# Temporal-Distributed Backdoor Attack against Video Based Action Recognition

**Xi Li*, Songhe Wang*, Ruiquan Huang, Mahanth Gowda, George Kesidis**

The Pennsylvania State University
{xzl45, sxw5765, rzh5514, mkg31, gik2}@psu.edu

## Abstract

Deep neural networks (DNNs) have achieved tremendous success in various applications including video action recognition, yet remain vulnerable to backdoor attacks (Trojans). The backdoor-compromised model will mis-classify to the target class chosen by the attacker when a test instance (from a non-target class) is embedded with a specific trigger, while maintaining high accuracy on attack-free instances. Although there are extensive studies on backdoor attacks against image data, the susceptibility of video-based systems under backdoor attacks remains largely unexplored. Current studies are direct extensions of approaches proposed for image data, e.g., the triggers are **independently** embedded within the frames, which tend to be detectable by existing defenses. In this paper, we introduce a *simple* yet *effective* backdoor attack against video data. Our proposed attack, adding perturbations in a transformed domain, plants an **imperceptible, temporally distributed** trigger across the video frames, and is shown to be resilient to existing defensive strategies. The effectiveness of the proposed attack is demonstrated by extensive experiments with various well-known models on two video recognition benchmarks, UCF101 and HMDB51, and a sign language recognition benchmark, Greek Sign Language (GSL) dataset. We delve into the impact of several influential factors on our proposed attack and identify an intriguing effect termed "collateral damage" through extensive studies.

## 1 Introduction

Deep neural networks (DNNs) have shown impressive performance in various applications, yet remain susceptible to adversarial attacks. Recently, backdoor (Trojan) attacks on DNNs have garnered attention in multiple domains, including image classification (Gu, Dolan-Gavitt, and Garg 2017; Chen et al. 2017; Nguyen and Tran 2021; Saha, Subramanya, and Pirsiavash 2020; Li et al. 2021a), speech recognition (Liu et al. 2018), text classification (Dai, Chen, and Li 2019), point cloud classification (Xiang et al. 2021), and even deep regression (Li et al. 2021b). The attacker plants a backdoor in the victim model, which is fundamentally a mapping from a specific trigger to the attacker-chosen target class. During inference, the compromised model will misclassify a test instance embedded with the same trigger to the target class. Moreover, the attacked model still maintains high accuracy on users' (backdoor-free) validation sets, rendering the attack stealthy. The typical backdoor attack is implemented by poisoning the training set for the victim DNN using a few instances embedded with the trigger, while intentionally mislabeling them to the target class.

Video recognition systems are increasingly integrated into various domains, such as surveillance systems (Elharrouss et al. 2021), autonomous vehicles (Saleh, Hossny, and Nahavandi 2019), and video-based sign language recognition (Li et al. 2020). The threat posed by backdoor attacks on video recognition systems can be significant and multifaceted. A compromised surveillance system could allow undetected crimes by mis-identifying malicious events or villains. Similarly, a backdoored autonomous vehicle may misread moving pedestrians, risking fatal accidents. Moreover, a tampered sign language system might dangerously misinterpret an emergency sign such as confusing "help" for "fine".

Considering the widespread application of video recognition systems, it is crucial to study their robustness under potential threats. However, there is a noticeable gap in the literature addressing this. In this work, we aim to investigate the robustness of video recognition systems by probing their vulnerability to backdoor attacks. Current studies such as Hammoud et al. (2023); Zhao et al. (2020) directly extend backdoor attacks against images to videos by embedding identical triggers into every frame. These attacks demonstrate efficacy, as the repeated embedding reinforces the model's training on the backdoor mapping. However, they present two primary challenges: (1) Some of the triggers are perceptible to humans; (2) More importantly, the triggers are *independently* embedded in each frame, therefore, can be caught by existing backdoor defenses originally proposed for image data.

In this paper, we leverage the temporal dimension of videos and plant an **imperceptible temporal-distributed** trigger within videos to address the two challenges. The proposed attack is *simple* yet *effective*. The trigger is embedded by appropriately altering certain components of a representation of a video, thus the trigger is imperceptible to human observers in the video space. Also, by carefully choosing the basis of the transformed space, this backdoor trigger is distributed across the whole video. Hence, the proposed attack could circumvent the existing backdoor defenses which

*These authors contributed equally.

examine the frames individually. Furthermore, we reveal a phenomenon brought by the proposed attack, termed "collateral damage". We analyze this phenomenon and its possible reasons on various transforms and model structures. In summary, our contributions are as follows:

1. We propose the **first general framework** to embed an **imperceptible temporal-distributed** backdoor trigger in **videos**, closing the gaps left by previous works. This is achieved by delicately selecting a basis of the input space and perturbing certain components in the representation on the basis. We specialize the framework to Fourier, cosine, wavelet transform, and random transforms.
2. We empirically validate the efficacy of our proposed attack across diverse benchmark datasets and different model architectures in the realm of video recognition. Moreover, our evaluations underscore the stealthiness of the attack, not only to human observers but also to current backdoor detection and mitigation techniques.
3. We further conduct extensive studies to explore the impacts of several key factors on the effectiveness of the proposed attack, providing insights on robustness verification of DNNs. Also, we reveal and analyze an interesting phenomenon, termed "collateral damage", associated with the proposed attack.

# 2 Related Work

## 2.1 Backdoor Attacks and Defenses

Backdoor attacks are one type of poisoning attacks initially proposed against DNN image classifiers. There are various ways of designing effective triggers in image classification: (1) Embedding patterns directly in the *input space* (Li et al. 2021a; Xiang, Miller, and Kesidis 2020; Gu, Dolan-Gavitt, and Garg 2017; Liu et al. 2018; Li et al. 2021a; Chen et al. 2017; Barni, Kallas, and Tondi 2019); (2) Introducing triggers in *an alternative space*, which leads to imperceptible input-specific triggers in the input space (Wang et al. 2022; Li et al. 2022a; Nguyen and Tran 2021).

