Secure?

Before introducing the details of our proposed SFDA technique, in this section, first, we demonstrate the problem (i.e., security threat) that our work aims to address. Here, we for the first time, empirically demonstrate the threat of backdoor attacks in existing SFDA methods.

Attack Formulation. Here, we consider an attacker who trains the source model, denoted as $\mathbf{F}_s(\cdot)$, using the source dataset $\mathcal{D}_s$. To execute a backdoor attack, the attacker poisons a subset of the source dataset $\mathcal{D}_s$ with a specific input pattern known as a 'trigger' at a ratio of ρ, which represents the proportion of poisoned training samples. During source model training, each clean pair (x_i^s, y_i^s) in the subset is substituted with a backdoor pair $(\mathcal{B}(x), c(y))$, where $\mathcal{B}(\cdot)$ represents the backdoor injection function, and $c(\cdot)$ is the poisonous label function. After training, the source model generates malicious predictions for source instances with a trigger input, as shown in the following equation:

$$c(y_i^s) = \mathbf{F}_s(\mathcal{B}(x_i^s)), \quad \forall i = 1, \ldots, n_s$$

The next step involves training the target model, denoted by $\mathbf{F}_t(\cdot)$. This model is initialized with the weights and architecture of the source model, $\mathbf{F}_s(\cdot)$, and then trained using the target instances in $\mathcal{D}_t$ without any labels. However, even after the target model has been trained without any poisonous samples, the attacker's goal is to ensure that the backdoor attack still persists in the target domain, i.e.,

$$c(y_i^t) = \mathbf{F}_t(\mathcal{B}(x_i^t)), \quad \forall i = 1, \ldots, n_t$$

Observations. The experimental results of the above threat model and attack scenario is illustrated in Table 2, which shows that a malicious source model can result in an equally malicious target model after SFDA, even though the target owner trains the target model without any intrusion from the source attacker. Thus, this leads to our first observation:

Observation I. *Performing a backdoor attack in the source domain is sufficient to attack the target domain, making existing SFDA highly vulnerable to the risk of a malicious source adversary.*

Table 3: *Performance of existing defense methods against backdoor [10] attack in SFDA on Office-Home [38]. Some methods are not applicable here which is denoted by N/A, since they either need access to source data or target labels.*

Methods	Source data	Target labels	*Rw* → *Ar*	
			ACC	ASR
Baseline SFDA [24]	✗	✗	74.21	99.59
SPECTRE [11]	✓	✗	N/A	N/A
Neural Cleanse [39]	✗	✓	N/A	N/A
NAD [23]	✗	✓	N/A	N/A
CLP [47]	✗	✗	56.53	15.00
SSDA (Ours)	✗	✗	74.17	3.05

Given the significant security risk posed by backdoor attacks in SFDA, owners of target domains must consider backdoor threats during adaptation and attempt to incorporate existing defenses into the adaptation process. However, existing defense methods for backdoor attacks have significant limitations, as illustrated in Table 3. For instance, existing strong defense methods such as SPECTRE [11], Neural Cleanse [39], and NAD [23] are not practical in their current form in SFDA, as they require either the source dataset or target labels for defense. As we discussed previously, given the SFDA setting, the recently proposed data-free CLP [47] method can potentially be a good candidate. However, as shown in Table 3, CLP provides reasonable security at the expense of clean (i.e., no attack) accuracy. CLP fails because it is a static post-pruning scheme after completing the target domain training. It fails to take advantage of available target data and training phase, which are additional resources the defender can leverage in defending the backdoor for SFDA. We hypothesize that instead of applying static techniques like CLP in SFDA, the defender needs to utilize the target domain training phase and develop a dynamic defense against the backdoor while performing the adaptation. This leads to our second observation, which is as follows:

Observation II. *The existing backdoor defense methods are ineffective in mitigating the security threat demonstrated in SFDA.*

These observations confirm the vulnerability of current SFDA methods to backdoor attacks and the inadequacy of existing defense methods in mitigating this security threat.

