Look, Listen, and Attack: Backdoor Attacks Against Video Action Recognition

Hasan Abed Al Kader Hammoud[1] Shuming Liu[1] Mohammed Alkhrashi[2]
Fahad AlBalawi[3] Bernard Ghanem[1]
[1] KAUST [2] SDAIA [3] Taif University

Abstract

Deep neural networks (DNNs) are vulnerable to a class of attacks called "backdoor attacks", which create an association between a backdoor trigger and a target label the attacker is interested in exploiting. A backdoored DNN performs well on clean test images, yet persistently predicts an attacker-defined label for any sample in the presence of the backdoor trigger. Although backdoor attacks have been extensively studied in the image domain, there are very few works that explore such attacks in the video domain, and they tend to conclude that image backdoor attacks are less effective in the video domain. In this work, we revisit the traditional backdoor threat model and incorporate additional video-related aspects to that model. We show that poisoned-label image backdoor attacks could be extended temporally in two ways, statically and dynamically, leading to highly effective attacks in the video domain. In addition, we explore natural video backdoors to highlight the seriousness of this vulnerability in the video domain. And, for the first time, we study multi-modal (audiovisual) backdoor attacks against video action recognition models, where we show that attacking a single modality is enough for achieving a high attack success rate.

1. Introduction

A fundamental requirement for the deployment of deep neural networks (DNNs) in real-world tasks is their safety and robustness against possible vulnerabilities and security breaches. This requirement is, in essence, the motivation behind exploring adversarial attacks. One particularly interesting adversarial attack is "backdoor attacks". Backdoor attacks or neural trojan attacks explore the scenario in which a user with limited computational capabilities downloads pretrained DNNs from an untrusted party or outsources the training procedure to such a party that we refer to as the adversary. The adversary provides the user with a model that performs well on an unseen validation set, but produces a pre-defined class label in the presence of an attacker-defined trigger called the backdoor trigger. The association between the backdoor trigger and the attacker-specified label is created by training the DNN on poisoned training samples, which are samples polluted by the attacker's trigger [40]. In poisoned-label attacks, unlike clean-label attacks, the attacker also switches the label of the poisoned samples to the intended target label.

Considerable attention has been paid to explore backdoor attacks and defenses for 2D image classification models [6, 23, 26]. However, little attention has been paid to exploring backdoor attacks and defenses against video action recognition models. The disappointing conclusion uncovered by [88] regarding the limited effectiveness of image backdoor attacks on videos stunted further development of video backdoor attacks. Unfortunately, the attacks considered in [88] were limited to only visible patch-based clean-label attacks. Moreover, [88] directly adopted the 2D backdoor attack threat model without incorporating important video-specific considerations.

To this end, and as opposed to [88]. we first revisit and revise the commonly adopted 2D *poisoned-label* backdoor threat model by incorporating additional constraints that are inherently imposed by video systems. These constraints arise due to the presence of the temporal dimension. We then explore two ways to extend image backdoor attacks to incorporate the temporal dimension into the attack to enable more video-specific backdoor attacks. In particular, image backdoor attacks could be either extended statically by applying the same attack to each frame of the video or dynamically by adjusting the attack parameters differently for each frame. Then, three novel natural video backdoor attacks are presented to highlight the seriousness of the risks associated with backdoor attacks in the video domain. We then test the attacked models against three 2D backdoor defenses and discuss the reason behind the failure of those methods. We also study, for the first time, audiovisual backdoor attacks, where we ablate the importance and contribution of each modality on the performance of the attack for both late and early fusion settings. We show that attacking a single modality is enough to achieve a high attack success rate.

Contributions. Our contributions are twofold. **(1)** We revisit the traditional backdoor attack threat model and incor-

porate video-related aspects, such as video subsampling and spatial cropping, into the model. We also extend existing image backdoor attacks to the video domain in two different ways, statically and dynamically, after which we propose three novel natural video backdoor attacks. Through extensive experiments, we provide evidence that the previous perception of image backdoor attacks in the video domain is not necessarily true, especially in the poisoned-label attack setup. **(2)** To the best of our knowledge, this work is the first to investigate audiovisual backdoor attacks against video action recognition models.

2. Related Work

Backdoor Attacks. Backdoor attacks were first introduced in [23]. The attack, called BadNet, was based on adding a patch to the corner of a subset of training images to create a backdoor that could be triggered by the attacker at will. Following BadNet, [45] proposed optimizing for the values of the patch to obtain a more effective backdoor attack. Shortly after the development of patch-based backdoor attacks, the community realized the importance of adding an invisibility constraint to the design of backdoor triggers to bypass any human inspection. Works such as [10] proposed blending the backdoor trigger with the image rather than stamping it. [38] generated backdoor attacks using the least significant bit algorithm. [53] generated warping fields to warp the image content as a backdoor trigger. [15] went one step further and designed learnable transformations to generate optimal backdoor triggers. After many attacks were proposed in the spatial domain [11, 38, 41, 46, 57, 58, 68, 73, 76], and others in the latent representation domain [14, 54, 81, 89, 92], [20, 26, 72, 83, 85] proposed to switch attention to the frequency domain. [26] utilized frequency heatmaps proposed in [82] to create backdoor attacks that target the most sensitive frequency components of the network. [20] proposed blending low frequency content from a trigger image with training images as a poisoning technique. *In our work, we extend the 2D backdoor threat model to the video domain by incorporating video-related aspects into it. We also extend five image backdoor attacks into the video domain and propose three natural video backdoor attacks.*

Backdoor Defenses. Backdoor attack literature was immediately opposed by various defenses. Backdoor defenses are generally of five types: preprocessing-based [13, 48, 56], model reconstruction-based [39, 43, 75, 84, 90], trigger synthesis-based [24, 25, 30, 44, 55, 59, 63, 69], model diagonsis-based [16, 35, 47, 77, 91], and sample-filtering based [1, 9, 21, 27, 31, 62, 64]. Early backdoor defenses such as [69] hypothesized that backdoor attacks create a shortcut between all samples and the poisoned class. Based on that, they solved an optimization problem to find whether a trigger of an abnormally small norm exists that would flip all samples to one label. Later, multiple improved iterations of this method were proposed, such as [24, 44, 84]. Fine pruning [43] suggested that the backdoor is triggered by particular neurons that are dormant in the absence of the trigger. Therefore, the authors proposed pruning the least active neurons on clean samples. STRIP [21] showed that blending clean samples with other clean samples would yield a higher entropy compared to when clean images are blended with poisoned samples. Activation clustering [9] uses KMeans to cluster the activations of an inspection, a potentially poisoned data set, into two clusters. A large silhouette distance between the two clusters would uncover the poisoned samples. *In our work, we show that current image backdoor attacks have limited effectiveness in defending against backdoor attacks in the video domain, especially against the proposed natural video attacks.*