*However, the backdoor attacks against video data remain largely unexplored.* Hammoud et al. (2023) extend backdoor attacks from images to video recognition by independently embedding identical triggers in each frame, effectively compromising several models but vulnerable to existing defenses due to uncorrelated triggers. Zhao et al. (2020), following the existing framework (Shafahi et al. 2018; Turner, Tsipras, and Madry 2019; Saha, Subramanya, and Pirsiavash 2020), use clean-label poisoning to attack video systems, under the impractical assumption of attacker access to the clean training set. By contrast, we introduce a more practical, effective backdoor attack against video data, employing a **temporal-distributed** trigger across frames, which evades existing defenses through its strong inter-frame trigger correlation.

Existing backdoor defenses are deployed either before/-during the DNN's training stage or post-training. *Pre-training* defenses, such as Tran, Li, and Madry (2018); Chen et al. (2019), are based on anomaly detection techniques. Methods such as Du, Jia, and Song (2020); Huang et al. (2022) are deployed *during DNN training*. On the other hand, *post-training detection* methods, such as Wang et al. (2019); Xiang, Miller, and Kesidis (2020); Guo et al. (2019), detect whether a given classifier has been backdoor-compromised; Gao et al. (2019); Chou, Tramèr, and Pellegrino (2020); Doan, Abbasnejad, and Ranasinghe (2020); Li et al. (2022b) catch triggered test instances in the act. Besides, *post-training backdoor mitigation* approaches are proposed to mitigate backdoor attacks at test time, such that the model behaves normally on both clean and triggered inputs. Backdoor mitigation methods include Liu, Dolan-Gavitt, and Garg (2018); Wu and Wang (2021); Guan et al. (2022); Zheng et al. (2022); Li et al. (2021c); Xia et al. (2022); Zeng et al. (2022); Madry et al. (2018). Recently, Wang et al. (2023) propose an advanced backdoor trigger estimation strategy, UNICORN. They define a backdoor trigger as a predefined perturbation in a particular space, and approximate the transform and its inversion to this space by neural nets, which are jointly optimized with the backdoor trigger. However, their work is infeasible, especially on video data, due to the extremely high computation cost for estimating the trigger and the possible transform methods.

## 2.2 Video Action Recognition

Over the years, researchers have formulated three categories of video recognition models: 2D CNN + RNN, 3D-CNN, and Transformer-based models. The 2D CNN + RNN approach uses 2D CNN for frame feature extraction and RNN for capturing temporal dependencies between them (Donahue et al. 2015; Ng et al. 2015; Baccouche et al. 2011; Liu, Liu, and Chen 2016; Zhu et al. 2016). Later, 3D-CNNs evolved to concurrently process spatial and temporal dimensions, enabling motion pattern recognition across successive frames (Tran et al. 2015; Qiu, Yao, and Mei 2017; Carreira and Zisserman 2017; Hara et al. 2018). Inspired by the success in natural language processing, transformer-based models have entered this domain, using self-attention to gauge the relevance of different frames (Bertasius, Wang, and Torresani 2021; Liu et al. 2021).

# 3 Methodology

**Notations.** In this paper, we consider video action recognition tasks. The classifier, denoted by $f : \mathcal{V} \to \mathcal{A}$, is learned from a training dataset $\mathcal{D}_{\text{Train}} = \{(\mathbf{v}_\iota, a_\iota)\}_{\iota \in \mathcal{I}'}$, where $\mathcal{I}'$ is an index set, $\mathcal{V}$ denotes the input space, and $\mathcal{A}$ is the label space. We use $[N]$ to denote the set of integers from 0 to $N-1$. For simplicity, we only consider one channel of the video. The input space is then defined as $\mathcal{V} := [256]^{N_0 \times N_1 \times N_2}$, where $N_0$ is the number of frames[1], $N_1 \times N_2$ is the size of a frame. $\mathbf{v}(n_0, n_1, n_2) \in [256]$ is the pixel value at the frame $n_0$ and position $(n_1, n_2)$ of a video $\mathbf{v}$. Finally, $\mathbf{1}_E$ is the indicator function of an event $E$.

## 3.1 Threat Model

We consider classic mis-labelling backdoor poisoning attacks. We assume the attacker has the following **abilities**: (1) knows the classification domain $\mathcal{A}$ to collect valid samples

[1]For simplicity, we assume all videos have the same length. In experiments, shorter videos are padded with blank frames.

$\mathcal{D}_{\mathcal{S}} = \{(\mathbf{v}_\iota, s_\iota) | s_\iota \in \mathcal{A} \setminus \{t\},\ \iota \in \mathcal{I}_0\}$ from all classes other than the target class $t$ desired by the attacker (i.e., an all-to-one attack); (2) has access to the training set and can inject mis-labeled backdoor-triggered samples into it, i.e., $\mathcal{D}_{\text{Train}} = \mathcal{D}_{\text{Clean}} \cup \mathcal{D}_{\text{Attack}}$, where $\mathcal{D}_{\text{Attack}} = \{(\mathcal{B}(\mathbf{v}), t) | (\mathbf{v}, \cdot) \in \mathcal{D}_{\mathcal{S}}\}$, and $\mathcal{B} : \mathcal{V} \to \mathcal{V}$ is the attacker-specific trigger embedding function that embeds trigger into a given video $\mathbf{v}$; (3) is not aware of the structure of the target model (i.e., a black-box attack). After poisoned training, the attacker **aims** to: (i) have the victim classifier learn the "backdoor mapping" – the backdoor-attacked classifier will predict the attacker's desired target class $t$ when a test instance $\mathbf{v} \in \mathcal{V}$ is embedded with the backdoor trigger using $\mathcal{B}$; (ii) have the victim classifier achieve the accuracy on the user's (attack-free) validation set that is close to that of a non-poisoned classifier; (iii) have the trigger in the input space be visually imperceptible to a human.