5. Our Proposed SSDA Method

To improve the security of SFDA against backdoor attacks, we propose a novel training scheme called secure source-free domain adaptation (SSDA). Our proposed SSDA consists of two critical components for improving the transferred domain's security.

The first component is called *Single-shot Static Defensive Compression*, which takes the pre-trained source model as input and produces a compressed version of the initial source model. We define *compression* as setting some sensitive malicious channels' weights to zero, but the channels remain part of the model (i.e., no pruning). Next question, how do we quantify malicious or sensitive channels responsible for backdoor transfer? In our work, we adopt *Spectral Norm* as a metric to evaluate the sensitivity of each channel. Since it is well-established [32, 47] that spectral-norm can accurately measure the channel sensitivity for adversarial input attack or backdoor attack. We rank the channels of a pre-trained source model using spectral norm and set the sensitive channels' (i.e., high spectral norm) weight to zero.

Our first defense component is a post-source training static defense approach. However, in SFDA, the defender has additional target data and an adaptation stage, which the defender can use to their advantage. Hence, we need a dynamic (i.e., training stage) defense component to ensure that we can train the target model w/o compromising accuracy and security. Thus, the goal of our training stage defense component would be to maintain the benign test accuracy after adaptation. Additionally, to improve security, we also want to make sure that during the adaptation process, we suppress the malicious channels that contribute to the model's backdoor behavior.

The above motivation has led us to develop our second novel defense component, called *Knowledge Transfer with Dynamic Channel Suppression*. Here, we train another *uncompressed* target model that ultimately leverages the original pre-trained source model (i.e., the exact source model provided by the source owner). We propose to train this auxiliary target model to generate pseudo-labels, later used to train our primary compressed model. The intuition behind this knowledge transfer is that an uncompressed source model can generate more accurate pseudo-labels specifically for benign inputs than the compressed counterpart, thus helping to improve the benign accuracy. To generate the pseudo-labels, we adopt a standard prior SFDA technique [24]. Finally, on top of our static compression, where we compress the model with a high spectral norm (i.e., higher likelihood of being malicious), during training, we propose to add a novel regularization loss that penalizes the channels with a high spectral norm. However, computing spectral norm during training is computationally intensive. To address this, we provide a novel theoretically validated safe upper bound of this norm to compute it during training efficiently.

5.1. Single-Shot Static Defensive Compression

This component will perform a single-shot compression of the model by setting the weights of the sensitive backdoor channels to zero. Specifically, we will perform the compression on the feature encoder of the source model $\mathbf{G}_s(\cdot)$, keeping the classifier module $\mathbf{H}_s(\cdot)$ unmodified which is consistent with prior works of UDA [20, 24, 21].

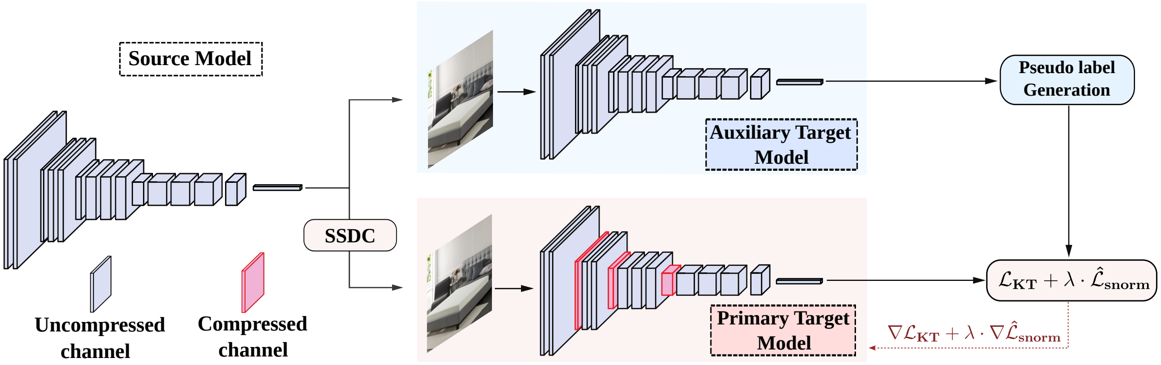
Source Model
Auxiliary Target Model
Pseudo label Generation
SSDC
Primary Target Model
$\mathcal{L}_{\mathbf{KT}} + \lambda \cdot \hat{\mathcal{L}}_{\text{snorm}}$
$\nabla\mathcal{L}_{\mathbf{KT}} + \lambda \cdot \nabla\hat{\mathcal{L}}_{\text{snorm}}$
Uncompressed channel
Compressed channel

