REVIEW ARTICLE

A Survey on Real-Time Adversarial Attack Detection and Robustness for Real-Time Systems

Sharanya Peri[1,*], **Navya Vadla**[1], **Yeshwanth Panuganti**[1] **and Kadiyala Ramana**[1]

[1] Department of Artificial Intelligence and Data Science, Chaitanya Bharathi Institute of Technology, Hyderabad, India

Abstract

The use of deep neural networks in modern surveillance systems enables real-time object detection, facial recognition, and anomaly detection, but they remain vulnerable to adversarial attacks, creating critical security risks. This survey reviews detection methods tailored for real-time surveillance, categorizing domain-specific attacks including gradient-based methods (FGSM, PGD, C&W), physical patches, and temporal attacks on video data. We evaluate detection approaches across six categories: feature-based (LID, frequency analysis), reconstruction-based (autoencoders, GANs), auxiliary model-based, uncertainty-based (Bayesian Networks, MIAD), steganalysis-based, and attention-based (ViTGuard, SHAP). Timeliness was a key focus—LSTM-AD achieved detection within 0.001 seconds at approximately 90% accuracy, while NutNet increased inference time by only 8% for patch detection. Key limitations include poor detection of novel attacks, computational burden on edge devices, and the accuracy–clean detection trade-off. Underexplored areas include video streaming attacks (relative to still images), weak integration with alerting systems, and vulnerability to adaptive attacks. Future work should focus on unified threat detection, online learning for evolving threats, and certified robustness for operational deployment.

Keywords: adversarial attacks, deep learning security, surveillance systems, real-time detection, object detection, video anomaly detection, adversarial defence.

1 Introduction

Deep Learning has become increasingly used for applications in various areas of life such video surveillance using face recognition and computer vision [1]. In the area of security, there has been a significant rise in the number of automated anomaly detection systems that can detect burglaries and other crimes using these technologies [2]. Deep Neural Networks can also be applied to Object Detection and Image Classification for Applications in Autonomous Driving [3] and Control of Unmanned Aerial Vehicles (UAVs) [4].

With all the successful cases of how Deep Learning has been utilized, they still have vulnerabilities to attacks, such as Adversarial Attacks [5, 6]. When an Image/Frame is altered by adding Adversarial Examples, there can be misclassifications by the

Submitted: 24 February 2026
Accepted: 28 April 2026
Published: 09 May 2026

Vol. 2, **No.** 2, 2026.
10.62762/JIAP.2026.481078

***Corresponding author:**
✉Sharanya Peri
sharanyaperi20@gmail.com

Citation
Peri, S., Vadla, N., Panuganti, Y., & Ramana, K. (2026). A Survey on Real-Time Adversarial Attack Detection and Robustness for Real-Time Systems. *ICCK Journal of Image Analysis and Processing*, 2(2), 104–120.

Model run on that Image/Frame by resulting in incorrect identifications [5, 7]. Adversarial examples are input samples manipulated via small, often imperceptible perturbations to fool ML models [1, 5]. While these perturbations cannot be detected by the human eye, deep learning models are sensitive to these minor changes, thereby causing the model to make an incorrect prediction with high confidence [1, 5, 7], as illustrated in Figure 1. This threat is non-trivial; in safety-critical systems like autonomous vehicles, misclassifying a stop sign as a speed limit sign can have severe, life-endangering consequences [3]. Adversarial attacks are a security threat to several real-world systems, limiting their real-world application [5, 8].

An attack could also be launched using adversarial patches, such as attaching stickers that will confuse the model [9–11]. In surveillance, the attack could take place either through an Appearing Attack, whereby the attacker will trick the detector by confusing it with regard to the objects' identities, or a Hiding Attack, where there will be a security breach and any crime remains unnoticed [12]. Face recognition models are not safe from attacks, where they have been used to launch security breaches [13].

To tackle the above problems, this survey will discuss all pertinent techniques in terms of detection and protection against such attacks. Our emphasis here will be on finding the necessity for real-time usability and suitable methods for video streaming data because these are the more important needs usually neglected in studies limited to only images. The rest of the paper is organized as follows: Section 2 includes the literature survey; Section 3 explains the taxonomy of attack types; Section 4 deals with defense and detection mechanisms; Section 5 outlines real-time detection issues concerning surveillance; Section 6 explains the evaluation methodology; and Section 7 is the analysis and research direction section.

2 Literature Survey

This section reviews the key works in the field of adversarial attack detection and robustness, synthesising their contributions, limitations, and reported inference performance.

2.1 Benchmark and Survey Works

Several foundational survey and benchmark contributions establish the broader context of adversarial robustness research. The *BlackboxBench* framework [5] provides a flexible benchmark that consolidates 59 varieties of black-box attacks—29 query-based and 30 transfer-based—into a unified evaluation platform, enabling fair cross-method comparisons on datasets such as CIFAR-10 and ImageNet. While valuable for benchmarking, its focus on image classification with norm-bounded noise means it does not address adversarial detection in video surveillance contexts. Similarly, the survey *Adversarial Attacks & Defenses in Computer Vision* [5] reviews advances from 2018 onwards across more than 3,000 publications, covering certified defenses, detection-based defenses, and complex computer vision tasks such as object detection and tracking, though the sheer volume of the field makes complete coverage infeasible. The *Systematic Review of Adversarial Machine Learning Attacks, Defense Mechanisms, and Technologies* [14] catalogues attack types and real-world illustrations while recommending open-source toolkits such as ART and CounterFit, but stops short of practical implementation and testing, and notes that formal verification—though theoretically rigorous—suffers from severe scalability constraints.

The survey on *Autonomous Driving Systems Based on Deep Learning* [3] provides an extensive review of physical, cyber, and adversarial attacks against camera-dependent driving systems, stressing the importance of monitoring systems as a last line of defense; it explicitly flags real-time necessity while acknowledging that adversarial training is resource-intensive and may degrade performance against novel attacks.

2.2 Attack-Focused Studies

A number of papers concentrate primarily on constructing or characterising adversarial attacks rather than on detection. The work on *Physical Adversarial Attacks Using Meta-GAN* [9] demonstrates how Meta-GAN can produce physically robust, generalisable adversarial perturbations of high perceptual quality that transfer across model architectures and classes; ResNet-based models show greater resilience than VGG-based ones, although no real-time detection or defense mechanism is proposed. The *Gradient Sign-Based Adversarial Attack Algorithm by Differential Evolution* [16] introduces a black-box attack that approximates gradient signs via differential evolution without requiring any knowledge of the target model, but its focus on evasion strategy leaves detection unaddressed. The paper on *Universal Adversarial Patch Attacks to Object*

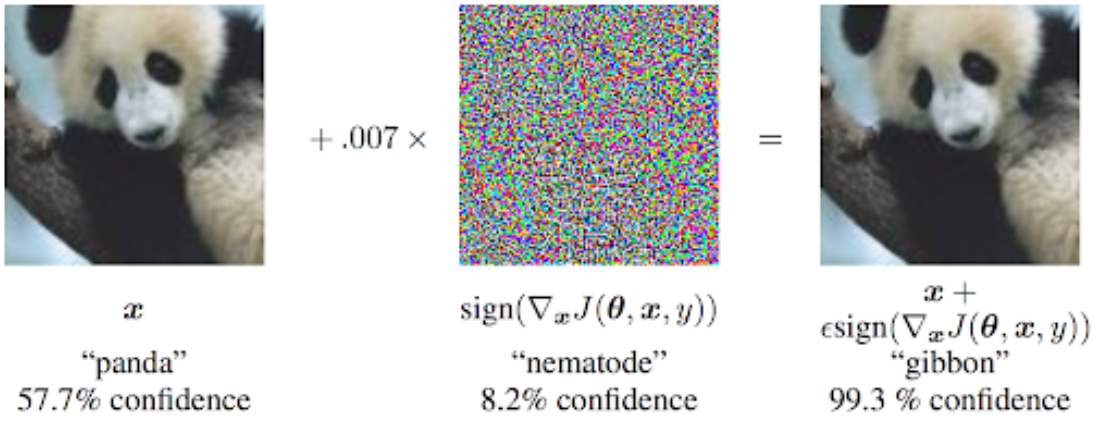

Figure 1. Illustration of an adversarial perturbation: a panda image classified at 57.7% confidence is misclassified as a gibbon at 99.3% confidence after adding an imperceptible FGSM perturbation [1].