### 3.2 Backdoor Attacks against Video: A Higher Level of Stealthiness

Unlike images, videos incorporate an additional dimension: time. This provides the possibility of a higher level of stealthiness against the current backdoor defense strategies. Studies such as Hammoud et al. (2023); Zhao et al. (2020) trivially extend image backdoor attacks (*e.g.*, the ones proposed by Gu, Dolan-Gavitt, and Garg (2017); Chen et al. (2017); Turner, Tsipras, and Madry (2019)) to videos. Hammoud et al. (2023) **independently** embed the classic backdoor triggers for images into each frame of a video. Although these attacks are effective against video data (as shown in Tab. 2 and Hammoud et al. (2023)), there are two major problems: (1) some of the backdoor triggers are human perceptible (*e.g.*, Gu, Dolan-Gavitt, and Garg (2017); Chen et al. (2017)), *i.e.*, can be detected by a human without advanced defenses. (2) More importantly, since they apply frame-wise trigger embedding strategy, which is fundamentally the same as the ones applied on images, the attacks are susceptible to existing backdoor defense strategies in image domain, as illustrated in Tab. 3 and Tab. 2.

To address problems (1) and (2), we need to design a trigger that satisfies the following properties: First, it introduces minor variation to each pixel so that the trigger is human imperceptible (Barten 1999; Wang et al. 2004); Second, it is temporally distributed, *i.e.*, the trigger spans the entire video, and thus evades existing backdoor defense mechanisms originally proposed for image data. A trigger satisfying the two properties could be generated by making perturbations in a transformed space. Such transformed space is defined on a basis that has non-identical entries across time. Specifically, an (appropriate amount of) perturbation added to certain components of such a representation may introduce minor variation to each pixel of the original representation, so effectively "spreading out" the energy of the perturbations across the entire video. As a result, only a combination of subtle patterns in consecutive frames is able to trigger the attack, and thus the attack is able to evade current backdoor defenses which individually examine the frames.

Notably, designing triggers for videos presents unique challenges compared to images. First, designing a backdoor pattern learnable by video action recognition systems, *e.g.*, 3D convolutional neural networks, is difficult since they extract different features than 2D image classifiers; Second, video data is more complicated than image due to the additional time dimension. Thus, effective video triggers require thorough exploration of key factors such as the number of perturbed components and perturbation magnitude (will be discussed in Sec. 4.6), which are hardly derived from the backdoor attacks proposed in image domain.

### 3.3 Imperceptible Temporal-Distributed Backdoor Attack against Video Data

Our general framework of constructing poisoned instances $\mathcal{D}_{\text{Attack}}$ from clean samples of all non-target classes $\mathcal{D}_{\mathcal{S}}$ consists of three steps: (1) Select a basis of the transformed space; (2) Embed the trigger in the transformed representation of video data; (3) Reconstruct video data from the perturbed transformed representations. We now provide details of the trigger embedding function $\mathcal{B}$.

**Step 1: Selection of a transform basis.** To generate a temporally distributed trigger in the original space, we need to design a basis $\mathbf{B} = \{\mathbf{b}_0, \ldots, \mathbf{b}_{N-1}\} \subset \mathcal{V}$ of the input space. Then, a video sample $\mathbf{v} \in \mathcal{D}_{\mathcal{S}}$ can be represented by the linear combination of the basis videos: $\mathbf{v} = \sum_{n=0}^{N-1} r_n^{\mathbf{v}} \mathbf{b}_n$. We denote the transformed representation (coordinates) of the video $\mathbf{v}$ by $\mathbf{R}^{\mathbf{v}} = \{r_0^{\mathbf{v}}, \ldots, r_{N-1}^{\mathbf{v}}\}$.

**Step 2: Backdoor trigger embedding.** We then embed the backdoor trigger in the transformed space by perturbing certain components in the representation $\mathbf{R}_{\mathbf{v}}$. Let $\delta \geq 0$ be the perturbation magnitude, and $\mathcal{I} \subset [N]$ be the index-set of the components to be perturbed. We add a perturbation of $\delta$ to each component $r_n^{\mathbf{v}}$, $\forall n \in \mathcal{I}$. So, The perturbed representation is $\tilde{\mathbf{R}}^{\mathbf{v}} = \{r_n^{\mathbf{v}} + \delta \mathbf{1}_{\{n \in \mathcal{I}\}}\}_{n=0}^{N-1}$.

**Step 3: Video reconstruction.** We then reconstruct a valid video from the perturbed representation $\tilde{\mathbf{R}}^{\mathbf{v}}$ to get a poisoned sample for $\mathcal{D}_{\text{Train}}$. The resulting instance in the original space is expressed as $\mathbf{v}' = \sum_{n=0}^{N-1} (r_n^{\mathbf{v}} + \delta \mathbf{1}_{\{n \in \mathcal{I}\}}) \mathbf{b}_n$. Since certain components in the representation are perturbed by $\delta$, the inverse-transformed instance $\mathbf{v}'$ might not be a valid instance in the space $\mathcal{V}$. For example, certain entries of $\mathbf{v}'$ could be non-integer valued or fall outside the range of valid pixel intensities $[256]$. Hence, we apply a projection function $\Pi_{\mathcal{V}}$ to the resulting instance $\mathbf{v}'$ to obtain a valid video $\tilde{\mathbf{v}}$. In other words, the range of $\Pi_{\mathcal{V}}$ is $\mathcal{V}$.

In summary, we define the backdoor trigger embedding function $\mathcal{B}_{\mathcal{I},\delta}$ as

$$\mathcal{B}_{\mathcal{I},\delta}(\mathbf{v}) = \Pi_{\mathcal{V}} \left( \sum_{n=0}^{N-1} (r_n^{\mathbf{v}} + \delta \mathbf{1}_{\{n \in \mathcal{I}\}}) \mathbf{b}_n \right). \quad (1)$$

The attacker chooses the parameters $\mathcal{I}$ and $\delta$ of the backdoor trigger, generates backdoor-triggered samples by applying the trigger embedding function $\mathcal{B}_{\mathcal{I},\delta}$ to videos of $\mathcal{D}_{\mathcal{S}}$, mislabels them to the target class $t$, and injects them into the training set $\mathcal{D}_{\text{Train}}$. That is,

$$\mathcal{D}_{\text{Train}} = \mathcal{D}_{\text{Clean}} \cup \{(\mathcal{B}_{\mathcal{I},\delta}(\mathbf{v}), t) | (\mathbf{v}, \cdot) \in \mathcal{D}_{\mathcal{S}}\}.$$

We now present two classic transforms and their basis construction, and defer discrete wavelet transform

(DWT) and random transform (RT) to Apdx. A[2]. Let $\{\mathbf{e}_{n_0,n_1,n_2}\}_{n_\iota\in[N_\iota],\iota\in[3]}$ denote the standard basis of single-channel video, i.e., $\mathbf{v} = \sum_{n_0,n_1,n_2} \mathbf{v}(n_0,n_1,n_2)\mathbf{e}_{n_0,n_1,n_2}$.