Figure 3: *The proposed training pipeline of SSDA includes Single-Shot Defensive Compression (SSDC) to compress the source model's malicious (red-colored) channels. The primary target model is then trained using our knowledge transfer (KT) scheme and novel spectral norm penalty.*

We use spectral norm to identify the sensitive backdoor channels in the model, considering each channel as a linear function. In general, spectral norm of a linear function is defined as,

$$\sigma(\mathbf{A}) = \max_{||\mathbf{x}||_2 \neq 0} \frac{||\mathbf{A}\mathbf{x}||_2}{||\mathbf{x}||_2},$$

where, $\mathbf{A}$ is the transformation matrix of the linear function. Now, consider the source feature encoder $\mathbf{G}_s(\cdot)$ of L convolution layers with a set of convolution weight tensor $\mathcal{W} = \{\mathcal{W}^l : l = 1, 2, \ldots, L\}$. $\mathcal{W}^l \in \mathbb{R}^{K^l \times c^l \times h^l \times w^l}$ represents weight tensor of the l-th convolutional kernel, where K^l, c^l, h^l, and w^l are the number of output channels, input channels, height, and width of the convolutional kernel, respectively. To compute spectral norm of the k^{th} channel of the l^{th} convolution layer, denoted by $\mathcal{W}^l_k \in \mathbb{R}^{c^l \times h^l \times w^l}$, we reshape the tensor into a matrix as follows,

$$\mathsf{reshape} : \mathcal{W}^l_k \in \mathbb{R}^{c^l \times h^l \times w^l} \rightarrow \mathbf{W}^l_k \in \mathbb{R}^{c^l \times h^l w^l} \qquad (1)$$

Note that the spectral norm of this reshaped matrix approximates the spectral norm of convolution operation performed using the weight tensor validated in prior works [45, 47]. Next, we calculate the mean and variance of the spectral norm of the l^{th} layer as follows:

$$\mu^l = \frac{1}{K^l} \sum_{k=1}^{K^l} \sigma(\mathbf{W}^l_k), \quad s^l = \frac{1}{K^l} \sum_{k=1}^{K^l} (\sigma(\mathbf{W}^l_k) - \mu^l)^2$$

Subsequently, we compress the channels of l^{th} layer if its spectral norm exceeds a certain threshold, i.e.,

$$\mathcal{W}^l_k = \mathbf{0}, \quad \forall \sigma(\mathbf{W}^l_k) > \mu^l + \gamma \cdot \sqrt{s^l},$$

where γ is a hyper-parameter. Finally, we repeat the above steps for each layer of the source model and obtain the compressed source model $\widehat{\mathbf{F}}_s = \mathbf{H}_s \circ \widehat{\mathbf{G}}_s$, where $\widehat{\mathbf{G}}_s$ is the compressed source feature encoder. Overall, by utilizing the concept of spectral norm, we identify and compress the most sensitive channels of the source model initially before training for target domain adaptation. However, as the name implies, it is a one-time static (i.e., pre-target training) method on the pre-trained source model. Thus, during the target model's training, we need a dynamic (i.e., training stage penalty) approach for suppressing the malicious channels as well.

