Beyond Boundaries: A Comprehensive Survey of Transferable Attacks on AI Systems

Guangjing Wang, Ce Zhou, Yuanda Wang, Bocheng Chen, Hanqing Guo and Qiben Yan
SEIT Lab, Michigan State University, USA

Abstract

Artificial Intelligence (AI) systems such as autonomous vehicles, facial recognition, and speech recognition systems are increasingly integrated into our daily lives. However, despite their utility, these AI systems are vulnerable to a wide range of attacks such as adversarial, backdoor, data poisoning, membership inference, model inversion, and model stealing attacks. In particular, numerous attacks are designed to target a particular model or system, yet their effects can spread to additional targets, referred to as transferable attacks. Although considerable efforts have been directed toward developing transferable attacks, a holistic understanding of the advancements in transferable attacks remains elusive. In this paper, we comprehensively explore learning-based attacks from the perspective of transferability, particularly within the context of cyber-physical security. We delve into different domains – the image, text, graph, audio, and video domains – to highlight the ubiquitous and pervasive nature of transferable attacks. This paper categorizes and reviews the architecture of existing attacks from various viewpoints: data, process, model, and system. We further examine the implications of transferable attacks in practical scenarios such as autonomous driving, speech recognition, and large language models (LLMs). Additionally, we outline the potential research directions to encourage efforts in exploring the landscape of transferable attacks. This survey offers a holistic understanding of the prevailing transferable attacks and their impacts across different domains.

Contents

1 Introduction

Deep learning has initiated a revolutionary era of AI. Beginning with the advancement in computer vision (CV), current large language and large multimodal models demonstrate significant improvements in information processing capabilities. For example, empowered by Generative Pre-trained Transformers (GPT), GPT-4 [1] excels in multimodality tasks ranging from content creation to problem-solving. Deep learning has also been a pivotal technology in advancing autonomous driving systems, which leverage complex algorithms (*e.g.*, object detection, semantic segmentation) to process data from sensors in real-time, making informed decisions to maneuver vehicles safely through various driving conditions and environments.

However, despite their remarkable success, AI systems are vulnerable to numerous attacks. Researchers have been developing new attacks to compromise deep learning models. Particularly, adversarial attacks [2, 3, 4, 5, 6, 7, 8, 9, 10] involve the generation of adversarial perturbation or patches to mislead the model output. The adversarial perturbation or patch can be regarded as a type of noise, which can be merged with different clean samples to deceive the model. Poisoning [11, 12, 13] and backdoor [14, 15, 16, 17] attacks also have a substantial impact on the AI systems. A data poisoning attack corrupts the training data, causing the model to yield errors in predictions or classifications. The poisoning attack compromises the model's overall behavior, detrimentally affecting its performance across a broad spectrum of inputs. Similarly, a backdoor attack embeds a trigger into the model via malicious training data. As a result, when the model encounters inputs containing this trigger during deployment, it generates incorrect or unauthorized results. Model-stealing attacks [18, 19, 20, 21, 22, 23] aim to steal the pretrained proprietary models. Beyond the theft of the model itself, attackers also attempt to steal the underlying data, known as an inference attack. There are two primary forms of inference attacks. The first is the model inversion attack [24, 25, 26, 27], which aims to reconstruct or approximate the model's original training or input data. This is done by exploiting the model's outputs or architecture to reverse-engineer the training process. The second form is the membership inference attack [28, 29, 30, 31, 32], where the objective is to determine if a specific data point was used in the training set of a machine learning model. Attackers scrutinize the model's outputs to infer whether a given data entry was part of the training dataset.

Numerous learning-based attacks show great transferability, where an attack model or method developed for one system is effective against another system, even though the two systems may have different architectures or configurations. In essence, transferable attacks encompass a broad spectrum of attacks that are transferable across various samples within the same or different domains, different models, and even across different systems. This transferability makes such attacks particularly concerning in the realm of Cyber-physical security. In this survey, we provide a high-level classification of transferable attacks, and the survey is structured as follows. First, we introduce the background knowledge about transfer learning, optimization methods and divergence measurement in Section 2, which lays the foundation for learning-based transferable attacks. Then, we summarize existing transferable attacks according to their attack designs as shown in Fig. 1. (i) We study transferable attacks from a data-centric perspective in Section 3. Existing research [33, 18, 23] has developed data argumentation mechanisms by synthesizing new data or transforming data. Data augmentation improves the diversity of the training data, leading to better generalization of learning-based attack models. Moreover, other studies [34, 35, 36] have revealed that disentangling the content and style features can boost the transferability of attacks. (ii) We review transferable attacks from the perspective of the learning process in Section 4, using the universal adversarial attack as a case study. The universal adversarial attacks aim to generate universal perturbations or patches applicable across various inputs, resulting in misclassification for any given input. To find such universal perturbations, researchers utilize different learning methods including gradient-based [37, 38, 39] and heuristic-based methods [40, 41, 42]. Additionally, generative models such as Generative Adversarial Networks (GANs) are also employed to generate these universal perturbations [43, 44, 45].