**Discrete Fourier Transform (DFT).** DFT provides a comprehensive view of the frequency information of videos. The basis $\{\mathbf{b}_{k_0,k_1,k_2}\}_{k_\iota\in[N_\iota],\iota\in[3]}$ of DFT is defined as follows (with $i=\sqrt{-1}$):

$$\mathbf{b}_{k_0,k_1,k_2} = \sum_{n_0,n_1,n_2} \mathbf{e}_{n_0,n_1,n_2} \prod_{\iota=0}^{2} \exp(-in_\iota k_\iota/N_\iota).$$

**Discrete Cosine Transform (DCT).** DCT is similar to DFT. The basis of DCT is defined as follows.

$$\mathbf{b}_{k_0,k_1,k_2} = \sum_{n_0,n_1,n_2} \mathbf{e}_{n_0,n_1,n_2} \prod_{\iota=0}^{2} \cos(\pi n_\iota k_\iota/(2N_\iota)).$$

We remark that the above basis have non-identical entries across time due to their dependencies on $k_0$ (time).

# 4 Experiments

## 4.1 Experimental Setup

**Datasets**: We consider two benchmark datasets used in video action recognition, **UCF-101** (Soomro, Zamir, and Shah 2012) and **HMDB-51** (Kuehne et al. 2011), and a sign language recognition benchmark, Greek Sign Language (**GSL**) dataset (Adaloglou et al. 2022). UCF-101 encompasses 13,320 video clips sorted into 101 distinct action categories. Similarly, HMDB-51 contains 7,000 video clips categorized into 51 classes of action. GSL incorporates 40,785 gloss instances across 310 unique glosses[3].

**Target Model Architectures**: In our main experiments, we consider four popular CNN-based model architectures used for video action recognition: **SlowFast** (Feichtenhofer et al. 2019), **Res(2+1)D** (Tran et al. 2018), **S3D** (Xie et al. 2018) and **I3D** (Carreira and Zisserman 2017). These models utilize 3D kernels to jointly leverage the spatial-temporal context within a video clip. The results on transformer-based networks, *e.g.*, timesformer (Bertasius, Wang, and Torresani 2021) are shown in Apdx. F.

**Training Settings**: We train all the models on all the datasets for 10 epochs, using the AdamW optimizer (Loshchilov and Hutter 2019) with an initial learning rate of $0.0003$. Following the common training strategy in video recognition (Hammoud et al. 2023) and for reducing computation cost, we down-sample the videos into 32 frames.

**Attack Settings**: We consider *all-to-one* attacks. We arbitrarily choose **class 0** as the target class, and randomly select 20% of the training samples per class for the attacker's manipulation. For the proposed attack, we apply **DFT** for trigger generation in the main experiments. The results of using **other transform** methods (*e.g.*, **DCT**, **DWT**, and **RT**) are shown in Tab. 5 in Apdx. F. In the frequency domain of the video, we select a subset $\mathcal{I} = \{35,36,\ldots,44\} \times X \times Y \subset [N_0]\times[N_1]\times[N_2]$ with both $X, Y$ randomly selected and size 25. The perturbation size $\delta = 50,000$. After inverting the altered representations, we create valid videos by taking the magnitude of complex numbers and clipping pixel intensities within $[256]$. For comparison, we embed classic triggers proposed for image data, including **BadNet** (Gu, Dolan-Gavitt, and Garg 2017), **Blend** (Chen et al. 2017), **SIG** (Barni, Kallas, and Tondi 2019), **WaNet** (Nguyen and Tran 2021), **FTtrojan** (Wang et al. 2022), in each frame of the video (these are the attacks proposed by Hammoud et al. (2023)). We follow their poisoning pipeline and appropriately modify the attack hyper-parameters to achieve effective attacks. For all the attacks, the triggers are embedded into the down-sampled videos. We defer the detailed attack settings in Apdx. C.

**Evaluation Metrics**: The effectiveness of backdoor attacks is evaluated by 1) accuracy (**ACC**) – the fraction of clean test samples that are correctly classified to their ground truth classes; and 2) attack success rate (**ASR**) – the fraction of backdoor-triggered samples that are mis-classified to the target class. The ACC and ASR are measured on the *same* test set. For an effective backdoor attack, the ACC of the poisoned model is close to that of the clean model, and the ASR is as high as possible. Besides, we evaluate the imperceptibility of the proposed trigger by the peak signal-to-noise ratio (**PSNR**) (Horé and Ziou 2010) and structural similarity index (**SSIM**) (Wang et al. 2004). For both metrics, a higher value indicates better imperceptibility to humans.

**Defenses**: To further demonstrate the effectiveness of the proposed attack, we examine its stealthiness against several classic backdoor detection and mitigation methods, including **NC** (Wang et al. 2019), **PT-RED** (Xiang, Miller, and Kesidis 2020), **TABOR** (Guo et al. 2019), **AC** (Chen et al. 2019), **STRIP** (Gao et al. 2019), **NAD** (Li et al. 2021c), **FP** (Liu, Dolan-Gavitt, and Garg 2018), and **DBD** (Huang et al. 2022)[4]. NC proposes both methods for detection and mitigation, we respectively denote them as **NC-D** and **NC-M**. For all the methods, we set their hyper-parameters following the suggestions in their original papers. More details, including pattern estimation, detection statistics, and hyper-parameter settings are shown in Apdx. D and E.