5.2. Knowlege Transfer with Dynamic Channel Suppression

The second component of our proposed SSDA method performs knowledge transfer from an auxiliary model (non-compressed) to the primary compressed model to recover the clean accuracy, as shown in Fig. 3. We call it knowledge transfer because the auxiliary model generates a more accurate pseudo-label and transfers the label information to the primary compressed model. Then, the compressed model is trained using the pseudo-label in a supervised setting. Next, to improve security during the knowledge transfer, we want to ensure our target training still suppresses the high spectral norm channels (i.e., backdoor sensitive). However, computing the spectral norm during training is computationally intensive. Hence, we derive a novel upper bound of the spectral norm and use this to efficiently penalize the high spectral norm channels during target training.

Our knowledge transfer process involves training two target models sequentially. The primary model, denoted as $\widehat{\mathbf{F}}_t(\cdot)$, is adapted from the compressed source model $\widehat{\mathbf{F}}_s(\cdot)$, while the auxiliary model, denoted as $\mathbf{F}_t(\cdot)$, is adapted from the non-compressed source model $\mathbf{F}_s(\cdot)$. Training the auxiliary model first aims to generate more accurate pseudo-labels. The primary model suffers from inaccurate pseudo-label generation due to its initialization with a compressed source model. The auxiliary model can be trained using existing SFDA methods to generate the pseudo-labels. In our approach, we generate the pseudo-label set, denoted as

Table 4: *Evaluation of SFDA (Baseline) [24] and SSDA on Office-Home dataset [38] for three different attacks for a subset of domains (rest are in supplementary).*

Attack	Method	$Ar \to Cl$		$Ar \to Pr$		$Ar \to Rw$		$Cl \to Ar$		$Cl \to Pr$		$Cl \to Rw$	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	SFDA [24]	55.60	74.23	77.31	88.08	80.70	83.22	67.37	96.33	77.29	85.40	78.24	92.72
	SSDA (Ours)	55.60	***1.56***	77.27	***1.80***	80.54	***1.72***	67.33	***3.05***	***77.40***	***1.91***	78.13	***1.77***
Blended	SFDA [24]	55.99	95.03	77.74	88.13	81.62	47.42	66.75	77.46	77.65	90.00	78.63	27.04
	SSDA (Ours)	55.88	***1.72***	***77.83***	***1.78***	81.32	***1.72***	66.71	***3.42***	***77.72***	***1.89***	78.38	***1.72***
WaNet	SFDA [24]	56.70	93.10	76.21	90.65	81.96	88.27	67.94	98.93	79.03	98.51	79.07	82.86
	SSDA (Ours)	***56.75***	***4.31***	***76.32***	***1.91***	81.59	***1.79***	***68.03***	***14.34***	***79.05***	**2.00**	78.98	***1.79***

$\widehat{\mathcal{Y}}_t$ for the target instance set $\mathcal{X}_t$, by training the auxiliary model using [24]. While training the auxiliary model, we only need to store the pseudo labels after each epoch, which results in a negligible memory overhead (~ 1 % worst case) for our defense.

Next, we will use these pseudo-labels to train the primary model using the standard cross-entropy loss function. Therefore, the loss function for our knowledge transfer approach can be defined as follows

$$\mathcal{L}_{\mathbf{KT}}(\mathcal{X}_t, \widehat{\mathcal{Y}}_t; \widehat{\mathbf{F}}_t) = \frac{1}{n_t} \sum_{x_t, \hat{y}_t \in \mathcal{X}_t, \widehat{\mathcal{Y}}_t} \ell(\hat{y}_t, \widehat{\mathbf{F}}_t(x_t)) \tag{2}$$

where $\ell(\cdot)$ is the standard cross-entropy loss.

Finally, to ensure that channel sensitivity is not further amplified during target training, we still want sensitive backdoor channels to get suppressed. We impose a constraint on the knowledge transfer loss function to achieve this by adding a penalty based on the spectral norm. The spectral norm penalty is calculated for the feature encoder of target model $\widehat{\mathbf{G}}_t(\cdot)$ during target training as:

$$\mathcal{L}_{\mathsf{snorm}}(\widehat{\mathbf{G}}_t) = \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{K^l} \sigma(\mathbf{W}_k^l)$$

where $\mathbf{W}_k^l$ is the reshaped weight tensor (defined in Equation 1) of k-th channel of the l-th convolution layer of $\widehat{\mathbf{G}}_t(\cdot)$.