Detection: Revisiting [17] reexamines suppression- and misclassification-style patch attacks on object detectors, yet provides no information on countermeasures or attack inference timing.

Attacks on Video Anomaly Detection Using Adversarial Machine Learning [2] proposes novel attack methods—including freeze attacks and frame-skipping attacks—that can be implemented via real-world cyberattacks such as Wi-Fi deauthentication, reducing the confidence of video anomaly detection systems or concealing physical anomalies entirely. The study demonstrates the damage potential of such attacks through simulation but deliberately defers the design of countermeasures. *IBAttack* [6] presents a stealthy backdoor attack based on adaptive triggers that survives fine-tuning and transfers to unseen classifiers, though its scope is restricted to data poisoning rather than real-time evasion. *SingleADV* [5] describes a target-specific attack that preserves interpretability while fooling a single class, with defense techniques evaluated only on general image classification models.

2.3 Adversarial Training and Robustness Enhancement

Several works focus on hardening models through improved training strategies. The *Hardness-Guided Attack Strategy* (*HGSD-AT*) [18] progressively increases attack difficulty during training using perturbation magnitude and loss values as difficulty indicators, improving standard accuracy; however, evaluation is limited to image classification, and over-reliance on robust features may reduce efficacy in real-world deployment. The *Hybrid Adversarial Attack Defense* [19] combines FGSM and PGD perturbations into a hybrid adversarial training framework evaluated on MNIST, CIFAR-10, and CelebA, though the additional parameters introduced complicate application to other architectures. The work on *Topology Aligning Adversarial Training* [20] improves robustness by emphasising inter-class relationships through topology alignment and knowledge-guided training, though specific limitations are not discussed by the authors. Research on *Neuron Sensitivity* links adversarial robustness to the stability of sensitive neurons and proposes training modifications to mitigate this sensitivity, but does not address real-time evasion scenarios.

Adversarial Training with Booster Signal Injection [18] demonstrates that input-preprocessing and randomisation defenses are vulnerable in white-box settings, and proposes signal injection to strengthen adversarial training; inference was measured at 25.34 ms on an NVIDIA A6000 GPU. The study on *Adversarial Robustness of Randomized Neural Networks Using GradDiv Regularization* [1] introduces gradient diversity regularisation to prevent gradient alignment in randomised networks, overcoming a known weakness in prior randomness-based approaches that were exploitable via expected optimal transport attacks.

2.4 Preprocessing and Purification Defenses

A distinct line of work seeks to neutralise adversarial perturbations before they reach the classifier. *APR-Net* [21] proposes a universal dual-module architecture comprising a denoising module and an image restoration module that removes adversarial perturbations without modifying the classifier structure; it achieves 0.21 s per image with 8.90M parameters and 265.6M MACs on ImageNet, though preprocessing can degrade feature extraction quality.

FM-Defense (*Semantic Feature Manipulation*) [15] employs a Combo-VAE to classify and purify adversarial samples, achieving accurate sample reconstruction; however, adversarial noise cannot transfer to other samples, and high-fidelity VAE generation remains challenging on large datasets. *Defense Patterns for Video Recognition* [22] introduces universally optimised perturbations appended to video frames that transfer across time steps to boost model robustness with minimal retraining overhead, although clean-frame accuracy may suffer and individualised patterns could outperform universal ones in some scenarios. The *Transformer-Based Defense-GAN for Palm-Vein Recognition* [13] uses a local transformer GAN to purify perturbations in palm-vein biometric images at 0.02 s per image, but the domain specificity limits generalisation to surveillance footage.

The A^3D *Platform* [5] integrates neural architecture search with adversarial attack optimisation, achieving practical results on CIFAR and ImageNet with runtimes between 3.05 and 657.9 s and model sizes from 2.30M to 11.17M parameters; validation on video or surveillance data is absent. *AEDPL-DL* (*Protection Layer Framework*) [14] inserts a conventional machine learning protection layer—using random forest and XGBoost classifiers—ahead of the DNN classifier to detect and exclude adversarial examples under four evasion attack types, with performance degrading as the perturbation magnitude epsilon decreases.

2.5 Detection-Focused Methods

A large body of work is dedicated specifically to detecting adversarial examples. *On Detecting Adversarial Perturbations* [23] proposes training a detector sub-network alongside the main classifier via a binary classification task; the detector generalises well to weaker adversaries but is susceptible to dynamic adversaries that simultaneously fool both the classifier and the detector. *Feature Squeezing* [24] detects adversarial inputs by comparing model predictions on original and transformed versions of an image (using bit-depth reduction and smoothing); detection is computationally cheap and requires no additional training, with accuracy reaching above 98% on MNIST and around 85% on CIFAR-10 and ImageNet, though robustness against adaptive attackers remains uncertain.

Local Intrinsic Dimensionality (*LID*) [25] characterises adversarial subspaces by observing that adversarial samples exhibit significantly higher LID scores than clean or corrupted inputs, outperforming kernel density and Bayesian uncertainty metrics in accuracy; its reliability degrades with small minibatch sizes, and adaptive attackers can circumvent LID-based detection in complex settings. *Detecting Adversarial Samples from Artifacts* [26] explores kernel density estimation and Bayesian neural network uncertainty as detection signals based on artefacts in intermediate model representations, providing an early foundation for uncertainty-based detection.

The *Mutual Information-based Adversarial Detection* (*MIAD*) [27] method uses a twin autoencoder with mutual information estimation and distribution matching to generate discriminative features in a model-agnostic manner, removing the dependency on target model details that other techniques such as LID require. *Pseudorandom Classification* (*PRC*) [28] draws on cryptographic principles, injecting a keyed pseudorandom function to disrupt backpropagation-based attacks and achieve attack-agnostic detection; training is more time-consuming than standard classifiers due to the additional input operations involved. *Bayesian Adversarial Example Detector* (*BATER*) [29] uses a Bayesian neural network to detect adversarial examples by analysing distributional discrepancies in hidden layer outputs, but requires multiple forward passes, imposing computational overhead at test time. The theoretical analysis of *Bayesian Neural Network Robustness* [30] establishes a positive correlation between Bayesian accuracy and adversarial robustness, though the theoretical guarantees apply to infinite-size networks and accurate Bayesian inference in non-linear models remains computationally hard.

The *Fusion-Based Feature Selection* [31] method for second-round attack detection fuses handcrafted and deep features under an expansion-compression selection framework, achieving 0.42–0.50 s per image on an RTX 2080 Ti; steganalysis residual map (SRM) features perform poorly on sparse adversaries. The *Steganalysis-Based Detection* [32] approach exploits the structural similarity between adversarial perturbations and accidental steganography, using estimated modification probability maps for detection; non-neural-network detectors tend to perform strongly, though more sophisticated adaptive techniques remain a challenge. *Local Histogram Equalization* (*LHE*) [7] amplifies high-frequency perturbation characteristics as a non-network preprocessing step, improving detection without altering the base classifier, though cross-model detection under I-FGSM reduces accuracy.

ViTGuard [33] is tailored for Vision Transformers (ViTs) and uses a Masked AutoEncoder (MAE) to reconstruct inputs and compare attention maps and CLS-token representations for feature divergence detection, handling both ℓ_p-norm and patch attacks in just 2.4 ms (MAE) plus 0.2 ms (ViT-16 inference); the choice of optimal detection layer must be determined empirically per attack type. The *LSTM-AD and CNN-AD* models proposed for drone guidance [4] use SHAP explainability values to detect adversarial attacks on deep reinforcement learning controllers, with LSTM-AD achieved approximately 90% detection accuracy with a latency of only 0.001 s on the UAV guidance system, making it highly suitable for real-time deployment; CNN-AD requires 0.162 s.

2.6 Real-Time and Patch-Specific Defenses

Among the surveyed works, *NutNet* [12] stands out for its explicit real-time design. It is a lightweight autoencoder trained exclusively on clean samples that treats adversarial patches as out-of-distribution inputs, providing generalised defense against both Hiding Attacks (HA) and Appearing Attacks (AA) for object detectors at 34.9 FPS with only 8.6% additional inference overhead (5k-parameter autoencoder). Optional inpainting via PICnet can restore detection to clean-image precision but at substantial cost and a reduction in detection rate of over 50%. NutNet requires careful tuning of reconstruction thresholds and exhibits reduced patch detection accuracy in low-light environments. *SecureVQA* [34] defends video quality assessment models using random spatial grid sampling and pixel-level randomness (intra-frame) combined with temporal inter-frame features, achieving 0.369 s per frame in full mode and 0.00045 s in intra-frame-only mode on an RTX 4090; intra-frame defense degrades VQA accuracy and the inter-frame approach may be ineffective against future motion-based temporal attacks.