## 4.2 Attack Effectiveness

The ACCs and ASRs of all victim models trained on various video recognition datasets poisoned by the proposed attack using *DFT* are shown in Tab. 1. The results of using other transform methods including DCT, DWT, and RT, are shown in Tab. 5 in Apdx. F. The proposed attack successfully compromises all models, achieving an ASR (as indicated by DFT in Tab.1) of over 95% in most scenarios. *This highlights the susceptibility of representative video recognition models to adversarial threats.* On the other hand, the ACCs of the compromised models remain close to clean baselines in most cases, with an average decrease of less than 5%. The subtle drop in ACC, especially when benchmarked against image backdoor attacks such as BadNet, makes it difficult for

[2]The appendix is available at https://arxiv.org/abs/2308.11070.

[3]To reduce computation cost, we form a subset of GSL by instances from 50 randomly selected classes.

[4]Due to extremely expensive computational cost, we are not able to apply several popular mitigation methods, such as I-BAU (Zeng et al. 2022).

| Model | | UCF-101 | | HMDB-51 | | GSL | |
|---|---|---|---|---|---|---|---|
| | | Clean | DFT | Clean | DFT | Clean | DFT |
| Slow-Fast | ACC | 84.5 | 81.0 | 60.6 | 59.8 | 95.3 | 89.6 |
| | ASR | - | 97.9 | - | 97.6 | - | 99.9 |
| Res-(2+1)D | ACC | 77.4 | 69.9 | 53.6 | 53.0 | 95.6 | 91.1 |
| | ASR | - | 99.4 | - | 99.6 | - | 100.0 |
| S3D | ACC | 90.6 | 90.3 | 69.3 | 67.5 | 95.4 | 93.8 |
| | ASR | - | 96.9 | - | 90.4 | - | 100.0 |
| I3D | ACC | 89.0 | 87.5 | 66.6 | 59.0 | 94.2 | 92.2 |
| | ASR | - | 97.3 | - | 85.0 | - | 99.5 |

Table 1: ACCs and ASRs (in %) of SlowFast, Res2+1D, S3D, and I3d trained on UCF-101, HMDB-51, and GSL datasets poisoned by the proposed attack using DFT.

users to notice abnormal behaviors of the DNN during training. Besides, the proposed attack utilizes a complex temporal pattern and performs comparably to classic backdoor triggers (as shown in the first column in Tab.2).

In our current trigger generation, we simply assume the attacker is aware of the down-sampling strategy applied during model training and utilizes the same strategy before trigger embedding. However, in practice, the attacker might have no information of the down-sampling strategy during training. Hence, to simulate the realistic attacking scenario, we embed the backdoor trigger in the original 32-frame video, then the triggered samples are down-sampled to 16 frames. The ACC and ASR of S3D trained on UCF-101 under the above attack scenario are 90.98% and 96.36%, respectively, highlighting the effectiveness of the proposed attack and the vulnerability of the current video recognition systems in realistic attack scenarios.

### 4.3 Resistance to Backdoor Defenses

To further demonstrate the effectiveness of the proposed attack, we apply classic backdoor detection and mitigation methods to the SlowFast models trained on UCF-101 poisoned by all attacks. The details of these defense techniques are shown in Apdx. D and E. Following the suggestion in NC, we set the threshold of NC-D, PT-RED and TABOR at 2 – a class with an index larger than 2 is deemed as the true target class. The anomaly indices of the true target class (class 0) computed by the three detection methods are shown in Tab. 3. All attacks except ours are detected by existing methods, while our proposed attack, distributing the trigger throughout the video, successfully evades all detections We apply STRIP with a threshold set to achieve a 15% false positive rate (FPR) – the fraction of clean test instances mis-identified as triggered instances. The corresponding true positive rate (TPR) – the fraction of triggered instances correctly detected – is presented in Tab. 3. The instances embedded with salient patterns (BadNet, Blending, and SIG) are easily detected by STRIP, with TPRs higher than 90%, while the instances with less perceptible triggers (the proposed trigger, WaNet, and FT-trojan) are not. Furthermore, we apply AC on the poisoned training sets. Following their suggested detection threshold, the TPR and FPR of AC are shown in Tab. 2. It fails on all the attacks. AC is unable to detect any poisoned samples for most cases, while falsely detects around 10% samples on all the attacks.

We then deploy several backdoor mitigation methods on the compromised models, including DBD, NAD, FP, and NC-M. The ACCs and ASRs of the victim models after mitigation are shown in Tab. 2 DBD seems less effective on video recognition models than image classifiers possibly due to the complexity of both models and datasets. It fails to suppress the ASR, but reduces the ACC on all the attacks. NAD effectively counters WaNet and FTtrojan. In contrast, other attacks exhibit resistance to distillation-based mitigation methods. FP successfully mitigates FTtrojan, while failing on the remaining attacks. Although NC-M degrades the ACC due to fine-tuning on a small dataset, it suppresses the ASR of most of the attacks, except for ours and Blend. We also demonstrate that the proposed attack could survive the random video pre-processing methods in Apdx. J.

The advanced defense method, UNICORN, is infeasible due to the excessive computational cost required to optimize the potential transform methods and the associated triggers. Besides, even if the defender is aware of the transformed space, such as the frequency domain determined by DFT, reverse-engineering the trigger remains challenging. The potentially perturbed frequency range extends infinitely. Without knowledge of the attacker-specified frequencies, trigger estimation becomes prohibitively expensive.

### 4.4 Resistance to Human Observers

All the backdoor triggers in this paper are visualized in Fig. 1. We evaluate the imperceptibility of triggers to human perception using PSNR and SSIM, standard metrics in image quality assessment. For a more accurate evaluation, we employ localized quality metrics, with further details provided in Apdx. H. Tab. 4 shows the results of all triggers. Triggers from BadNet, Blend, SIG, and WaNet are relatively obvious to the human eye, while those from the proposed attack and FT-trojan are more imperceptible. The heightened imperceptibility arises since both attacks introduce triggers by perturbing the frequency domain, leading to minimal alterations per pixel. FTtrojan is slightly more imperceptible than ours, due to its gentler frequency domain perturbations. This also explains its relatively lower ASR.