However, computing the spectral norm for the whole target encoder $\widehat{\mathbf{G}}_t(\cdot)$ during training can be computationally expensive as it requires Singular Value Decomposition (SVD). To alleviate this computational burden, we propose an approximation method to replace the spectral norm with its upper bound derived through the following lemma,

Lemma 5.1. *The spectral norm of a matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *is upper-bounded by,*

$$\boldsymbol{\sigma}(\mathbf{A}) \leq \mathsf{trace}(\mathbf{A}^T \mathbf{A})$$

Proof. Consider the SVD of matrix $\mathbf{A} = \mathbf{U\Sigma V}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices and $\mathbf{\Sigma}$ is a diagonal matrix with singular values in the diagonal entries. Now,

$$\mathbf{A}^T \mathbf{A} = (\mathbf{U\Sigma V}^T)^T \cdot (\mathbf{U\Sigma V}^T) = \mathbf{V\Sigma}^T \mathbf{\Sigma V}^T$$

Applying Spectral theorem [14], we get $\mathbf{A}^T \mathbf{A} = \mathbf{S\Lambda S}^T$, where $\mathbf{S}$ is an orthonormal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with eigen values of $\mathbf{A}^T\mathbf{A}$ in the diagonal entries. Then it follows that $\mathbf{\Lambda} = \mathbf{\Sigma}^T \mathbf{\Sigma}$. Therefore,

$$\boldsymbol{\sigma}(A) = \max_i \mathbf{\Sigma}_{ii} \leq \mathsf{trace}(\mathbf{\Lambda}) = \mathsf{trace}(\mathbf{A}^T \mathbf{A}),$$

which proves the lemma. □

Hence, we approximate the spectral norm penalty as,

$$\hat{\mathcal{L}}_{\mathsf{snorm}}(\widehat{\mathbf{G}}_t) = \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{K^l} \mathsf{trace}\big(\mathbf{W}_k^{l\,T} \mathbf{W}_k^l\big) \tag{3}$$

Therefore, the overall loss function for training the target model $\widehat{\mathbf{F}}_t(\cdot)$ is as follows,

$$\mathcal{L}_{\mathbf{SSDA}}(\mathcal{X}_t, \widehat{\mathcal{Y}}_t; \mathbf{H}_t \circ \widehat{\mathbf{G}}_t) = \mathcal{L}_{\mathbf{KT}}(\cdot) + \lambda \cdot \hat{\mathcal{L}}_{\mathsf{snorm}}(\cdot)$$

where $\mathcal{L}_{\mathbf{KT}}(\cdot)$ and $\hat{\mathcal{L}}_{\mathbf{snorm}}(\cdot)$ are defined in Equation 2 and Equation 3, respectively. Here, λ controls the weight of the approximate spectral norm penalty. During target training, the defender's objective is to update $\mathcal{W}$ by minimizing $\mathcal{L}_{\mathbf{SSDA}}(\cdot)$.

6. Experimental Setup

We evaluate our proposed SSDA on three commonly used visual benchmarks for assessing DA methods: Office [33], Office-Home [38] and VisDA-C [30] (in the supplementary). We are following standard SFDA setting [24] including network architecture (e.g., ResNet-50 [13] for Office-Home and Office, ResNet-101 [13] for VisDA-C) and hyper-parameters. We evaluate the effectiveness of our proposed SSDA against three well-known and recent attack methods: BadNets [10], Blended Backdoor Attack [3] and WaNet [28]. All attacks employ an All-to-One strategy where the poisoning label for a subset of the source dataset is set to a specific label C_k, i.e., $y_i^s = C_k$. We set γ to 1, and λ to 100 for all of our experiments. We direct the reader to the Supplementary section for additional implementation details and ablation studies of hyper-parameters (e.g., λ). In our evaluation, we report the standard test accuracy (ACC in %) in the target domain and attack success rate (ASR in %), which is the percentage of test samples classified to the target class for test samples with embedded attacker triggers.