2.7 Domain-Specific and Specialist Works

Several studies address adversarial robustness in specialised domains that inform, but do not directly map to, general-purpose video surveillance. Research on *Salient Object Detection in Optical Remote Sensing* [35] demonstrates the strength of Adversarial Fallback Defense (AFD) against PGD attacks in remote sensing imagery, though the domain differs substantially from ground-level video and no latency requirements are discussed. The *S^3ANet* architecture [1] combines pyramid spatial attention and a global spectral transformer for hyperspectral image classification defense, with runtimes of 234–578 s on an RTX 3090—unsuitable for real-time surveillance. The *Double-Level Attack* (*DLA*) framework for air transport signal recognition uses dynamic step sizes and feature-level and decision-level losses to generate adversarial examples and enhance model robustness, but the application domain of jamming signal recognition differs significantly from CCTV footage. Studies on *adversarial attacks in infrared deep learning surveillance* [14] confirm the effectiveness of adversarial training against mixed attacks on infrared CNNs, but scalability and real-time applicability for sophisticated surveillance devices are not discussed. Adversarial attacks and defenses for *deep ranking and metric learning* propose candidate and query attacks alongside an anti-collapse triplet defense and an Empirical Robustness Score (ERS); the focus on image retrieval and face recognition limits direct applicability to real-time detection pipelines.

3 Taxonomy of Adversarial Attacks

Attack categories are fundamentally defined by the attacker's knowledge (White-box, Black-box, Grey-box [1, 5]), the transferability of the attack across models [5], and the desired outcome (Targeted vs. Untargeted [5, 39]). For surveillance, attacks are primarily categorised by their nature and method. Figure 2 presents a hierarchical taxonomy of the attack categories discussed in this section.

3.1 Attack Categories

Attacks against surveillance systems primarily fall into two categories. The first category is evasion attacks, which occur during the deployment phase. In such attacks, the adversary crafts a perturbation and adds it to a clean input to manipulate the model's predictions at the time of inference [5, 6]. The second category is poisoning attacks, wherein malicious data, such as triggers or misleading labels, are injected into the training dataset, leading the model to produce incorrect predictions on the ground-truth training set [6].

Attacks can also be classified according to their objectives. Targeted attacks aim to craft an input that forces the model to predict a specific class chosen by the adversary. This type of attack is generally more difficult to generate [5, 39]. In contrast, untargeted attacks seek to degrade the model's performance by causing the input to be classified as any class other than its original ground-truth or the intended target class [1, 7]. Untargeted attacks typically require less

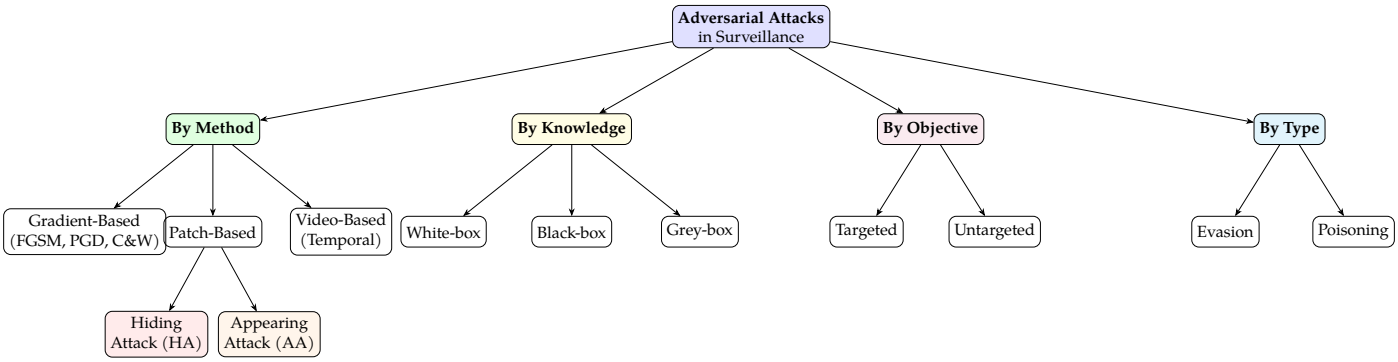

Figure 2. Taxonomy of adversarial attacks in surveillance systems, categorised by attack method, attacker knowledge, objective, and type.

perturbation and achieve successful attempts more easily compared to targeted attacks [5].

3.2 Attack Methods Relevant to Surveillance

Gradient-Based Attacks Gradient-based attacks represent a fundamental class of white-box methods that exploit the gradients of a target model to increase the loss associated with the model's predictions [1, 38]. Among these, the Fast Gradient Sign Method (FGSM) is a simple, one-step approach that rapidly modifies an input image by adding a distortion aligned with the gradient of the model at that input. FGSM can be extended into iterative multi-step methods, such as the Projected Gradient Descent (PGD) or DeepFool, which employ more advanced optimization techniques to generate adversarial inputs with minimal distortion [1, 18, 38]. Another prominent approach is the Carlini & Wagner (C&W) method, widely regarded as one of the most effective white-box attack techniques. C&W attacks formulate the adversarial generation as an optimization problem that seeks to produce the optimal distortion while preserving minimal changes to the original image, using a loss function defined with ℓ_0, ℓ_2, or ℓ_∞ distortion metrics.

Patch-Based Attacks Adversarial patches pose a serious risk to surveillance systems due to their high effectiveness and feasibility of physical implementation [10–12, 17]. A patch attack is accomplished by modifying pixel values within a specific, localized area of an image. Consequently, such patches can be physically produced as stickers and placed in the real world, enabling adversarial perturbations in physical environments [9–11]. In the context of object detection and surveillance, patches are commonly employed to carry out two types of attacks [11, 12, 17]. The first is the hiding attack (HA), in which the adversarial patch covers the target object to minimize the objectness score predicted by the detector. This causes the region proposal network to eliminate the proposal, effectively hiding the object from the detector [11, 12]. For instance, a person may use such an adversarial patch on their clothing to evade detection by a surveillance system. The second is the appearing attack (AA), where the patch is placed in the scene to increase the classification score of a target object, thereby tricking the detector into recognizing an object that is not actually present [11, 12, 17]. Repeated AA attacks can lead to operator fatigue and reduced trust in the surveillance system.

Attacks Related to Video Attacks targeting video recognition models can be categorized into those that exploit the spatial and temporal characteristics of video data [2, 22]. A straightforward cyber attack against a video anomaly detection (VAD) system is the Wi-Fi deauthentication attack, which disrupts the video feed by slowing it down, freezing frames, or reducing video quality. Such conditions create opportunities to conceal physical anomalies or trigger false alarms in the VAD system [2]. Beyond simple cyber attacks, more sophisticated methods take advantage of the temporal characteristics of video analysis models. These approaches sparsely insert perturbations based on the temporal propagation properties of sequence-based models, thereby compromising the model's performance while minimizing perceptible alterations to the video stream [2, 22].

3.3 Attack Types Specific to a Domain

There are major threats to Autonomous Driving Systems (ADS), including physical attacks on traffic signals using adversarial patches/stickers [3, 9, 10], as well as Poisoning/Backdoor Attacks through injecting malicious data during training that could result

in undesirable actions being taken when making inference from trained models [3].

4 Defense and Detection Mechanisms

4.1 Defence Paradigms

Defence paradigms focus on strengthening model robustness against adversarial perturbations by improving either the model training process or the input data preprocessing stage [1, 5].

Adversarial Training Adversarial training is a proactive heuristic defense method in which models are trained using adversarially perturbed data as part of their training sets, aiming to achieve better generalization and add an additional level of regularization [6, 18, 19]. However, this approach does not provide certified protection against attacks and may only defend against specific attack types [5, 18]. Training models using adversarial examples generated by Projected Gradient Descent (PGD) has been shown to produce robust models [18, 19]. Nevertheless, adversarial training incurs substantial computational costs associated with generating, labeling, and training on adversarial examples [3, 6, 18, 19].