### 4.5 Collateral Damage

Collateral damage refers to a phenomenon where perturbations in specific areas of the transformed domain could unintentionally activate the attack, even if they *mismatch* those intentionally introduced during training. We observe this phenomenon in our experiments and illustrate it by presenting the test ASR as a function of $k_0$ (in Fig. 2), where the test instances are triggered on the set $\{k_0, \ldots, k_0 + 9\} \times X \times Y$. The triggered instances are fed to four compromised models trained on UCF-101[5] and the results are shown in Fig. 2(a). The black dashed line denotes the frequencies perturbed during training ($k_0 = 35$). The figure shows that the backdoor is successfully activated by lower-frequency perturbations

[5]Note that we only vary the frequencies for perturbation at test-time, and the compromised models are untouched.

| Attack | No Defense | | DBD | | NAD | | FP | | NC-M | | AC | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | ACC | ASR | TPR | FPR |
| DFT(ours) | 81.0 | 97.9 | 41.9 | 97.3 | 75.4 | 86.1 | 80.8 | 86.8 | 81.0 | 97.9 | 0.0 | 9.0 |
| BadNet | 83.5 | 98.6 | 31.8 | 99.7 | 82.2 | 98.3 | 83.0 | 61.7 | 77.4 | 49.3 | 0.0 | 9.8 |
| Blend | 83.3 | 99.4 | 39.4 | 99.6 | 64.5 | 99.9 | 83.1 | 92.9 | 77.8 | 84.9 | 0.0 | 7.8 |
| SIG | 83.8 | 99.9 | 37.2 | 99.9 | 80.1 | 99.9 | 83.4 | 96.8 | 80.8 | 22.9 | 34.9 | 11.6 |
| WaNet | 82.0 | 95.3 | 46.9 | 50.1 | 80.5 | 2.1 | 80.8 | 89.7 | 80.8 | 1.1 | 22.3 | 8.8 |
| FT-trojan | 76.6 | 83.4 | 22.2 | 68.7 | 83.1 | 1.2 | 79.7 | 8.0 | 75.8 | 0.7 | 0.0 | 8.7 |

Table 2: ACCs and ASRs (in %) of the victim model before and after the mitigation methods are applied, and the TPR and FPR of AC. All the mitigation and detection methods are applied to the SlowFast trained on poisoned UCF-101 datasets.

(a) clean (b) ours (c) BadNet (d) SIG (e) Blend (f) WaNet (g) FT-trojan

Figure 1: Examples of Backdoor Triggers.

| Detection | DFT | BadNet | Blend | SIG | WaNet | FTtrojan |
|---|---|---|---|---|---|---|
| NC-D | **0.1** | 134.1 | 173.1 | 269.2 | 166.7 | 3.4 |
| PT-RED | **1.6** | 13.5 | 9.6 | 2.3 | 23.5 | 2.7 |
| TABOR | **0.2** | 7.2 | 18.7 | 15.5 | 120.6 | **1.9** |
| STRIP | **25.7%** | 93.5% | 96.8% | 98.8% | **24.5%** | **24.9%** |

Table 3: Anomaly index of the true target class (class 0) computed by NC-D, PT-RED, and TABOR, and the TPR of STRIP at test-time. All the detection methods are applied to the SlowFast trained on poisoned UCF-101 datasets.

| Metric | DFT | BadNet | Blend | SIG | WaNet | FTtrojan |
|---|---|---|---|---|---|---|
| PSNR | **41.6** | 36.3 | 20.2 | 28.7 | 34.3 | **47.4** |
| SSIM | **0.972** | 0.173 | 0.435 | 0.515 | 0.842 | **0.982** |

Table 4: Imperceptibility of all backdoor triggers measured by PSNR and SSIM.

for any model. The ASR gradually declines as the perturbation affects higher frequencies. Specifically, the ASR of SlowFast drops rapidly when the perturbed frequency exceeds 50, whereas the ASR of Res(2+1)D remains high. This suggests that Res(2+1)D might be more vulnerable to adversarial perturbations compared to other model architectures, while SlowFast demonstrates relatively higher robustness.

We attribute this phenomenon to additional operations other than trigger embedding, such as pixel clipping. These operations would introduce unintended perturbations to all components in the transformed representation. Hence, during poisoned training, the victim DNN might inadvertently associate these unintended perturbations with the target class. We further illustrate the results for attacks using DFT and DCT in Fig. 2(b) While both DFT and DCT exhibit this collateral damage, their effects manifest differently across frequency bands. Specifically, DFT's unintended effects are

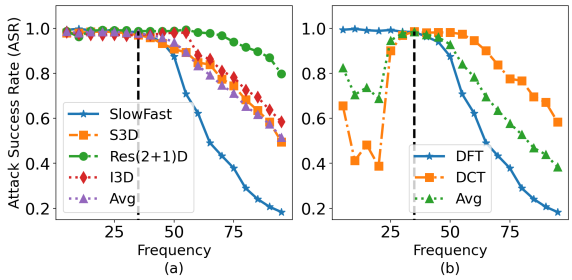


Figure 2: (a) Collateral damage on SlowFast, Res(2+1)D, S3D, and I3D trained on UCF-101 poisoned by the our DFT attack. (b) Collateral damage on SlowFast trained on UCF-101 poisoned by the proposed attack using DFT and DCT.

primarily concentrated in the lower frequencies, whereas DCT shows these effects more in the mid-frequency range.

### 4.6 Case Study: ASR v.s. Influential Factors

In this section, we examine the impact of various influential factors of the attack on the learning of backdoor mapping and model vulnerability. Fig. 3 illustrates how ASR is affected by four different factors. These experiments were conducted on the UCF-101 dataset compromised by a DFT-based attack, with more detailed settings available in Apdx. I. The results suggest that *attackers can easily choose suitable attack hyper-parameters for an effective attack across various models*, highlighting the importance of strengthening the defenses of video recognition systems.