Table 5: *Evaluation of SFDA [24] and SSDA on Office dataset [33] across all possible attack combinations and domains.*

Attack	Method	$A \rightarrow D$		$A \rightarrow W$		$D \rightarrow A$		$D \rightarrow W$		$W \rightarrow A$		$W \rightarrow D$	
		ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR
BadNets	SFDA [24]	92.37	99.80	88.30	98.74	73.91	68.51	98.74	100.00	73.45	38.62	99.80	96.59
	SSDA (Ours)	92.37	***2.61***	88.30	***3.65***	73.87	***3.41***	98.74	***3.65***	73.38	***3.27***	99.80	***2.61***
Blended	SFDA [24]	91.77	99.80	89.69	93.08	74.01	38.05	98.62	79.87	73.55	80.51	100.00	100.00
	SSDA (Ours)	91.77	***6.63***	89.69	***3.65***	74.01	***4.01***	98.62	***5.91***	***73.62***	***5.61***	100.00	***31.73***
WaNet	SFDA [24]	92.17	100.00	89.43	100.00	74.80	42.60	98.74	99.62	73.55	78.20	100.00	99.80
	SSDA (Ours)	92.17	***56.02***	89.43	***36.98***	74.76	***3.44***	98.74	***18.36***	73.41	***3.76***	100.00	***55.42***

Table 6: *Evaluation of the performance of SFDA and proposed SSDA under two different cases: i) Considering a benign source training (Left-half); ii) Considering a malicious source training (Right-half).*

Method	$Ar \rightarrow Cl$	$Cl \rightarrow Ar$	Method	$Ar \rightarrow Cl$	$Cl \rightarrow Ar$
	ACC	ACC		ACC	ACC
SFDA [24] **(benign source)**	56.66	68.03	SFDA [24] **(malicious source)**	55.60	67.37
SSDA **(benign source)**	56.56	67.99	SSDA **(malicious source)**	55.60	67.33

7. Evaluation and Discussion

Evaluation of SSDA. The evaluation of SSDA is presented in Table 4 on the office-home dataset. This table compares our SSDA with the baseline SFDA [24] case. First, we observe that SFDA is highly vulnerable to backdoor attacks with high ASR for all the attacks across all the domains. In contrast, our SSDA can significantly lower the attacking threat by achieving less than *4%* ASR against BadNets and Blended attacks and less than *5%* ASR against WaNet attack across all the tasks in Table 4. We achieve this improved security against the backdoor attack without compromising test accuracy compared to baseline SFDA. The results further indicate that, although all attacks pose significant security threats for SFDA, WaNet is the most potent attack, achieving higher ASR on most tasks ($\sim$ 90 % or above). Regardless, SSDA can still successfully defend against this strong attack by lowering the attack success rate by more than *80 %* from the baseline case.

We also present results on the Office benchmark dataset in Table 5 and the VisDA-C dataset in the supplementary materials. A general conclusion across different datasets, tasks and attacks is that our proposed SSDA can significantly improve the defense against backdoor (i.e., much lower ASR than SFDA) with little to no degradation in test accuracy. In addition, in Table 6, we demonstrate that in the case of benign source training (i.e., no attack), our SSDA's adaptation performance does not deteriorate (with 0.1 %) compared to the baseline SFDA [24]. *To summarize, our comprehensive evaluation across the different datasets and multiple tasks confirm that the backdoor is a critical security concern for SFDA. Our proposed SSDA provides a feasible solution with no compromise in test accuracy.*

Table 7: *Performance of different SFDA methods against BadNets [10] on Office-Home dataset.*

Method	$Ar \rightarrow Cl$		$Cl \rightarrow Ar$	
	ACC	ASR	ACC	ASR
SHOT [24]	55.60	74.23	67.37	96.33
G-SFDA [43]	54.96	97.16	64.07	99.55
NRC [42]	56.93	60.69	67.94	83.93
AaD [44]	58.42	78.95	66.50	67.70
SSDA (ours)	55.60	***1.56***	67.33	***3.05***