Several advanced adversarial training strategies have been proposed to enhance robustness. Reactive strategies involve training a secondary, simultaneous model specifically designed to detect adversarial examples before they reach the primary classifier [23, 41]. The Hardness-Guided Strategy for Adaptive Training (HGSD-AT) is a training method that adapts based on the strength of attack samples, using metrics such as the perturbation magnitude required to fool the model or loss values in critical regions. As training progresses from weak to strong adversarial examples, this method increases the difficulty for adversaries while simultaneously preserving higher standard accuracy [18]. Topology alignment is another approach that enhances robustness by emphasizing the interrelations between classes rather than focusing on individual data points [20]. Knowledge-guided training further improves this method by employing a pre-existing robust model to direct a target (student) model in aligning its output with how the original model classifies input data [20].

Input Preprocessing Techniques Input preprocessing techniques defend against adversarial examples by applying transformations, disruptions, or alterations to input data before they reach the classifier, thereby diminishing the influence of adversarial perturbations. Feature squeezing removes unnecessary features from the input using squeezing methods, thereby reducing the overall feature space while preserving the integrity of the underlying model. When comparing model predictions between the original input and the corresponding squeezed input, any discrepancy is flagged as a potential adversarial example. Noise or perturbation removal methods denoise or purify the data before it enters the Deep Neural Network (DNN). By removing perturbations from the input data and restoring images to their original state, the impact of the attack is effectively reduced. The Universal Perturbation Removal Network (APR-Net) is a model-agnostic defense against adversarial examples that employs a dual-module architecture consisting of a denoising module and an image restoration module. Transformation techniques (TT) encompass a variety of input transformations, including scaling, noise addition, JPEG compression, cropping, and rotation. Defensive distillation is a robust training method that uses the output of soft labels (representing confidence levels) from an initial model to assist in training another model. This method incorporates a temperature parameter into the Softmax layer to effectively mask the gradients of both models. Finally, adversarial example detection methods are implemented to assist in identifying adversarial examples before classification is performed by the model [23, 24, 41].

4.2 Detection-Based Approaches

Detection methods find differences between normal examples and those that are considered to be adversarial through means such as anomaly detection based on the statistical characteristics and inconsistency in the output [41].

4.2.1 Detection Based on Features Used as a Basis of Detection

This category encompasses a set of methods that leverage distinctive features to distinguish adversarial examples (AEs) from clean inputs. These approaches detect AEs at various levels, including the input layer, intermediate stages of a layer, or across different layers of the neural network.

One representative technique in this category is local intrinsic dimensionality (LID), which exploits an intrinsic property of adversarial examples. Empirical studies have demonstrated that AEs exhibit significantly higher LID values compared to clean data. Consequently, LID can serve as a basis for

training an auxiliary detector to classify whether an input is adversarial or benign.

Another group of feature-based detection techniques involves analyzing the differences in how convolutional neural network (CNN) neurons represent inputs when comparing AEs against clean images. A notable implementation of this approach is SafetyNet, which evaluates how ReLU activation neurons in the last layer of a CNN produce two distinct types of distributions. By training a classifier, such as a radial basis function support vector machine (RBF SVM), on these two neuronal response patterns, SafetyNet can effectively discriminate between adversarial and legitimate inputs.

The third class of feature-based detection methods operates in the frequency domain to identify AEs. By transforming both watermarked and original images into the frequency domain using techniques such as Fourier or wavelet transformations, it becomes possible to distinguish between AEs and clean examples. This differentiation is feasible because certain types of adversarial attacks significantly manipulate lower frequency coefficients to enhance their transferability across models, thereby leaving detectable artifacts in the frequency representation.

4.2.2 *Reconstruction-Based Detection*

Reconstruction-based detection techniques aim to either purify or identify adversarial examples by mapping them back onto the manifold of clean data. These methods employ generatively trained models that are trained exclusively on benign or clean data inputs, enabling them to detect perturbations by measuring reconstruction fidelity [15, 41].

Autoencoder and VAE-Based Methods Autoencoders trained solely on clean datasets produce significantly larger reconstruction errors when presented with perturbed samples. This phenomenon occurs because the encoder struggles to map an adversarially perturbed input back to its original representation, thereby indicating the presence of an adversarial example in the data stream [15, 41]. Models such as MagNet utilize autoencoders to approximate the clean representation of original data and assist in identifying potential adversarial samples [41].

Generative Models for Purification Generative models, such as Defense-GAN, attempt to purify an example once it has been detected as adversarial [8, 15]. This purification process is accomplished by training the defense mechanism solely on clean data or samples drawn from the original data distribution. After an autoencoder identifies an example as adversarial, the Generative Adversarial Network (GAN) attempts to map the perturbed sample back onto the learned clean data manifold, seeking a clean image that is as close as possible to the adversarial input while preserving semantic content [8, 13, 15].

Universal Perturbation Removal Network The Universal Perturbation Removal Network (APR-Net), previously discussed in Section 4.1, employs a dual denoising and restoration mechanism to remove perturbations from input samples. This purification mechanism operates independently of the specific attack type, making it a model-agnostic defense against universal adversarial perturbations [21].

4.2.3 *Auxiliary Model Detection*

Auxiliary model detection approaches determine whether an input is adversarial by developing separate models that perform classification assessment on each input independently [14, 23, 41]. These methods introduce additional model components that work alongside the primary classifier to detect adversarial examples before they reach the final decision stage.

One common implementation involves training sub-networks for detection that learn to distinguish between clean and harmful inputs by leveraging intermediate representation features extracted from the main classifier's output [23, 32]. These sub-networks effectively serve as binary classifiers that flag potentially adversarial inputs based on internal representations that may reveal subtle perturbations invisible to the final output layer.

Another approach employs a dual-classifier framework consisting of two detection models: one operating on the original input and another on a transformed version of the same input, such as a feature-squeezed representation [24]. When the prediction produced by the original input differs from that of its transformed counterpart beyond a specified threshold, the input is classified as adversarial. This discrepancy arises because adversarial perturbations are often sensitive to input transformations, whereas clean inputs tend to produce consistent predictions across such transformations [24].

Feature fusion methods offer a more sophisticated detection strategy by combining multiple sources of information, including texture features and deep features extracted from Deep Neural Networks

(DNNs) [31, 32]. The objective of this approach is to identify small distortions that may not be captured by deep feature extraction alone. By fusing complementary feature representations, feature fusion techniques can detect subtle adversarial manipulations that might otherwise evade detection, thereby providing resilience against adaptive second-round attacks [31].

4.2.4 *Uncertainty and Probabilistic Methods*

Uncertainty and probabilistic methods distinguish adversarial samples from conventional inputs by analyzing model prediction variance or model confidence levels. Rather than relying solely on deterministic outputs, these approaches leverage the inherent uncertainty of model predictions to detect subtle perturbations indicative of adversarial manipulation [41].

Bayesian Neural Networks (BNNs), also referred to as Bayesian Adversarial Example Detectors (BATNs), employ randomness to generate output distributions from hidden layers. This probabilistic framework improves adversarial example detection compared to traditional Deep Neural Networks (DNNs), particularly when exposed to specific attacks such as Fast Gradient Sign Method (FGSM) or Carlini & Wagner (C&W) attacks. The inherent uncertainty estimates provided by BNNs offer a natural mechanism for identifying out-of-distribution inputs, including adversarial examples [29, 30].

Mutual Information-based Adversarial Detection (MIAD) is a model-agnostic method that utilizes a dual autoencoder architecture with self-supervision to develop a mapping from the image space into a latent space [27]. MIAD detects adversarial examples by comparing its output projections into latent space with the output projections into hidden spaces of similar models. Discrepancies between these projections signal the presence of adversarial perturbations [27].

Pseudorandom Classification (PRC) employs pseudorandom injection mechanisms to disrupt backpropagation-based attacks. By introducing controlled randomness into the classification process, PRC develops probabilistic adversarial example detection methods that are independent of the specific nature or type of attack. This approach provides robust detection across diverse attack strategies without requiring prior knowledge of the attack methodology [28].

Enhancing Image Differences Using Local Histogram Equalization (LHE) serves as a non-network preprocessing technique that amplifies and enhances the high-frequency characteristics caused by adversarial perturbations. By applying local histogram equalization, subtle frequency-domain artifacts introduced by perturbations become more pronounced, thereby improving detector performance without altering the underlying performance characteristics of the base classification model [7].