Video Action Recognition. Video action recognition models, which only leverage the raw frames of a video, can be categorized into two categories, CNN-based networks and transformer-based networks. 2D CNN-based methods are built on top of pretrained image recognition networks with well-designed modules to capture the temporal relationship between multiple frames [42, 50, 70, 71]. Those methods are computationally efficient as they use 2D convolutional kernels. To learn stronger spatial-temporal representations, 3D CNN-based methods were proposed. These methods utilize 3D kernels to jointly leverage the spatio-temporal context within a video clip [18, 19, 65, 66]. To better initialize the network, I3D [8] inflated the weights of 2D pretrained image recognition models to adapt them to 3D CNNs. Realizing the importance of computational efficiency, S3D [79] and R(2+1)D [67] proposed to disentangle spatial and temporal convolutions to reduce computational cost. Recently, transformer-based action recognition models were able to achieve better performance in large training data sets compared to CNN-based models, *e.g.* [5, 7, 17, 49]. *In this work, we test backdoor attacks against three action recognition architectures, namely I3D, SlowFast, and TSM.*

Audiovisual Action Recognition. In addition to frames, a line of action recognition models [2, 28, 29, 52] has used the accompanying audio to better understand activities such as "playing music" or "washing dishes". To take advantage of existing CNN and transformer-based models, the Log-Mel spectrogram was introduced to convert audio data from a non-structured signal into a 2D representation in time and frequency usable by these models [3, 4, 36, 78]. Current audiovisual action recognition methods are divided into two categories based on when the audio and visual signals are merged in the recognition pipeline: early fusion and late fusion. Early fusion combines features before classification, which can better capture features [33, 78]. The disadvantage of early fusion is that there is a higher risk of overfitting to the training data [60]. Late fusion, on the other hand, treats the video and audio networks separately, and the predictions

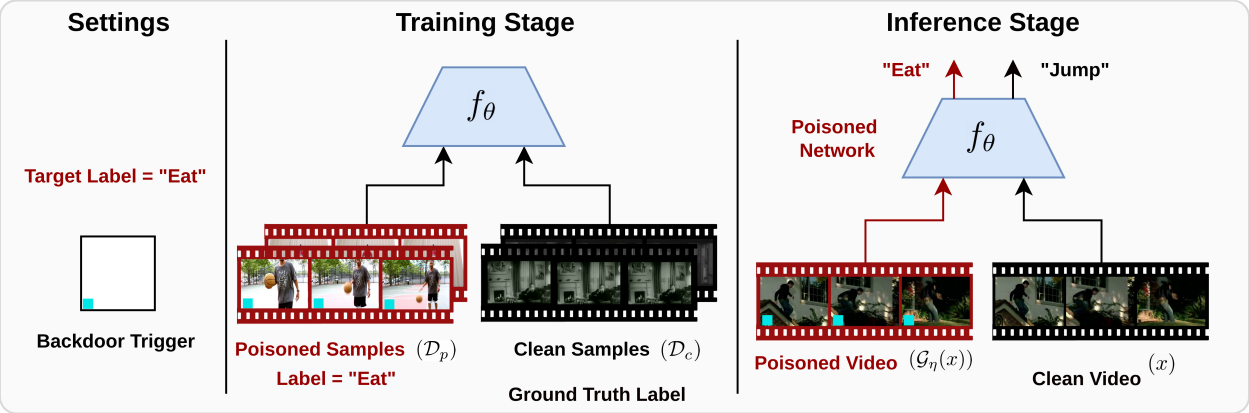

Figure 1. **Traditional Backdoor Attack Pipeline.** After selecting a backdoor trigger and a target label, the attacker poisons a subset of the training data referred to as the poisoned dataset ($\mathcal{D}_p$). The label of the poisoned dataset is fixed to a target poisoning label specified by the attacker. The attacker trains jointly on clean (non-poisoned) samples ($\mathcal{D}_c$) and poisoned samples leading to a backdoored model, which outputs the target label in the presence of the backdoor trigger.

of each network are carried out independently, after which the logits are aggregated to make a final prediction [22]. *For the first time, we test backdoor attacks against audiovisual action recognition networks in both late and early fusion setups.*

3. Video Backdoor Attacks

3.1. The Traditional Threat Model

The commonly adopted threat model for backdoor attacks dates back to the works that studied those attacks against 2D image classification models [23]. The victim outsources the training process to a trainer who is given access to both the victim's training data and the network architecture. The victim only accepts the model provided by the trainer if it performs well on the victim's private validation set. The attacker aims to maximize the effectiveness of the embedded backdoor attack [40]. We refer to the model's performance on the validation set as clean data accuracy (CDA). The effectiveness of the backdoor attack is measured by the attack success rate (ASR), which is defined as the percentage of test examples not labeled as the target class that are classified as the target class when the backdoor pattern is applied. To achieve this goal, the attacker applies a backdoor trigger to a subset of the training images and then, in the poisoned-label setup, switches the labels of those images to a target class of choice before training begins. A more powerful backdoor attack is one that is visually imperceptible (usually measured in terms of ℓ_2/ℓ_∞-norm, PSNR, SSIM, or LPIPS) but achieves both a high CDA and a high ASR. This is summarized in Figure 1.