Table 8: *Effect of each component of our proposed SSDA. Here, SSDC denotes Single-shot static defensive compression (the first component of our defense).*

Method	$Ar \rightarrow Cl$		$Cl \rightarrow Ar$	
	ACC	ASR	ACC	ASR
SFDA [24] (Baseline)	55.60	74.23	67.37	96.33
SFDA [24] + SSDC	10.84	42.08	16.73	66.09
SSDC + $\mathcal{L}_{\text{KT}}$ (Ours)	56.54	34.78	67.66	47.55
SSDC + $\mathcal{L}_{\text{KT}}$ + $\lambda \cdot \hat{\mathcal{L}}_{\text{snorm}}$ (Ours)	56.75	4.31	68.03	14.34

Comparison to SOTA SFDA techniques and Backdoor defenses. In Table 7, we compare the proposed SSDA with existing SOTA SFDA techniques. The result further confirms that the backdoor is a persistent threat to all existing SFDA methods. Our SSDA provides the best protection against backdoor attacks (i.e., lowest ASR) with similar test accuracy. However, to be fair to the prior methods [24, 42, 43], our method requires twice (worst-case) the training time as we do need to train two different models sequentially. Additionally, we have a slight memory overhead ($\sim$ 1% for the worst case) as we need to store the 8-bit pseudo labels for primary model training, which is tiny compared to the model size (e.g., ResNet-101).

We have already compared our proposed SSDA with existing backdoor defenses (in the context of SFDA) in Section 4 in Table 3. *Therein, we showed that the existing backdoor defenses either do not apply in their current form or are ineffective in countering backdoor attacks in SFDA, thereby making SSDA the first and only successful defense against backdoor attacks in SFDA*.

Ablation Study. In Table 8, we summarize the effect of each component of our defense and its role in improving test accuracy and defense. First, we show the baseline SFDA [24] case, which suffers from poor defense performance. Next, adding the first component of our defense (SSDC) improves defense performance at the expense of an accuracy drop in the target model due to pre-training compression. Next, we can recover test accuracy to the baseline

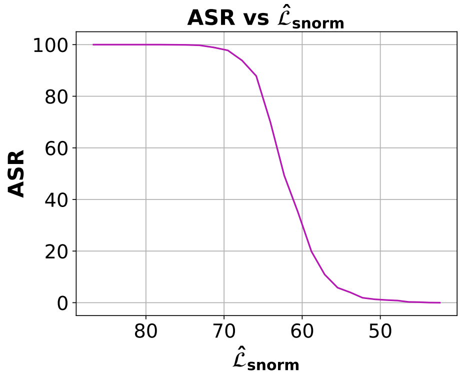

Figure 4: *Effect of spectral norm minimization on ASR.*

level using better pseudo-labels from our novel knowledge transfer loss strategy (SSDC + $\mathcal{L}_{\text{KT}}$), but still with poor security performance. Finally, to defend against backdoor attacks successfully while maintaining benign accuracy, our novel upper bound of the snorm loss penalty ensures that the malicious channels are suppressed during target training to reduce ASR. This observation is further validated in Fig. 4, clearly exhibiting the relation between minimizing the snorm loss term and ASR.

8. Conclusion

We are the first to demonstrate a major security threat for SFDA by considering a source adversary. Hence, we develop a novel target domain training method by compressing and penalizing the sensitive channels with high spectral norms augmented by a knowledge transfer scheme. Our novel SSDA is the first and only successful defense against a backdoor in SFDA, completely mitigating the threat in the target domain. We perform extensive experiments on multiple datasets and tasks to comprehensively evaluate the impact of our proposed method in mitigating the risk. Our proposed defense is the first successful defense against a backdoor in SFDA while ensuring the performance of the domain adaptation does not suffer regardless of benign/malicious source. In the future, we hope more investigations will be undertaken to secure vulnerable SFDAs against malicious adversaries.