4.2.5 *Attention and Explainability-Based Methods*

Attention and explainability-based approaches leverage interpretability aspects of the underlying model's behavior to identify anomalies introduced by adversarial examples. These methods exploit the internal mechanisms of models, such as attention weights or feature importance scores, to detect discrepancies between clean and perturbed inputs.

ViTGuard ViTGuard is specifically designed to detect adversarial attacks on Vision Transformers (ViTs) and can identify both ℓ_p norm-based adversarial attacks and patch-based attacks [33]. The method relies on Masked Autoencoders (MAEs) to detect the presence of adversarial manipulations by comparing the outputs of the original input against those of a reconstructed input. This comparison is performed based on attention maps and CLS-token representations, which capture the model's focus and global image understanding, respectively. Discrepancies between these representations indicate potential adversarial perturbations [33]. As shown in Figure 3, the ViT architecture processes input through patch embedding, a transformer encoder with multi-head attention, and an MLP classification head.

SHAP for Adversarial Detection Shapley Additive Explanations (SHAP) is a method grounded in game-theoretic principles that provides a mathematically sound framework for explaining predictions made by machine learning models [4]. SHAP proves effective for identifying adversarial perturbations because such perturbations often manipulate features that differ substantially from those that are important for correctly classifying benign examples. Adversarial attacks typically introduce changes in features that are irrelevant or less relevant to the model's natural decision boundary, leading to explainability patterns that deviate from those of clean inputs. Consequently, a Convolutional Neural Network (CNN)-based classifier can be trained on SHAP values derived from model predictions,

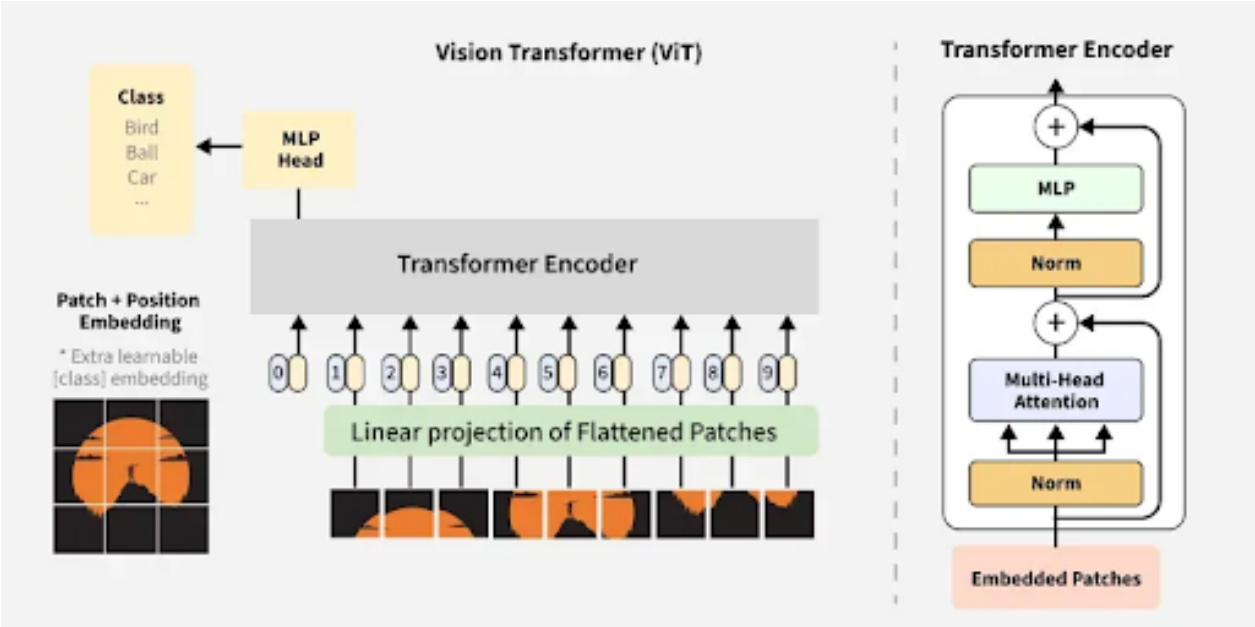

Figure 3. Architecture of the Vision Transformer (ViT) model, showing patch embedding, the transformer encoder with multi-head attention, and the MLP classification head [33].

enabling it to distinguish adversarial examples from legitimate ones based on their explanation signatures [4].

5 Real-Time Detection for Surveillance Systems

There are several important limitations that arise when developing autonomously-operated sensing systems with adversarial resilience. First, latency thresholds are essential to enable real-time use of the system [3]. Systems employing adversarial methods requiring large amounts of computational resources will not be feasible because an Autonomous Driving System (ADS) with significant computational requirements could exceed highly regulated timing limits and therefore delay the detection of an event [3]. Second, the defence method to be employed must support systems with limited processing, memory, and network resources [3]. Using methods that have no additional models or layers is easier to implement in resource-constrained peripheral systems than developing models and/or layers that require extensive processing capability [3, 12]. Third, high throughput detection, which tends to be the primary goal for many organisations using an adversarial-resilient system, can negate low rates of false-positive detections that occur when a clean input is misclassified [41].

5.1 Surveillance-Specific Solutions

Specific defense mechanisms have been developed to address the vulnerabilities inherent in visual surveillance tasks across different application domains, including object detection, video recognition, tracking systems, and video quality assessment.

Object Detection Defense A notable solution for object detection is NutNet [12], which employs a lightweight autoencoder trained exclusively on clean data using a reconstruction-based approach. By leveraging reconstruction error to identify adversarial patches, this autoencoder successfully defends against both Hiding Attacks (HA) and Appearing Attacks (AA). A key advantage of NutNet is its parameter efficiency, requiring only approximately 5,000 parameters while maintaining effective detection performance [12].

Video Recognition Defense Defense Patterns (DPs) represent a solution designed for video recognition models [22]. This system enhances video robustness by generating and incorporating divergent frames into existing video sequences, thereby increasing the overall quality and resilience of the video stream. The generation of divergent frames occurs over a pre-trained network, which minimizes computational costs as no additional training is required concurrently with inference [22].

Tracking Systems For tracking systems, a frequency-based detection method has been

proposed that draws inspiration from steganalysis techniques [37]. This approach exploits the changes in relationships between neighboring pixels caused by adversarial perturbations. The detection model expresses the interdependence of adjacent pixels in the filtered image as a higher-order Markov chain, from which a transition probability matrix is constructed. This matrix serves as a feature vector for detecting adversarial manipulations in tracking systems [32, 37].

Video Quality Assessment SecureVQA provides a security-focused framework for protecting Video Quality Assessment (VQA) models against adversarial attacks [34]. The defense mechanism employs randomization as an intra-frame defense strategy through spatial grid sampling or pixel-level randomization within individual frames. Additionally, temporal information is leveraged as an inter-frame defense by reducing input resolution during both training and inference phases of VQA models. This multi-level approach enhances robustness without significantly compromising assessment accuracy [34].

5.2 Performance Analysis

Real-Time Capable Methods Among the surveyed defense mechanisms, several methods demonstrate the capability to operate under real-time constraints. NutNet exhibits a high level of efficiency, with its inference time increasing by only 8% over the baseline detector, thereby sufficiently meeting the real-time requirements associated with autonomous driving and surveillance applications [12]. Additionally, LSTM-AD (Long Short-Term Memory-based Adversarial Detection) has been shown to provide 91% detection accuracy while introducing minimal computational overhead. When evaluated on a UAV guidance system across ten time points, this method achieved a substantial reduction in computation time. To date, LSTM-AD remains the only method that has demonstrated practical real-time functionality for the specified detection task [4].

Comparative Performance Table 1 presents a runtime comparison between CNN-AD and LSTM-AD. While the base model run times are comparable between the two approaches (0.052 seconds for CNN-AD versus 0.056 seconds for LSTM-AD), substantial differences emerge in SHAP value computation and overall detection runtime. CNN-AD requires 13.33 seconds for SHAP value extraction, whereas LSTM-AD accomplishes this in only 0.019 seconds. Consequently, the total detection runtime for CNN-AD is 0.162 seconds, compared to just 0.001 seconds for LSTM-AD. According to prior studies, LSTM-AD achieves a 99.4% reduction in detection computation time relative to CNN-AD, highlighting the efficiency gains enabled by recurrent architectures in adversarial detection tasks [4].

Table 1. Run Time Comparison for Different Detectors.