More formally, we denote the classifier which is parameterized by θ as $f_\theta : \mathcal{X} \to \mathcal{Y}$. It maps the input $x \in \mathcal{X}$, such as images or videos, to class labels $y \in \mathcal{Y}$. Let $\mathcal{G}_\eta : \mathcal{X} \to \mathcal{X}$ indicate an attacker-specific poisoned image generator that is parameterized by some trigger-specific parameters η. The generator may be image-dependent. Finally, let $\mathcal{S} : \mathcal{Y} \to \mathcal{Y}$ be an attacker-specified label shifting function. In our case, we consider the scenario in which the attacker is trying to flip all the labels into one particular label, *i.e.* $\mathcal{S} : \mathcal{Y} \to t$, where $t \in \mathcal{Y}$ is an attacker-specified label that will be activated in the presence of the backdoor trigger. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ indicate the training dataset. The attacker splits $\mathcal{D}$ into two subsets, a clean subset $\mathcal{D}_c$ and a poisoned subset $\mathcal{D}_p$, whose images are poisoned by $\mathcal{G}_\eta$ and labels are poisoned by $\mathcal{S}$. The poisoning rate is the ratio $\alpha = \frac{|\mathcal{D}_p|}{|\mathcal{D}|}$, generally a lower poisoning rate is associated with a higher clean data accuracy. The attacker typically trains the network by minimizing the cross-entropy loss on $\mathcal{D}_c \cup \mathcal{D}_p$, *i.e.* minimizes $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_c\cup\mathcal{D}_p}[\mathcal{L}_{CE}(f_\theta(\mathbf{x}), y)]$. The attacker aims to achieve high accuracy on the user's validation set $\mathcal{D}_{val}$ while being able to trigger the poisoned-label, t, in the presence of the backdoor trigger, *i.e.* $f_\theta(\mathcal{G}_\eta(\mathbf{x})) = t,\ \forall x \in \mathcal{X}$ (ideally).

3.2. From Images to Videos

Unlike images, videos have an additional dimension, the temporal dimension. This dimension introduces new rules to the game between the attacker and the victim. More precisely, the attacker now has an additional dimension to hide the backdoor trigger, leading to a higher level of imperceptibility. The backdoor attack could be applied to all the frames or a subset of the frames statically, *i.e.* the same trigger is applied to each frame, or dynamically, *i.e.* a different trigger is applied to each frame. On the other hand, the testing pipeline now imposes harsher conditions

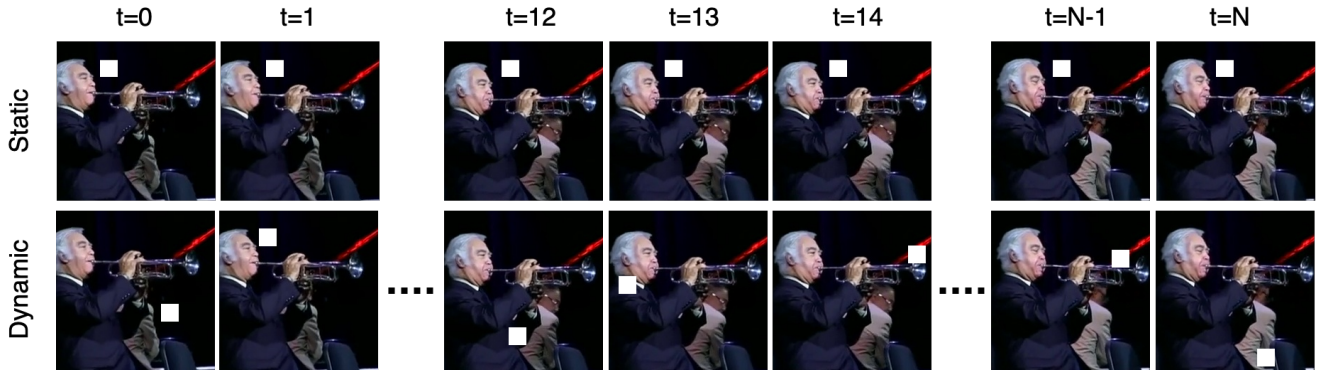

Figure 2. **Static vs Dynamic Backdoor Attacks.** Static backdoor attacks apply the same trigger across all frames along the temporal dimension. On the other hand, dynamic attacks apply a different trigger per frame along the temporal dimension.

against the backdoor attack. Video recognition models tend to test the model on multiple sub-sampled clips with various crops [8, 19, 42] which might, in turn, destroy the backdoor trigger. For example, if the trigger is applied to a single frame, it might not be sampled, and if the trigger is applied to the corner of the image, it might be cropped out. The threat model presented in Subsection 3.1 was directly adopted in [88], which to the best of our knowledge, is the only previous work that considered backdoor attacks for video action recognition.

Our work sheds light on the aforementioned video-related aspects. In Section 4.2, we show the effect of the number of frames poisoned on CDA and ASR. We also show how existing 2D methods could be extended both statically and dynamically to suit the video domain. For example, BadNet [23] applies a fixed patch as a backdoor trigger. The patch could be applied statically using the same pixel values and the same position along the temporal dimension or applied dynamically by changing the position and possibly the pixel values of the patch for each frame. Figure 2 shows a BadNet attack when applied in a static and dynamic way. Additionally, we show how simple yet natural video "artifacts" could be used as backdoor triggers. More specifically, lag in a video, motion blur, and compression glitches could all be used as naturally occurring backdoor triggers.

3.3. Audiovisual Backdoor Attacks

Videos are naturally accompanied by audio signals. Similarly to how the video modality could be attacked, the audio signal could also be attacked. The interesting question that arises is how backdoor attacks would perform in a multimodal setup. In the experiments of Section 4.4, we answer the following questions: (1) What is the effect of having two attacked modalities on CDA and ASR?; (2) What happens if only one modality is attacked and the other is left clean?; (3) What is the difference in performance between late and early fusion in terms of CDA and ASR?

4. Experiments

4.1. Experimental Settings

Datasets. We consider three standard benchmark datasets used in video action recognition: UCF-101 [61], HMDB-51 [37], and Kinetics-Sounds [32]. Kinetics-Sounds is a subset of Kinetics400 that contains classes that can be classified from the audio signal, *i.e.* classes where audio is useful for action recognition [3]. Kinetics-Sounds is particularly interesting for Sections 4.3 and 4.4, where we explore backdoor attacks against audio and audiovisual classifiers.