**Poisoning Ratio.** The poisoning ratio represents the fraction of training samples under the manipulation of the attacker. Unsurprisingly, the ASRs for all the models increase as the attack is strengthened. With just 5% of the training data poisoned, both SlowFast and Res(2+1)D are compro-

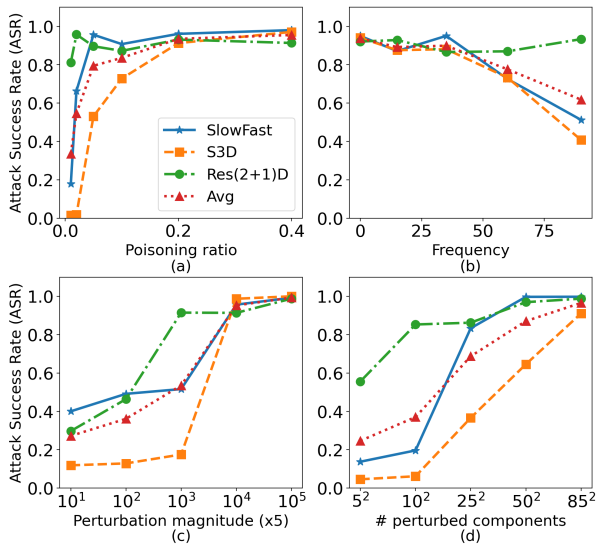


Figure 3: The ASR of various models trained on UCF-101 as a function of (a) poisoning ratio (b) frequencies for adding perturbation (c) perturbation magnitude (d) the number of perturbed components.

mised with ASRs greater than 90%. Notably, even if the attacker merely manipulates 1% of the training data, the attack is effective to Res(2+1)D with an ASR of around 80%, while it is hard for the other models to build the backdoor mapping. We believe Res(2+1)D is more susceptible to the adversarial attack compared with other CNN-based models. As a result, it prioritizes learning the backdoor mapping over the normal mapping during training, as shown in Fig. 6 in Apdx. I.

**Frequency.** The frequency refers to a range of frequencies $F = \{k_0, \ldots, k_0 + 9\}$ where the attacker adds perturbations during *training*, and we fix the length of the range at 10. Fig. 3 (b) displays ASR as a function of $f_s$. Generally speaking, low-frequency components in images and videos represent the primary content, including large-scale structures, broad shapes, and general illumination. By contrast, high-frequency components capture details, edges, and textures. It is not surprising that the attack becomes less effective as the perturbed frequency increases (except for Res(2+1)D), since perturbing low-frequency components generates more salient features than high-frequency components. However, there is still a sufficiently large range of frequencies (0-50) for devising effective attacks. Similar to the observation on the poisoning ratio, Res(2+1)D is vulnerable to a broader range of frequencies than the other models.

**Number of perturbed components.** The total number of perturbed components over all selected frequencies is the size of $\mathcal{I}$ defined in Eq. 1. These components are randomly chosen. Similar to the observation on perturbation magnitude, the ASR monotonically increases with the number of components being altered. With only 625 components (1.2% of the total components in a $224 \times 224$ spectrum) in each frequency being perturbed, the attack successfully compromises the Res(2+1)D and SlowFast with ASR of around 80%. S3D is resistant to the number of perturbed components – the backdoor is planted with 7225 components being perturbed (14% of the total components).

**Perturbation magnitude.** The perturbation magnitude (denoted as $\delta$ in the trigger embedding function given by Eq. 1) represents the amount of change applied to each selected component. The ASR monotonically increases with the perturbation magnitude: as the perturbation magnitude rises, the trigger becomes more salient in the original space. Besides, the ASR gets a significant boost from a magnitude of $5 \cdot 10^3$ to $5 \cdot 10^5$. Hence, from the aspect of the attacker, choosing the right perturbation magnitude is straightforward – with a few components being perturbed, applying a higher magnitude of perturbation can lead to more potent attacks.

## 5 Limitation and Future Work

In this paper, we focus on all-to-one attacks for the following reasons. First, training video recognition models is inherently challenging given the increased complexity of video data compared to traditional image data. Second, under attack settings such as many-to-one and one-to-one, there may not be a sufficient number of perturbed training samples to establish a solid mapping from the backdoor trigger(s) to the desired target classes. However, many-to-one backdoor attacks would be more practical in scenarios, *e.g.*, sign language translation. The attacker would only aim to map a set of few words to another word(s) to introduce misinformation to the expression. Besides, there is no *feasible* backdoor defense strategy proposed for complex triggers and complicated datasets. UNICORN (Wang et al. 2023) proposes a general framework to estimate a potential backdoor pattern embedded in any transformed space. However, it is practically infeasible due to the extremely high computation cost for estimating the trigger and the transform methods. Also, it fails to detect if a given model is backdoor compromised. We leave addressing the above problems as future works.

Finally, the observed collateral damage raises an unanswered question about its impact on attack stealthiness against defenses and model robustness. We suspect it does, as seen in WaNet (Nguyen and Tran 2021), where similar impact brought by clipping is observed. This work suggests that without robust designs like noise mode, the attacked model is easily detectable.

## 6 Conclusion

In this paper, we propose a general framework for embedding an imperceptible, temporal-distributed backdoor trigger in videos. Notably, it exhibits invisibility not only to human eyes, but also to current backdoor defense strategies. Empirical experiments across various benchmark datasets and popular video recognition model architectures demonstrate the effectiveness of our attack. Furthermore, we explore the impact of several factors on the effectiveness of the proposed attack, providing an enriched perspective on the vulnerabilities in video recognition systems, emphasizing the urgency for advancing robustness measures in this domain.

## Acknowledgments

This work was supported by a research gift from Cisco Systems. We thank Jiaqi Wang for additional computational resources and helpful discussions. We also thank all the individuals who supported us throughout this journey.

## References

Adaloglou, N.; Chatzis, T.; Papastratis, I.; Stergioulas, A.; Papadopoulos, G. T.; Zacharopoulou, V.; Xydopoulos, G. J.; Atzakas, K.; Papazachariou, D.; and Daras, P. 2022. A Comprehensive Study on Deep Learning-Based Methods for Sign Language Recognition. *IEEE Transactions on Multimedia*.

Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; and Baskurt, A. 2011. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*, 29–39. Springer.

Barni, M.; Kallas, K.; and Tondi, B. 2019. A New Backdoor Attack in CNNS by Training Set Corruption Without Label Poisoning. In *ICIP*.

Barten, P. G. 1999. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE press.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*.

Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I. M.; and Srivastava, B. 2019. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. In *AAAI*.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv:1712.05526*.