Field	CNN-AD [4]	LSTM-AD [4]
Base Model Run Time (s)	0.052	0.056
SHAP Values Run Time (s)	13.33	0.019
Detection Run Time (s)	0.162	0.001

Scalability Issues Despite the progress made in adversarial defense, several scalability challenges persist, particularly those associated with adversarial training approaches. First, training a model capable of withstanding adversarial attacks is computationally expensive in both time and computational resources, especially when large-scale, high-dimensional datasets are involved [2, 12, 14]. The computational intensity required for generating adversarial examples and performing adversarial training limits the applicability of these methods to resource-constrained surveillance environments.

Second, specific detection methods incur high overhead that restricts the scalability of the overall detection system [41]. A notable example is Neural-network Invariant Checking (NIC), which requires training and storing parameters associated with each classifying layer in the neural network. This approach results in extremely high computational complexity and latency, making it unsuitable for large-scale deployment. Additional detection methods that introduce substantial overhead for saving classifier parameters include Selective and Feature-based Adversary Detection (SFAD) and Deep Neural Rejector (DNR) [41]. These overheads pose significant barriers to deploying robust adversarial defenses in real-world surveillance systems that must process high-volume video streams under strict latency constraints.

6 Evaluation Framework

The assessment of security and robustness in deep learning models for safety-critical applications, particularly surveillance, is structured around standardised datasets, comprehensive metrics, and rigorous threat modelling.

6.1 Datasets

Evaluation relies on a combination of widely accepted general benchmarks and domain-specific datasets

relevant to video analysis:

Standard Benchmarks: The most frequently used datasets for evaluating adversarial robustness and detection methods include CIFAR-10 (a small-image recognition task with 10 classes and 60,000 images, widely used in research [1, 5]) and ImageNet (a large-scale recognition task with 1,000 classes [1]). Due to computational requirements, ImageNet images are often downsampled, and sometimes only a subset of the validation set is sampled [1, 38]. Other common general benchmarks include MNIST, CIFAR-100, SVHN, Tiny-ImageNet, and MS COCO (used for object detection evaluations [1, 12]).

Surveillance-Relevant and Video Datasets: For evaluating temporal attacks and video analysis, key benchmarks include the CUHK Avenue and ShanghaiTech Campus datasets for video anomaly detection [2], UCF-101 and Kinetics-400 for action recognition [22], and LIVE-VQC, KoNViD-1k, YouTube-UGC, and LSVQ for evaluating Video Quality Assessment (VQA) security and accuracy [34].

6.2 Metrics

Detection Metrics Simple accuracy may not provide a reliable comparison of detector performance, as it does not account for class imbalance or the asymmetric costs of different error types. An appropriate evaluation of a detection system requires balancing the trade-offs between false negatives (FN) and false positives (FP) [40, 41].

Key detection metrics include the following. The True Positive Rate (TPR), also known as Recall or Detection Rate (DR), represents the proportion of adversarial examples correctly classified as adversarial [41]. The False Positive Rate (FPR) denotes the proportion of clean samples misclassified as adversarial; maintaining a low FPR is critical in surveillance applications where false alarms can lead to operator fatigue and reduced trust in the system [41]. The Area Under the ROC Curve (AUC-ROC) measures the area under the ROC curve plotting TPR against FPR, providing an aggregate assessment of detector performance across all classification thresholds, with values ranging from 0 to 1 [26, 41]. The F1-score serves as a combined measure of Recall and Precision, offering a harmonic mean that balances both metrics [41]. The Fooling Rate (FR) is used for evaluating model robustness and is defined as Q/P, where Q is the number of initially correctly classified images that become misclassified after perturbation, and P is the initial count of correctly classified samples [38].

Metrics for Measuring Robustness Robustness metrics quantify the resilience of a model when subjected to adversarial attacks [41]. Robust Accuracy refers to the accuracy of the classifier in the presence of attacks, measuring how well the model maintains correct predictions under adversarial conditions [18, 19]. The Attack Success Rate (ASR), also referred to as Robustness Failure, represents the proportion of attack instances that successfully satisfy the adversary's objective [5]. Average Distortion measures the magnitude of the perturbation added to the input, typically calculated using the ℓ_2 norm to quantify the minimal perturbation required to fool the model [38]. The Empirical Robustness Score (ERS) provides a comprehensive robustness assessment by taking the mean of normalized scores evaluated over a wide range of attack types and strengths [5].

Metrics for Evaluating Performance Evaluating the performance and practical deployability of defenses is imperative, particularly for real-time surveillance systems [41]. Inference Latency refers to the time taken by the combined model and detector to process an input; high levels of latency can be detrimental to real-time performance and may render a defense unsuitable for time-critical applications [3]. Frames Per Second (FPS) is a metric highly relevant to real-time object detection tasks, including autonomous driving and video surveillance, as it directly reflects the system's throughput and ability to process live video streams [12]. Overhead and Complexity encompass both the volume of parameters (model size) and the additional computational resources required for deployment. Complexity is closely related to the time and computational resources needed to train a detector, with lower complexity generally favoring practical adoption [41]. The Number of Queries is a critical variable for black-box attacks against database-driven systems, as the computational cost of executing an attack depends significantly on the number of query hits required to achieve the adversary's objective [16, 36].

6.3 Threat Model Considerations

The validity of any adversarial defense evaluation is conditional upon its assumptions regarding an aggressor's capability to execute attacks, a framework collectively referred to as the threat model [1, 5]. Two primary evaluation models exist based on the adversary's level of knowledge about the target system: white-box evaluation and black-box evaluation [1, 5].

In white-box attacks, also known as perfect knowledge scenarios, the adversary possesses complete access to the model's architecture, parameters, and deployed defenses. This comprehensive knowledge enables the attacker to compute gradients efficiently and craft adversarial examples with high precision. Attack methods such as Projected Gradient Descent (PGD) and Carlini & Wagner (C&W) serve as typical benchmarks for evaluating robustness under white-box conditions [1, 18, 38].

Conversely, black-box attacks operate under zero knowledge constraints, where the adversary can only access the model's outputs, such as class labels or confidence scores, without any insight into the internal structure or parameters [5, 16]. These attacks rely on two primary strategies. Transfer-based methods involve training a local substitute model to approximate the target model's decision boundary, with adversarial examples generated on the substitute transferring to the target. Query-based methods iteratively analyze output feedback to estimate gradients or directly optimize perturbations using only the model's responses, often requiring a large number of queries to succeed [16, 36].

Adaptive Attacks and Obfuscated Gradients Effective evaluation must consider an adversary who is aware of the deployed defenses and actively adapts their attack strategy accordingly [1, 5, 31]. Adaptive attacks represent scenarios where the attacker possesses knowledge of the defense mechanism and designs perturbations to jointly fool both the classifier and the detector [5, 12, 31]. Testing against adaptive attacks is essential because many defenses that appear robust under standard evaluation protocols fail when confronted with customized attacks specifically designed to circumvent the protective measures [5, 31].

Obfuscated gradients constitute a common defense pitfall wherein the method unintentionally masks its gradients, causing standard gradient-based attacks such as PGD to fail. This phenomenon can create a false sense of security, as defenses may appear robust when evaluated only against basic attacks while remaining vulnerable to more sophisticated approaches. Attacks such as C&W are necessary to test for this gradient masking and reveal the true robustness of a defense [5, 24].

Second-round adversarial attacks represent an advanced white-box threat where the adversary possesses knowledge of both the base model F_{base} and the detector D, and attacks both components simultaneously. This coordinated attack strategy poses a significantly greater challenge than targeting either component individually. Achieving robustness against second-round attacks requires complex feature manipulation and defense mechanisms that cannot be easily circumvented through joint optimization of the classifier and detector [31, 32].

7 Critical Analysis and Future Directions

Numerous unresolved Issues confront DNN (Deep Neural Network) security in many critical application areas such as surveillance due to limitations of available defence mechanisms and the relative ease with which adversaries can develop and exploit new types of adversarial attack models [1, 5].

7.1 Limitations

Today, there is tremendous difficulty in the use of adversarial defences and adversarial attack detection due to poor generalisation, high computational cost, and a loss of model performance on clean input [1, 5, 41].

Limited Generalisation Against Unseen Attacks: Some detection mechanisms demonstrate great sensitivity to the types of AEs used during taining and do not generalise well against unseen attacks [14, 23, 41]. Because the training of an auxiliary detector implicitly trains on a limited range of possible attacks, there is potential for future modifications to avoid being detected from the system [23, 41]. In addition, there are many input-transformation based defences that take advantage of obscured gradients, which demonstrates a lack of inherent robustness in these systems because there are stronger adaptive attacks that will be able to bypass them [1, 5, 24].