Network Architectures. Following common practice, for the visual modality, we use a dense sampling strategy to sub-sample 32 frames per video to fine-tune a pretrained I3D network on the target dataset [8]. In Section 4.2, we also show results using TSM [42] and SlowFast [19] networks. All three models adopt ResNet-50 as the backbone and are pretrained on Kinetics-400. Similarly to [3], for the audio modality, a ResNet-18 is trained from scratch on Mel-Spectrograms composed of 80 Mel bands sub-sampled temporally to a fixed length of 256.

Attack Setting. For the video modality, we study and extend the following image-based backdoor attacks to the video domain: BadNet [23], Blend [10], SIG [6], WaNet [53], and FTrojan [72]. We also explore three additional natural video backdoor attacks. For the audio modality, we consider two attacks: sine attack and high-frequency noise attack, both of which we explain later. Following [23, 26, 53], the target class is arbitrarily set to the first class of each data set (class 0), and the poisoning rate is set to 10%. Unless otherwise stated, the considered image backdoor attacks poison all frames of the sampled clips during training and evaluation.

Evaluation Metrics. As is commonly done in the backdoor literature, we evaluate the performance of the model using clean data accuracy (CDA) and attack success rate (ASR) explained in Section 3. CDA represents the usual

Figure 3. **Visualization of 2D Backdoor Attacks.** Image backdoor attacks mainly differ according to the backdoor trigger used to poison the training samples. They could be extended either statically or dynamically based on how the attack is applied across the frames.

validation/test accuracy on an unseen dataset hence measuring the generalizability of the model. On the other hand, ASR measures the effectiveness of the attack when the poison is applied to the validation/test set. In addition, we test the attacked models against some of the early 2D backdoor defenses, more precisely against activation clustering (AC) [9], STRIP [21], and pruning [43].

Implementation Details. Our method is built on MMAction2 library [12], and follows their default training configurations and testing protocols, except for the learning rate and the number of training epochs (check Supplementary). All experiments were run using 4 NVIDIA A100 GPUs.

4.2. Video Backdoor Attacks

Extending Image Backdoor Attacks to the Video Domain. As mentioned in Section 3.2, image backdoor attacks could be extended either statically by applying an attack in the same way across all frames or dynamically by adjusting the attack parameters for different frames. We consider five attacks that differ according to the applied backdoor trigger. BadNet applies a patch as a trigger, Blend blends a trigger image to the original image, SIG superimposes a sinusoidal trigger to the image, WaNet warps the content of the image, and FTrojan poisons a high- and mid- frequency component in the discrete cosine transform (DCT). Figure 3 visualizes all five attacks on the same video frame. Each of the considered methods could be extended dynamically as follows: **BadNet**: change the patch location for each frame; **Blend**: blend a uniform noise that is different per frame; **SIG**: change the frequency of the sine component superimposed with each frame; **WaNet**: generate a different warping field for each frame; **FTrojan**: select a different DCT basis to perturb at each frame. Note that Blend and FTrojan are generally imperceptible. Visualizations and saliency maps for each attack are found in the Supplementary.

Tables 1 and 2 show the CDA and ASR of the I3D models attacked using various backdoor attacks on UCF-101, HMDB-51, and Kinetics-Sounds. Contrary to the conclusion presented in [88], we find that backdoor attacks are actually highly effective in the video domain. The CDA of the attacked models is very similar to that of the clean unattacked model (baseline), surpassing it in some cases.

	UCF101		HMDB51		KineticsSound	
	CDA(%)	ASR(%)	CDA(%)	ASR(%)	CDA(%)	ASR(%)
Baseline	93.95	-	69.59	-	81.41	-
BadNet	93.95	99.63	69.35	98.89	82.97	99.09
Blend	94.29	99.26	68.37	86.73	82.12	97.54
SIG	93.97	99.97	68.50	99.80	82.84	99.87
WaNet	94.05	99.84	68.95	99.61	82.38	99.09
FTrojan	94.16	99.34	68.10	97.52	82.45	97.86

Table 1. **Statically Extended 2D Backdoor Attacks.** Statically extending 2D backdoor attacks to the video domain leads to high CDA and ASR across all three considered datasets.

	UCF101		HMDB51		KineticsSound	
	CDA(%)	ASR(%)	CDA(%)	ASR(%)	CDA(%)	ASR(%)
Baseline	93.95	-	69.59	-	81.41	-
BadNet	94.11	99.97	69.08	99.54	82.25	99.74
Blend	94.21	99.44	67.03	95.95	81.67	95.79
SIG	94.24	100.00	68.63	100.00	82.84	100.00
WaNet	94.29	99.79	69.22	99.80	82.25	99.61
FTrojan	94.16	99.34	67.19	98.69	82.25	95.27

Table 2. **Dynamically Extended 2D Backdoor Attacks.** Dynamically extending 2D backdoor attacks to the video domain leads to high CDA and ASR across all three considered datasets.

Extending attacks dynamically, almost always, improves CDA and ASR compared to extending them statically.

Natural Video Backdoors. A more interesting attack is one that seems natural and could bypass human inspection [51, 74, 80, 87]. There are several natural "glitches" that occur in the video domain and that one could exploit to design a natural backdoor attack. For example, videos might contain some frame lag, motion blur, video compression corruptions, camera focus/defocus, etc. In Table 3, we report the CDA and ASR of three natural backdoor attacks: frame lag (lagging video), video compression glitch (which we refer to as Video Corruption), and motion blur. Interestingly, these attacks could achieve both high clean data accuracy and high attack success rate. It is worth noting that for frame lag, a two-frame lag is used for UCF-101 and a three-frame lag is used for HMDB-51 and Kinetics-Sounds. More details are provided in the Supplementary.

Attacks Against Different Architectures. So far, all at-

	UCF101		HMDB51		KineticsSound	
	CDA(%)	ASR(%)	CDA(%)	ASR(%)	CDA(%)	ASR(%)
Baseline	93.95	-	69.59	-	81.41	-
Frame Lag	92.94	97.20	68.04	98.76	82.51	98.19
Video Corrupt.	94.26	99.87	69.22	99.22	81.74	98.51
Motion Blur	93.97	99.92	68.17	97.52	82.19	99.22

Table 3. **Natural Video Backdoor Attacks.** Natural attacks against video action recognition models could achieve high CDA and ASR while looking completely natural to human inspection.