Chou, E.; Tramèr, F.; and Pellegrino, G. 2020. SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems. In *2020 IEEE Security and Privacy Workshops*.

Dai, J.; Chen, C.; and Li, Y. 2019. A Backdoor Attack Against LSTM-Based Text Classification Systems. *IEEE Access*.

Doan, B. G.; Abbasnejad, E.; and Ranasinghe, D. C. 2020. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems. In *Annual Computer Security Applications Conference*.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Du, M.; Jia, R.; and Song, D. 2020. Robust Anomaly Detection and Backdoor Attack Detection Via Differential Privacy. In *ICLR*.

Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Bouridane, A.; and Beghdadi, A. 2021. A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence*.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. SlowFast Networks for Video Recognition. arXiv:1812.03982.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. STRIP: a defence against trojan attacks on deep neural networks. In *ACSAC*.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Guan, J.; Tu, Z.; He, R.; and Tao, D. 2022. Few-shot Backdoor Defense Using Shapley Estimation. In *CVPR*.

Guo, W.; Wang, L.; Xing, X.; Du, M.; and Song, D. 2019. TABOR: A Highly Accurate Approach to Inspecting and Restoring Trojan Backdoors in AI Systems. *arXiv:1908.01763*.

Hammoud, H. A. A. K.; Liu, S.; Alkhrashi, M.; Albalawi, F.; and Ghanem, B. 2023. Look, Listen, and Attack: Backdoor Attacks Against Video Action Recognition. *CVPR*.

Hara, K.; Hirayama, T.; Kashima, H.; and Satoh, Y. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *CVPR*.

Horé, A.; and Ziou, D. 2010. Image Quality Metrics: PSNR vs. SSIM. In *International Conference on Pattern Recognition*.

Huang, K.; Li, Y.; Wu, B.; Qin, Z.; and Ren, K. 2022. Backdoor Defense via Decoupling the Training Process. In *ICLR*.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T. A.; and Serre, T. 2011. HMDB: A large video database for human motion recognition. In *ICCV*.

Li, C.; Pang, R.; Xi, Z.; Du, T.; Ji, S.; Yao, Y.; and Wang, T. 2022a. Demystifying Self-supervised Trojan Attacks. *arXiv*, abs/2210.07346.

Li, D.; Rodriguez, C.; Yu, X.; and Li, H. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *WACV*.

Li, S.; Xue, M.; Zhao, B.; Zhu, H.; and Zhang, X. 2021a. Invisible Backdoor Attacks on Deep Neural Networks Via Steganography and Regularization. *IEEE Transactions on Dependable and Secure Computing*.

Li, X.; Kesidis, G.; Miller, D. J.; and Lucic, V. 2021b. Backdoor Attack and Defense for Deep Regression. *arXiv:2109.02381*.

Li, X.; Xiang, Z.; Miller, D. J.; and Kesidis, G. 2022b. Test-Time Detection of Backdoor Triggers for Poisoned Deep Neural Networks. In *ICASSP*.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021c. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. In *ICLR*.

Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *RAID*.

Liu, M.; Liu, H.; and Chen, C. 2016. Spatiotemporal LSTM with Trust Gates for 3D Human Action Recognition. In *European Conference on Computer Vision*, 816–833. Springer.

Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.; Zhai, J.; Wang, W.; and Zhang, X. 2018. Trojaning Attack on Neural Networks. In *NDSS*.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2021. Video Swin Transformer. arXiv:2106.13230.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

Ng, J. Y.-H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; and Toderici, G. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4694–4702.

Nguyen, T. A.; and Tran, A. T. 2021. WaNet - Imperceptible Warping-based Backdoor Attack. In *ICLR*.

Qiu, Z.; Yao, T.; and Mei, T. 2017. Learning spatiotemporal representation with pseudo-3d residual networks. In *ICCV*.

Saha, A.; Subramanya, A.; and Pirsiavash, H. 2020. Hidden Trigger Backdoor Attacks. In *AAAI*.

Saleh, K.; Hossny, M.; and Nahavandi, S. 2019. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *ICRA*.

Shafahi, A.; Huang, W. R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; and Goldstein, T. 2018. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *NeurIPS*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv*, abs/1212.0402.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral Signatures in Backdoor Attacks. In *NeurIPS*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*.

Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; and Paluri, M. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. arXiv:1711.11248.

Turner, A.; Tsipras, D.; and Madry, A. 2019. Label-Consistent Backdoor Attacks. *ArXiv*, abs/1912.02771.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *IEEE Symposium on Security and Privacy*.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An Invisible Black-Box Backdoor Attack Through Frequency Domain. In *ECCV*.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*.

Wang, Z.; Mei, K.; Zhai, J.; and Ma, S. 2023. UNICORN: A Unified Backdoor Trigger Inversion Framework. In *ICLR*.

Wu, D.; and Wang, Y. 2021. Adversarial Neuron Pruning Purifies Backdoored Deep Models. In *NeurIPS*.

Xia, J.; Wang, T.; Ding, J.; Wei, X.; and Chen, M. 2022. Eliminating Backdoor Triggers for Deep Neural Networks Using Attention Relation Graph Distillation. In *IJCAI*.

Xiang, Z.; Miller, D. J.; Chen, S.; Li, X.; and Kesidis, G. 2021. A Backdoor Attack against 3D Point Cloud Classifiers. *ICCV*.

Xiang, Z.; Miller, D. J.; and Kesidis, G. 2020. Detection of Backdoors in Trained Classifiers Without Access to the Training Set. *IEEE Transactions on Neural Networks and Learning Systems*.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *ECCV*.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z.; Jin, M.; and Jia, R. 2022. Adversarial Unlearning of Backdoors via Implicit Hypergradient. In *ICLR*.

Zhao, S.; Ma, X.; Zheng, X.; Bailey, J.; Chen, J.; and Jiang, Y. 2020. Clean-Label Backdoor Attacks on Video Recognition Models. In *CVPR*.

Zheng, R.; Tang, R.; Li, J.; and Liu, L. 2022. Data-Free Backdoor Removal Based on Channel Lipschitzness. In *ECCV*.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; and Xie, X. 2016. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 3291–3297.