Computational overhead vs real-time constraints: Many robust defence mechanisms will impose significant computational burdens making their use impractical for real-time applications with limited resources such as Autonomous Driving Systems (ADS) [3, 4, 41]. The use of methods that require expensive computational steps to perform, such as computing statistics (i.e. kernel density or LID) in intermediate layers or multiple classifiers/auxiliary models, will increase the memory and inference latencies associated with making decisions using these approaches [14, 25, 26, 41]. For example, AT would require on-the-fly generation of adversarial examples (AE), which is a time consuming task across larger models and/or higher resolution images [3, 6, 18, 19].

Accuracy of Trade-off against Clean Accuracy: The use of many defensive techniques for improving the robustness of the machine learning models will result in degradation of the accuracy of the models on the clean inputs [3, 5, 18, 19, 34, 41]. This being the case in many situations, it is considered as an intrinsic issue associated with the machine learning models' learning of the real concepts [1, 5]. In other words, SecureVQA technique would enhance the security of the models but at the expense of the clean accuracy of the VQA model [34]. Again, Defense Patterns (DPs) used for video recognition models would lead to minor reductions in the recognition accuracy of the models on clean videos [22]. Moreover, there is 10–15% reduction in the clean accuracy of machine learning models based on CIFAR-10 using adversarial training [18, 19].

7.2 Research Gaps

Focused research to improve the security and resilience of deep learning systems is needed in several areas [1, 41].

Video Streams vs Static Images: Most studies have been performed using static images exclusively, whereas considerably less attention has been paid to studying video streams, which form an important part of both surveillance and autonomous systems. Therefore, there is a lack of video studies [2, 3]. Moreover, several of the proposed defensive mechanisms (to counter attacks) for static images may not be applicable to videos due to their temporal nature, which is disregarded when coding videos [2, 22]. The need to study more defendable attack methodologies for video-based anomaly detection models is evident, as most current black-box attacks have been devised for action recognition tasks [2]. While researching more in-depth, future works must investigate the influence of AEs on other data types (time-series, graph, etc.) [1, 41].

Absence of Temporal and Video-based Benchmarks: Another problem with existing research work is that a majority of it uses static image datasets such as CIFAR-10, CIFAR-100, and ImageNet to measure the effectiveness of defense techniques against attacks. Even though these are standard benchmark datasets for measuring the performance of models against attacks, there is no way in which they can measure the effectiveness of a model that needs to operate in video feeds where frames have dependencies between each other. In the future, emphasis should be placed on creating video-based benchmarks based on well-known video datasets like ShanghaiTech Campus [2], CUHK Avenue [2], UCF-101, and Kinetics-400 [22].

Detection and Alert Systems are Not Unified: Safety-critical applications greatly suffer from detection mechanisms and alert/intervention systems being separate entities. Therefore, the lack of immediate fallback (e.g., requesting a human to intervene or disabling the autonomous operation of an automated driving system) after an adversarial perturbation is detected may limit the usefulness of the detection system. A primary direction of future research on detection systems should be the development of reliable detection systems that provide timely alerts while minimizing resource usage.

Detection of Physical Attacks: Attacks that take place in the real world (e.g., the use of adversarial Patch & changes to traffic signs) present a serious threat to safety and security; however, functional detection of these types of attacks remain a significant gap in knowledge. Since adversarial examples have been proven to translate very well into the physical world, there must be effective defense mechanisms established in order to reduce the large accuracy gap that exists between certified patch defense mechanisms and standard models.

7.3 Future Directions

Future research that focuses on the development of Detection and Defence Strategies that are Comprehensive, Efficient, and Robust [1, 5, 41].

Unified Detection Frameworks: Developing Detection Techniques to Generalize on Distinct Attack Types, Models, and Inputs, is a vital part of future research [5, 25, 41]. Using Unsupervised Detection Techniques Focusing on Modelling the Distribution of Normal samples, would provide an additional level of flexibility to an algorithm to better Defend against unknown Attacks [14, 27, 41]. Ensemble Detection Techniques (using like uncertainty prediction with bi-match prediction) are also suggested; however, these need to perform well without affecting FPR [41]. A promising avenue of future research is combining Bayesian Neural Networks (BNN) with Feature-Based Detection Techniques [29, 30].

Online Learning and Lightweight Detection Techniques: There is an urgent need for Algorithms that Complete Tasks Quickly and with Low Resource Usage in Real-Time (multi-model) and Low Resource (Designated Tasks) Environments [3, 4, 12, 41]. Future research should focus on Online Learning

and Lightweight Detection Techniques to Minimize Overhead for Completion of the Stringent Real-Time Constraints Required for Certain Duties such as UAV Flight [4, 12, 41].

Certified Robustness and Multi Layer Defences: For the advancement of AI Technologies, pursuing Non-Single/One Layered Techniques and Continuously Pursuing Mathematically Certifiable Defence Mechanisms is a requirement moving forward [5, 31, 35]. The application of Multi-Layer Detection Systems will provide more options, but will create Computation Intense Maintenance Challenges for Implementators/Researchers [25, 41]. Certified Robustness methods can be mathematically guaranteed against All Perturbations for a Specific Cost Limit, and provide an increased level of Trustworthiness for AI Systems against Future Unknown Attacks [5, 35].

8 Conclusion

Many researchers have studied deep learning systems but they are not very robust because deep learning generally operates on a restricted set of benign examples instead of learning actual underlying concepts [1, 5, 40]. Various gradient-based attacks (PGD, C&W) and decision-based attacks have been able to bypass defenses like defensive distillation [1, 16, 38]. We recommend that, for practical purposes, practitioners use a multilayered defense strategy that includes strong, lightweight detection systems at the beginning of the pipeline so that fallback solutions such as human review or shutting down autonomous operation can be triggered [3, 23, 24]. Empirical evidence exists to demonstrate that adversarial examples can be detected [24, 26, 41] and that highly effective processes (such as NutNet and LSTM-AD) can detect anomalies in real time and with little impact on performance [4, 12]. Finally, for maximum security, the detector should use unsupervised (MIAED, etc.) and ensemble techniques, because supervised techniques do not generalize well to unknown threats [14, 27, 41]. In order for the vision to build secure surveillance systems to become a reality, the deep learning models themselves (BNN and/or robust ViTs) must contain intrinsic robustness [29, 30, 35]. We are moving toward certifying guarantees against all forms of perturbation, given a set budget [5, 35]. Lastly, future research should broaden on how deep learning models handle complex data types, and specifically focus on video streams, archival time series, and graph data [1, 2, 22].

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE access, 9*, 155161-155196. [CrossRef]

[2] Mumcu, F., Doshi, K., & Yilmaz, Y. (2022). Adversarial machine learning attacks against video anomaly detection systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 206-213). [CrossRef]

[3] Deng, Y., Zhang, T., Lou, G., Zheng, X., Jin, J., & Han, Q. L. (2021). Deep learning-based autonomous driving systems: A survey of attacks and defenses. *IEEE Transactions on Industrial Informatics, 17*(12), 7897-7912. [CrossRef]

[4] Hickling, T., Aouf, N., & Spencer, P. (2023). Robust adversarial attacks detection based on explainable deep reinforcement learning for UAV guidance and planning. *IEEE Transactions on Intelligent Vehicles, 8*(10), 4381-4394. [CrossRef]

[5] Malik, J., Muthalagu, R., & Pawar, P. M. (2024). A systematic review of adversarial machine learning attacks, defensive controls, and technologies. *IEEE Access, 12*, 99382-99421. [CrossRef]

[6] Singh, T., Rai, P., Sood, S., Nigam, S., & Kumari, S. (2023, August). Adversarial Attacks and Defences of Various Artificial Intelligent Models. In *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)* (pp. 210-215). IEEE. [CrossRef]

[7] Yin, Z., Zhu, S., Su, H., Peng, J., Lyu, W., & Luo, B. (2025). Adversarial examples detection with enhanced image difference features based on local histogram equalization. *IEEE Transactions on Dependable and Secure Computing*. [CrossRef]

[8] Zhang, C., Yu, S., Tian, Z., & Yu, J. J. (2023). Generative adversarial networks: A survey on attack and defense perspective. *ACM Computing Surveys, 56*(4), 1-35. [CrossRef]