	SlowFast		TSM	
	CDA(%)	ASR(%)	CDA(%)	ASR(%)
Baseline	96.72	-	94.77	-
BadNet	96.64	99.47	94.69	97.78
SIG	96.70	99.97	94.77	99.47
FTrojan	96.25	98.52	94.21	100.00
Frame Lag	96.43	99.97	94.63	97.96
Video Corruption	96.54	99.76	95.08	98.97
Motion Blur	96.46	99.55	94.50	99.39

Table 4. **Video Backdoor Attacks Against Different Architectures (UCF-101).** When tested against network architectures other than I3D such as TSM and SlowFast, both image and natural backdoor attacks can still achieve high CDA and high ASR.

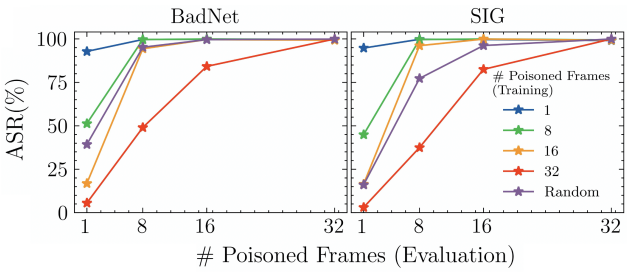

Figure 4. **Effect of the Number of Poisoned Frames (UCF-101).** Different colors refer to different number of frames poisoned during the training of the attacked model. Training the model with a single poisoned frame performs best for various choices of the number of frames poisoned during evaluation.

	Frame Lag	Motion Blur	SIG	BadNet	FTrojan
Elimination Rate(%)	0.00	0.00	34.21	33.77	34.12
Sacrifice Rate(%)	13.08	12.82	15.17	14.25	13.00

Table 5. **Activation Clustering Defense (UCF-101).** Whereas Activation Clustering provides partial success in defending against image backdoor attacks, it fails completely against natural attacks.

tacks have been experimented with against an I3D network. To further explore the behavior of backdoor attacks against other video recognition models, we test a subset of the considered attacks against a 2D based model, TSM, and another 3D based model, SlowFast, on UCF-101. Table 4 shows that all the aforementioned backdoor attacks perform significantly well in terms of CDA and ASR against both TSM and SlowFast architectures. Note that even though TSM is a 2D based model, our proposed natural video backdoor attacks still succeed in attacking it.

Recommendations for Video Backdoor Attacks. As mentioned in Section 3.2, the attacker must select a number of frames to poison per video, keeping in mind that the video will be sub-sampled and randomly cropped during evaluation. Since the attacker is the one who trained the network in the first place, he/she has access to the processing pipeline and could exploit this during the attack. For example, if video processing involves sub-sampling the video into clips of 32 frames and cropping the frames into 224×224 crops, the attacker could pass to the network an attacked video of a temporal length of 32 frames and a spatial size 224×224, hence bypassing sub-sampling and cropping. However, a system could force the user to input a video of a particular length, possibly greater than the length of the sub-sampled clips. This raises an important question regarding how many frames the attacker should poison. Clearly, the smaller the number of frames the attacker poisons, the less detectable the attack is, but does the attack remain effective? In Figure 4, we show the attack success rate of backdoor-attacked models **trained** on clips of 1, 8, 16, and 32 frames, and a randomly sampled number of poisoned frames (out of 32 total frames) when **evaluated** on clips of 1, 8, 16, and 32 poisoned frames (out of 32 total frames). Random refers to training on a varying number of poisoned frames per clip. Note that training the model against the worst-case scenario (single frame), which mimics the case where only one of the poisoned frames is sub-sampled, provides the best guarantees for achieving a high attack success rate.

Defenses Against Video Backdoor Attacks. We explore the effect of extending some of the existing 2D backdoor defenses against video backdoor attacks. Optimization-based defenses are extremely costly when extended to the video domain. For example, Neural Cleanse (NC) [69], I-BAU [84], and TABOR [24] involve a trigger reconstruction phase. The trigger space is now bigger in the presence of the temporal dimension, and therefore, instead of optimizing for a 224×224×3 trigger, the defender has to search for a 32×224×224×3 trigger (assuming 32 frame clips are used), which is both costly and hard to solve. The attacker has the spatial and temporal dimensions to design and embed their attack in, and, therefore, reverse engineering the trigger is quite hard.

We consider three well-known defenses that introduce no computational overhead when adopted to the video domain, namely Activation Cluster (AC) [9], STRIP [21], and pruning [43]. AC computes the activations of a neural network on clean samples (from the test set) and an inspection set of interest which may be poisoned. AC then applies PCA

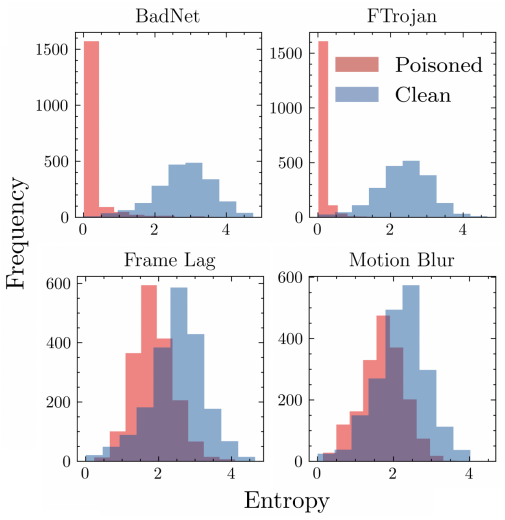

Figure 5. **STRIP Defense (UCF-101).** Whereas the entropy of image backdoor attacks is very low compared to that of clean samples, the proposed natural backdoor attacks have a natural distribution of entropies similar to that of clean samples.