References

[1] Sabbir Ahmed, Uday Kamal, and Md. Kamrul Hasan. Dfr-tsd: A deep learning based framework for robust traffic sign detection under challenging weather conditions. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5150–5162, 2022. 1

[2] Sabbir Ahmed, Moi Hoon Yap, Maxine Tan, and Md Kamrul Hasan. Reconet: Multi-level preprocessing of chest x-rays for covid-19 detection using convolutional neural networks. *medrxiv*, pages 2020–07, 2020. 1

[3] 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. 1, 3, 4, 7

[4] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12455–12464, 2020. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[6] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7212–7222, 2022. 1, 3

[7] Bao Gia Doan, Ehsan Abbasnejad, and Damith C Ranasinghe. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Annual Computer Security Applications Conference*, pages 897–912, 2020. 3

[8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[9] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013. 1

[10] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019. 1, 2, 3, 4, 5, 7, 8

[11] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, pages 4129–4139. PMLR, 2021. 2, 3, 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[14] A Horn Roger and R Johnson Charles. Matrix analysis cambridge university press. *New York*, 1985. 7

[15] Md Shamim Hussain, Sharmin R Ara, Uday Kamal, Sabbir Ahmed, and Md Kamrul Hasan. Breast lesion classification from bi-modal ultrasound images by deep cnn using transfer and multi-task learning. 1

[16] Taotao Jing and Zhengming Ding. Adversarial dual distinct classifiers for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 605–614, 2021. 2

[17] Taotao Jing, Haifeng Xia, and Zhengming Ding. Adaptively-accumulated knowledge transfer for partial domain adaptation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1606–1614, 2020. 2

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[19] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020. 1, 3

[20] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 615–625, 2021. 3, 5

[21] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9641–9650, 2020. 3, 5

[22] Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11516–11525, 2021. 1

[23] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 2, 3, 5

[24] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 1, 3, 5, 7, 8

[25] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 2

[26] Tasfin Mahmud, Mehedi Hossen Limon, Sabbir Ahmed, Mohammad Zunaed Rafi, Borhan Ahamed, Shadman Shahriar Nitol, Md Yeasin Mia, Rafat Emtiaz Choudhury, Adnan Sakib, Arik Subhana, et al. Non-invasive blood glucose estimation using multi-sensor based portable and wearable system. In *2019 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–5. IEEE, 2019. 1

[27] Jaemin Na, Heechul Jung, Hyung Jin Chang, and Wonjun Hwang. Fixbi: Bridging domain spaces for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1094–1103, 2021. 1

[28] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021. 1, 2, 3, 4, 7

[29] Michela Paganini. Prune responsibly. *arXiv preprint arXiv:2009.09936*, 2020. 2

[30] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017. 7

[31] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13198–13207, 2020. 3

[32] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *Advances in Neural Information Processing Systems*, 33:14973–14985, 2020. 5

[33] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010. 7, 8

[34] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 2

[35] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 3

[36] Luan Tran, Kihyuk Sohn, Xiang Yu, Xiaoming Liu, and Manmohan Chandraker. Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2672–2681, 2019. 1

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[38] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 4, 5, 7

[39] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben 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. 2, 3, 5

[40] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15074–15084, 2022. 3

[41] Haifeng Xia and Zhengming Ding. Hgnet: Hybrid generative network for zero-shot domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 55–70. Springer, 2020. 2

[42] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 34:29393–29405, 2021. 1, 8

[43] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain

adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 1, 3, 8

[44] Shiqi Yang, Yaxing Wang, Kai Wang, SHANGLING JUI, and Joost van de weijer. Attracting and dispersing: A simple approach for source-free domain adaptation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 8

[45] Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017. 6

[46] Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations*, 2020. 3

[47] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 175–191. Springer, 2022. 2, 3, 5, 6

[48] Ranyang Zhou, Sabbir Ahmed, Adnan Siraj Rakin, and Shaahin Angizi. Dnn-defender: An in-dram deep neural network defense mechanism for adversarial weight attack. *arXiv preprint arXiv:2305.08034*, 2023. 3