[9] Feng, W., Xu, N., Zhang, T., Wu, B., & Zhang, Y. (2023). Robust and generalized physical adversarial attacks via meta-GAN. *IEEE Transactions on Information Forensics and Security, 19*, 1112-1125. [CrossRef]

[10] Lee, M., & Kolter, Z. (2019). On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*. [arXiv]

[11] Liu, X., Yang, H., Liu, Z., Song, L., Li, H., & Chen, Y. (2018). Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*. [arXiv]

[12] Lin, Z., Zhao, Y., Chen, K., & He, J. (2024, December). I don't know you, but I can catch you: Real-time defense against diverse adversarial patches for object detectors. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 3823-3837). [CrossRef]

[13] Li, Y., Ruan, S., Qin, H., Deng, S., & El-Yacoubi, M. A. (2023). Transformer based defense GAN against palm-vein adversarial attacks. *IEEE Transactions on Information Forensics and Security, 18*, 1509-1523. [CrossRef]

[14] Al-Andoli, M. N., Tan, S. C., Sim, K. S., Goh, P. Y., & Lim, C. P. (2024). A framework for robust deep learning models against adversarial attacks based on a protection layer approach. *IEEE Access, 12*, 17522-17540. [CrossRef]

[15] Wang, S., Nepal, S., Rudolph, C., Grobler, M., Chen, S., Chen, T., & An, Z. (2021). Defending adversarial attacks via semantic feature manipulation. *IEEE Transactions on Services Computing, 15*(6), 3184-3197. [CrossRef]

[16] Li, C., Wang, H., Zhang, J., Yao, W., & Jiang, T. (2022). An approximated gradient sign method using differential evolution for black-box adversarial attack. *IEEE Transactions on Evolutionary Computation, 26*(5), 976-990. [CrossRef]

[17] Pavlitskaya, S., Hendl, J., Kleim, S., Müller, L. J., Wylczoch, F., & Zöllner, J. M. (2022, November). Suppress with a patch: Revisiting universal adversarial patch attacks against object detection. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* (pp. 1-6). IEEE. [CrossRef]

[18] He, S., Wei, J., Zhang, C., Xu, X., Song, J., Yang, Y., & Shen, H. T. (2024). Boosting adversarial training with hardness-guided attack strategy. *IEEE Transactions on Multimedia, 26*, 7748-7760. [CrossRef]

[19] Dhanaraj, R. S., & Sridevi, M. (2024). Building a robust and efficient defensive system using hybrid adversarial attack. *IEEE Transactions on Artificial Intelligence, 5*(9), 4470-4478. [CrossRef]

[20] Kuang, H., Liu, H., Lin, X., & Ji, R. (2024). Defense against adversarial attacks using topology aligning adversarial training. *IEEE Transactions on Information Forensics and Security, 19*, 3659-3673. [CrossRef]

[21] Liao, W., Liu, Z., Shen, M., Chen, R., & Liu, X. (2024). Apr-net: Defense against adversarial examples based on universal adversarial perturbation removal network. *IEEE Transactions on Artificial Intelligence, 6*(4), 945-954. [CrossRef]

[22] Lee, H. J., & Ro, Y. M. (2023). Defending video recognition model against adversarial perturbations via defense patterns. *IEEE Transactions on Dependable and Secure Computing, 21*(4), 4110-4121. [CrossRef]

[23] Metzen, J. H., Genewein, T., Fischer, V., & Bischoff, B. (2017). On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*. [arXiv]

[24] Xu, W., Evans, D., & Qi, Y. (2017). Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*. [arXiv]

[25] Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., ... & Bailey, J. (2018). Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*. [arXiv]

[26] Feinman, R., Curtin, R. R., Shintre, S., & Gardner, A. B. (2017). Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*. [arXiv]

[27] Gao, S., Wang, R., Wang, X., Yu, S., Dong, Y., Yao, S., & Zhou, W. (2023). Detecting adversarial examples on deep neural networks with mutual information neural estimation. *IEEE Transactions on Dependable and Secure Computing, 20*(6), 5168-5181. [CrossRef]

[28] Zhu, B., Dong, C., Zhang, Y., Mao, Y., & Zhong, S. (2023). Toward universal detection of adversarial examples via pseudorandom classifiers. *IEEE Transactions on Information Forensics and Security, 19*, 1810-1825. [CrossRef]

[29] Li, Y., Tang, T., Hsieh, C. J., & Lee, T. C. (2024). Adversarial examples detection with Bayesian neural network. *IEEE Transactions on Emerging Topics in Computational Intelligence, 8*(5), 3654-3664. [CrossRef]

[30] Bortolussi, L., Carbone, G., Laurenti, L., Patane, A., Sanguinetti, G., & Wicker, M. (2024). On the robustness of bayesian neural networks to adversarial attacks. *IEEE Transactions on Neural Networks and Learning Systems, 36*(4), 6679-6692. [CrossRef]

[31] Qin, C., Chen, Y., Chen, K., Dong, X., Zhang, W., Mao, X., ... & Yu, N. (2022). Feature fusion based adversarial example detection against second-round adversarial attacks. *IEEE Transactions on Artificial Intelligence, 4*(5), 1029-1040. [CrossRef]

[32] Liu, J., Zhang, W., Zhang, Y., Hou, D., Liu, Y., Zha, H., & Yu, N. (2019). Detection based defense against adversarial examples from the steganalysis point of view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4825-4834). [CrossRef]

[33] Sun, S., Nwodo, K., Sugrim, S., Stavrou, A., & Wang, H. (2024, December). Vitguard: Attention-aware detection against adversarial examples for vision transformer. In *2024 Annual Computer Security Applications Conference (ACSAC)* (pp. 1259-1275). IEEE. [CrossRef]

[34] Zhang, A. X., Wang, Y. G., Ran, Y., Tang, W., Guan, Q., & Yang, C. (2025). Secure video quality assessment resisting adversarial attacks. *IEEE Transactions on Broadcasting*. [CrossRef]

[35] Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, May). Certified adversarial robustness via randomized smoothing. In *international conference on machine learning* (pp. 1310-1320). PMLR.

[36] Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017, November). Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security* (pp. 15-26). [CrossRef]

[37] Li, C., Li, H., & Zhang, G. (2025). Detecting Adversarial Attacks Based on Tracking Differences in Frequency Bands. *IEEE Transactions on Multimedia, 27*, 4597-4612. [CrossRef]

[38] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*. [arXiv]

[39] Graese, A., Rozsa, A., & Boult, T. E. (2016, December). Assessing threat of adversarial examples on deep neural networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 69-74). IEEE. [CrossRef]

[40] Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*. [arXiv]

[41] Aldahdooh, A., Hamidouche, W., Fezza, S. A., & Déforges, O. (2022). Adversarial example detection for DNN models: A review and experimental comparison. *Artificial Intelligence Review, 55*(6), 4403-4462. [CrossRef]

Sharanya Peri is currently pursuing the B.E. degree in Artificial Intelligence and Data Science (AI&DS) from Chaitanya Bharathi Institute of Technology (CBIT), Gandipet, Hyderabad, India. (Email: sharanyaperi20@gmail.com)

Navya Vadla is currently pursuing the B.E. degree in Artificial Intelligence and Data Science (AI&DS) from Chaitanya Bharathi Institute of Technology (CBIT), Gandipet, Hyderabad, India. (Email: navyavadla098@gmail.com)

Yeshwanth Panuganti is currently pursuing the B.E. degree in Artificial Intelligence and Data Science (AI&DS) from Chaitanya Bharathi Institute of Technology (CBIT), Gandipet, Hyderabad, India. (Email: 7yeshwanth@gmail.com)

Kadiyala Ramana is currently an Associate Professor in the Department of Information Technology at Chaitanya Bharathi Institute of Technology (CBIT), Hyderabad, India. He received his Ph.D. and has over a decade of experience in teaching and research. His extensive research portfolio includes over 50 publications in prestigious journals such as IEEE Transactions on Intelligent Transportation Systems, Expert Systems, and Multimedia Tools and Applications. His primary research interests encompass the Internet of Things (IoT), Blockchain technology, Wireless Sensor Networks (WSNs), and Deep Learning applications for smart agriculture and traffic management. He serves as an active reviewer for numerous high-impact journals, including IEEE Transactions on Industrial Informatics and Applied Soft Computing. (Email: ramana.it01@gmail.com)