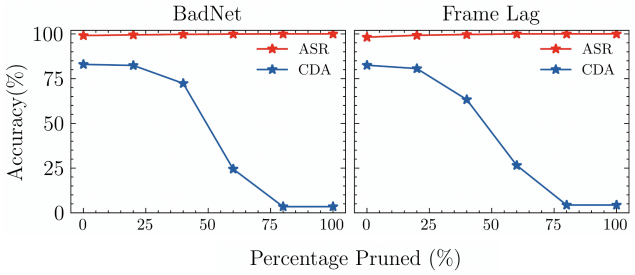

Figure 6. **Pruning Defense (Kinetics-Sounds).** Pruning is completely ineffective against image backdoor attacks extended to the video domain and natural video backdoor attacks. Even though the clean accuracy has dropped to random, the attack success rate is maintained at very high levels.

to reduce the dimension of the activations, after which the projected activations are clustered into two classes and compared to the activations of the clean set. STRIP blends clean samples with the samples of a possibly poisoned inspection set. The entropy of the predicted probabilities is then checked for any abnormalities. Unlike clean samples, poisoned samples tend to have a low entropy. Pruning suggests that the backdoor is usually embedded in particular neurons in the network that are only activated in the presence of the trigger. Therefore, those neurons are supposed to be dormant as far as the test set samples, *i.e*. clean samples, are concerned. This allows us to detect and prune those dormant neurons to eliminate the backdoor. Table 5 shows the elimination and sacrifice rates of AC when applied against some of the considered attacks. The elimination rate refers to the ratio of poisoned samples correctly detected as poisoned to the total number of poisoned samples, whereas the sacrifice rate refers to the ratio of clean samples incorrectly detected as poisoned to the total number of clean samples. Whereas AC has partial success in defending against image backdoor attacks, it fails completely against the proposed natural backdoor attacks. Figure 5 shows that the entropy of the clean and poisoned samples of the proposed natural attacks is very similar and therefore could evade the STRIP defense, while BadNet and FTrojan are detectable. Finally, Figure 6 shows that pruning the least active neurons causes a reduction in CDA without reducing ASR. This is observed not only for the natural attacks, but also for the extended image backdoor attacks, hinting that image backdoor defenses are not effective in the video domain.

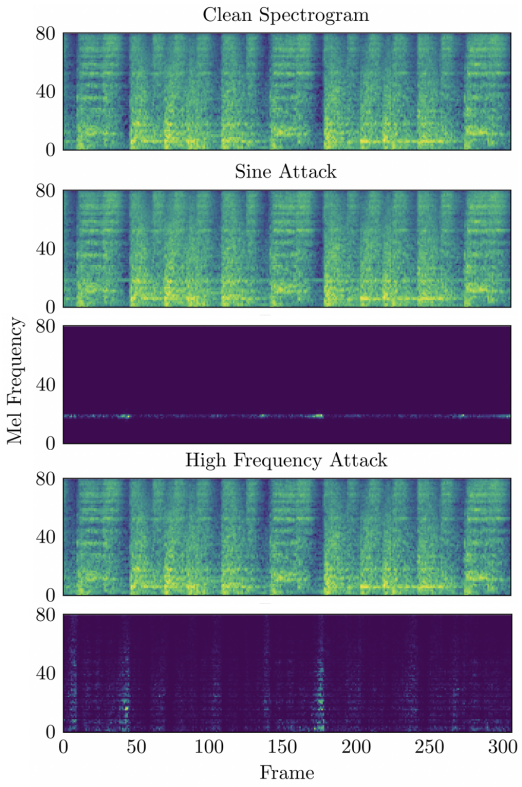

Figure 7. **Clean and Attacked Audio Spectrograms.** The utilized audio backdoor attacks are not only audibly imperceptible but also leave no perceptible artifacts in the Mel spectrogram. The spectrogram of each attack is followed by the absolute difference of the attacked spectrogram with the clean one.

	Baseline	Sine Attack	High Frequency Attack
CDA(%)	49.21	47.21	47.61
ASR(%)	-	96.36	95.96

Table 6. **Audio Backdoor Attacks (Kinetics-Sounds).** Both sine attack and the high-frequency band attack perform similarly to baseline in terms of CDA while being able to achieve high ASR.

	Late Fusion			Early Fusion		
	Clean Audio	**Sine Attack**	**High Freq. Attack**	**Clean Audio**	**Sine Attack**	**High Freq. Attack**
Clean Video	80.25 / -	81.74 / 70.98	80.96 / 77.91	84.72 / -	83.48 / 92.23	83.94 / 93.72
BadNet	77.33 / 66.97	78.63 / 99.74	77.33 / 99.87	87.50 / 99.29	85.10 / 99.87	85.75 / 100.00
Blend	79.60 / 75.06	80.76 / 99.68	79.08 / 99.61	86.08 / 98.19	83.55 / 99.81	85.43 / 99.87
SIG	78.50 / 68.33	80.12 / 99.87	79.02 / 100.00	86.92 / 99.81	84.97 / 100.00	85.95 / 100.00
WaNet	77.66 / 68.39	79.79 / 99.94	79.02 / 99.94	86.46 / 98.96	84.97 / 100.00	85.88 / 100.00
FTrojan	79.66 / 67.16	80.76 / 99.48	79.99 / 99.29	86.08 / 98.58	84.65 / 99.94	85.49 / 100.00
Frame Lag	79.08 / 63.41	80.57 / 99.74	79.47 / 99.87	86.08 / 98.19	84.59 / 99.94	84.65 / 100.00
Video Corruption	78.11 / 64.57	78.24 / 99.68	77.66 / 99.94	86.59 / 99.29	84.59 / 100.00	85.43 / 100.00
Motion Blur	79.79 / 69.24	80.70 / 99.68	79.86 / 99.94	86.40 / 98.58	84.65 / 100.00	85.62 / 100.00

Table 7. **Audiovisual Backdoor Attacks (Kinetics-Sounds).** The entries in the table report the CDA(%)/ASR(%) of attacking late and early fused audiovisual networks. When a single modality is attacked, late fusion has a low ASR compared to early fusion. When both modalities are attacked, the ASR of both late and early fusion are high.

4.3. Audio Backdoor Attacks

Attacks proposed against audio networks have been limited to adding a low-volume one-hot-spectrum noise in the frequency domain, which leaves highly visible artifacts in the spectrogram [86] or adding a human non-audible component [34], $f < 20$Hz or $f > 20$kHz, which is non-realistic, since spectrograms usually filter out those frequencies. We consider two attacks against the Kinetics-Sounds dataset; the first is to add a low-amplitude sine wave component with $f = 800$Hz to the audio signal, and the second is to add band-limited noise $5\text{kHz} < f < 6\text{kHz}$. The spectrograms and the absolute difference between the attacked spectrograms and the clean spectrogram are shown in Figure 7. Since no clear artifacts are observed in the spectrograms, human inspection fails to label the spectrograms as attacked. The CDA and ASR rates of the backdoor-attacked models for both attacks are shown in Table 6. The attacks achieve a relatively high ASR.

4.4. Audiovisual Backdoor Attacks

Now, we combine video and audio attacks to build a multi-modal audiovisual backdoor attack. The way we do it is by taking our attacked models from Sections 4.2 and 4.3 and applying early or late fusion. For early fusion, we extract video and audio features using our trained audio and video backbones, and we then train a classifier on the concatenation of the features. In late fusion, the video and audio networks predict independently on the input, and then the individual logits are aggregated to produce the final prediction. To answer the three questions posed in Section 3.3, we run experiments in which both modalities are attacked and others in which only a single modality is attacked for both early and late fusion setups (Table 7). We summarize the results as follows. **(1)** Attacking two modalities consistently improves ASR and even CDA in some cases. **(2)** Attacking a single modality is good enough to achieve a high ASR in the case of early fusion but not late fusion. **(3)** Early fusion enables the best of both worlds for the attacker, namely, a high CDA and an almost perfect ASR. On the other hand, late fusion experiences some serious drops in ASR in the unimodal attack setup. An interesting finding in these experiments is the following: if the outsourcer has the option to outsource the most expensive modality, training wise, while training other modalities in-house, applying late fusion could be used as a defense mechanism, especially in the presence of more clean modalities.

5. Conclusion

Backdoor attacks present a serious and exploitable vulnerability against both unimodal and multi-modal video action recognition models. We showed how existing image backdoor attacks could be extended either statically or dynamically to develop powerful backdoor attacks that achieve both a high clean data accuracy and a high attack success rate. Besides existing image backdoor attacks, there exists a set of natural video backdoor attacks, such as motion blur and frame lag, that are resilient to existing image backdoor defenses. Given that videos are usually accompanied by audio, we showed two ways in which one could attack audio classifiers in a human inaudible manner. The attacked video and audio models are then used to train an audiovisual action recognition model by applying both early and late fusion. Different combinations of poisoned modalities are tested, concluding that: **(1)** poisoning two modalities could achieve extremely high attack success rates in both late and early fusion settings, and **(2)** if a single modality is poisoned, unlike early fusion, late fusion could reduce the effectiveness of the backdoor.

6. Acknowledgements

This work was supported by SDAIA-KAUST Center of Excellence in Data Science, Artificial Intelligence.

References

[1] Hasan Abed Al Kader Hammoud, Adel Bibi, Philip H.S. Torr, and Bernard Ghanem. Don't freak out: A frequency-inspired approach to detecting backdoor poisoned samples in dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2338–2345, June 2023. 2

[2] Humam Alwassel, Dhruv Kumar Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *ArXiv*, abs/1911.12667, 2020. 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2, 4

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2

[6] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105, 2019. 1, 4

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 2

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2, 4

[9] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018. 2, 5, 6

[10] 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. 2, 4

[11] Xuan Chen, Yuena Ma, and Shiwei Lu. Use procedural noise to achieve backdoor attack. *IEEE Access*, 9:127204–127216, 2021. 2

[12] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 5

[13] 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. 2

[14] Khoa D Doan and Yingjie Lao. Backdoor attack with imperceptible input and latent modification. In *NeurIPS*, 2021. 2

[15] Khoa D Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11946–11956, 2021. 2

[16] Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of backdoor attacks with limited information and data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16482–16491, 2021. 2

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 2

[18] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 200–210, 2020. 2

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 4

[20] Yu Feng, Benteng Ma, Jing Zhang, Shanshan Zhao, Yong Xia, and Dacheng Tao. Fiba: Frequency-injection based backdoor attack in medical image analysis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20844–20853, 2022. 2

[21] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pages 113–125, 2019. 2, 5, 6

[22] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. The activitynet large-scale activity recognition challenge 2018 summary. *arXiv preprint arXiv:1808.03766*, 2018. 3

[23] 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

[24] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du, and Dawn Song. Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in ai systems. *arXiv preprint arXiv:1908.01763*, 2019. 2, 6

[25] Wenbo Guo, Lun Wang, Yan Xu, Xinyu Xing, Min Du, and Dawn Song. Towards inspecting and eliminating trojan backdoors in deep neural networks. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 162–171. IEEE, 2020. 2

[26] Hasan Abed Al Kader Hammoud and Bernard Ghanem. Check your other door! creating backdoor attacks in the frequency domain. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 1, 2, 4

[27] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: defending against backdoor attacks using robust statistics. *arXiv preprint arXiv:2104.11315*, 2021. 2

[28] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9240–9249, 2019. 2

[29] Di Hu, Zongge Wang, Haoyi Xiong, Dong Wang, Feiping Nie, and Dejing Dou. Curriculum audiovisual learning. *ArXiv*, abs/2001.09414, 2020. 2

[30] Xiaoling Hu, Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha, and Chao Chen. Trigger hunting with a topological prior for trojan detection. *arXiv preprint arXiv:2110.08335*, 2021. 2

[31] Mojan Javaheripi, Mohammad Samragh, Gregory Fields, Tara Javidi, and Farinaz Koushanfar. Cleann: Accelerated trojan shield for embedded neural networks. In *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pages 1–9. IEEE, 2020. 2

[32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[33] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2

[34] Stefanos Koffas, Jing Xu, Mauro Conti, and Stjepan Picek. Can you hear it?: Backdoor attacks via ultrasonic triggers. *Proceedings of the 2022 ACM Workshop on Wireless Security and Machine Learning*, 2022. 8

[35] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. Universal litmus patterns: Revealing backdoor attacks in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 301–310, 2020. 2

[36] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[37] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 4

[38] Yuezun Li, Y. Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16443–16452, 2021. 2

[39] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021. 2

[40] Yiming Li, Baoyuan Wu, Yong Jiang, Zhifeng Li, and Shutao Xia. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, PP, 2022. 1, 3

[41] Cong Liao, Haoti Zhong, Anna Cinzia Squicciarini, Sencun Zhu, and David J. Miller. Backdoor embedding in convolutional neural network models via invisible perturbation. *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020. 2

[42] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019. 2, 4

[43] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 273–294, 2018. 2, 5, 6

[44] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1265–1282, 2019. 2

[45] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. 2017. 2

[46] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *ECCV*, 2020. 2

[47] Yingqi Liu, Guangyu Shen, Guanhong Tao, Zhenting Wang, Shiqing Ma, and Xiangyu Zhang. Complex backdoor detection by symmetric feature differencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15003–15013, 2022. 2

[48] Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pages 45–48. IEEE, 2017. 2

[49] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 2

[50] Chenxu Luo and Alan L Yuille. Grouped spatial-temporal aggregation for efficient action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5512–5521, 2019. 2

[51] Hua Ma, Yinshan Li, Yansong Gao, Zhi Zhang, Alsharif Abuadbba, Anmin Fu, Said F. Al-Sarawi, Surya Nepal, and Derek Abbott. Macab: Model-agnostic clean-annotation backdoor to object detection with natural trigger in real-world. *ArXiv*, abs/2209.02339, 2022. 5

[52] Pedro Miguel Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12929–12940, 2021. 2

[53] A. Nguyen and A. Tran. Wanet - imperceptible warping-based backdoor attack. *ArXiv*, abs/2102.10369, 2021. 2, 4

[54] Xiangyu Qi, Ting Xie, Saeed Mahloujifar, and Prateek Mittal. Circumventing backdoor defenses that are based on latent separability. *ArXiv*, abs/2205.13613, 2022. 2

[55] Ximing Qiao, Yukun Yang, and Hai Li. Defending neural backdoors via generative distribution modeling. *Advances in neural information processing systems*, 32, 2019. 2

[56] Han Qiu, Yi Zeng, Shangwei Guo, Tianwei Zhang, Meikang Qiu, and Bhavani Thuraisingham. Deepsweep: An evaluation framework for mitigating dnn backdoor attacks using data augmentation. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 363–377, 2021. 2

[57] Yankun Ren, Longfei Li, and Jun Zhou. Simtrojan: Stealthy backdoor attack. *2021 IEEE International Conference on Image Processing (ICIP)*, pages 819–823, 2021. 2

[58] A. Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 703–718, 2022. 2

[59] Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. Backdoor scanning for deep neural networks through k-arm optimization. In *International Conference on Machine Learning*, pages 9525–9536. PMLR, 2021. 2

[60] Xiaoyu Song, Hong Chen, Qing Wang, Yunqiang Chen, Mengxiao Tian, and Hui Tang. A review of audio-visual fusion with machine learning. *Journal of Physics: Conference Series*, 1237, 2019. 2

[61] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 4

[62] Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of {DNNs} for robust backdoor contamination detection. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1541–1558, 2021. 2

[63] Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, Pan Li, and Xiangyu Zhang. Better trigger inversion optimization in backdoor scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13368–13378, 2022. 2

[64] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018. 2

[65] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[66] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5551–5560, 2019. 2

[67] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[68] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *ArXiv*, abs/1912.02771, 2019. 2

[69] 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, 6

[70] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021. 2

[71] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[72] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. Backdoor attack through frequency domain. *ArXiv*, abs/2111.10991, 2021. 2, 4

[73] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15054–15063, 2022. 2

[74] Emily Wenger, Roma Bhattacharjee, Arjun Nitin Bhagoji, Josephine Passananti, Emilio Andere, Haitao Zheng, and Ben Zhao. Natural backdoor datasets. *ArXiv*, abs/2206.10673, 2022. 5

[75] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34:16913–16925, 2021. 2

[76] Pengfei Xia, Hongjing Niu, Ziqiang Li, and Bin Li. Enhancing backdoor attacks with multi-level mmd regularization. *IEEE Transactions on Dependable and Secure Computing*, 2022. 2

[77] Zhen Xiang, David J Miller, and George Kesidis. Post-training detection of backdoor attacks for two-class and multi-attack scenarios. *arXiv preprint arXiv:2201.08474*, 2022. 2

[78] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[79] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 2

[80] Mingfu Xue, Can He, Shichang Sun, Jian Wang, and Weiqiang Liu. Robust backdoor attacks against deep neural networks in real physical world. *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 620–626, 2021. 5

[81] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019. 2

[82] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *NeurIPS*, 2019. 2

[83] Chang Yue, Peizhuo Lv, Ruigang Liang, and Kai Chen. Invisible backdoor attacks using data poisoning in the frequency domain. *ArXiv*, abs/2207.04209, 2022. 2

[84] Yi Zeng, Si Chen, Won Park, Z Morley Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. *arXiv preprint arXiv:2110.03735*, 2021. 2, 6

[85] Yi Zeng, Won Park, Zhuoqing Morley Mao, and R. Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16453–16461, 2021. 2

[86] Tongqing Zhai, Yiming Li, Zi-Mou Zhang, Baoyuan Wu, Yong Jiang, and Shutao Xia. Backdoor attack against speaker verification. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2560–2564, 2021. 8

[87] Feng Zhao, Li Zhou, Qi Zhong, Rushi Lan, and Leo Yu Zhang. Natural backdoor attacks on deep neural networks via raindrops. *Security and Communication Networks*, 2022. 5

[88] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14443–14452, 2020. 1, 4, 5

[89] Zhendong Zhao, Xiaojun Chen, Yu Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. Defeat: Deep hidden feature backdoor attacks by imperceptible perturbation and latent representation constraints. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15192–15201, 2022. 2

[90] Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Data-free backdoor removal based on channel lipschitzness. In *European Conference on Computer Vision*, pages 175–191. Springer, 2022. 2

[91] Songzhu Zheng, Yikai Zhang, Hubert Wagner, Mayank Goswami, and Chao Chen. Topological detection of trojaned neural networks. *Advances in Neural Information Processing Systems*, 34:17258–17272, 2021. 2

[92] Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. Imperceptible backdoor attack: From input space to feature representation. In *IJCAI*, 2022. 2