(iii) From the model perspective, we review model augmentation-based methods in Section 5. For black-

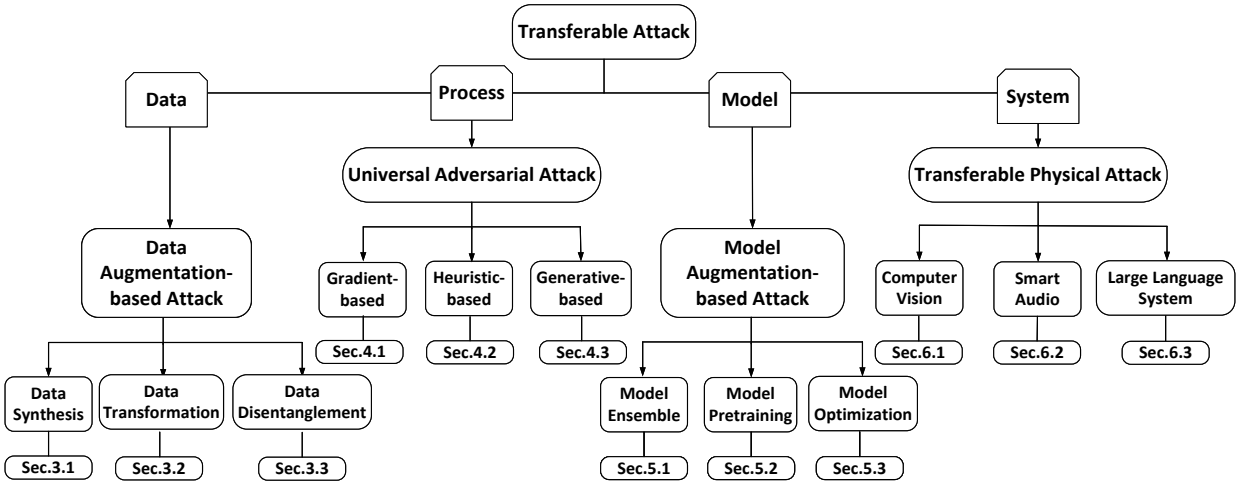

Figure 1: The taxonomy of transferable attacks.

box attacks, the adversary has no knowledge of the target system including the specific model architecture, model parameters, and training strategies. The black-box attacks can be executed by either attacking a surrogate model or using gradient estimation methods by querying the target model. We mainly focus on the black-box attacks that aim to construct a surrogate model that portrays a similar latent space as the potential target model [46, 40, 19]. This approach allows attackers to craft attacks against the surrogate model and subsequently transfer these attacks to the target model. To enhance the transferability of attacks, one strategy is to design attacks targeting ensembled surrogate models [47, 48, 21]. This approach leads to the generation of adversarial perturbations that are more effective across different models. Another strategy is to utilize model stealing attacks [49, 20, 50], which aim to obtain the model behind the black-box system. Then, attackers can utilize various white-box methods to compromise these systems. Furthermore, several researchers have targeted pre-trained foundation models [51], which could affect the downstream fine-tuned models. For example, data poisoning and backdoor attacks [52, 53, 54] are launched against the pre-trained foundation models, with the attacks remaining effective on the fine-tuned downstream models.

(iv) We illustrate the transferable attacks from the system perspective in Section 6. We focus on computer vision systems, speech recognition systems, and large language model (LLM) systems. For each system, the attack design depends on the features of the specific modality and the threat model of the system. For example, biometric security is one of the essential topics in computer vision and speech processing. Different biometric applications such as facial recognition, 3D face authentication, and voice authentication have different model designs, which require the adaptation of specific attacks. Finally, in Section 7, we summarize the limitations of current transferable attacks and propose future directions for their design. Our final goal is to motivate the development of robust, adaptive, and secure interconnected cyber-physical systems.

The Survey's Unique Contribution. Compared to the existing surveys, this survey comprehensively evaluates various attacks from the perspective of transferability. Table 1 lists the existing surveys on various attacks. Each type of learning-based attack has different design principles according to different data modalities such as image, text, audio, video, and graph. Most of the surveys [55, 64, 96, 82] only focus on a specific type of attack such as adversarial attacks or model stealing attacks. By contrast, we summarize all six widely studied attacks from the perspective of transferability. The concept of transferability is crucial because it reveals that the threat landscape is broader than anticipated, as machine learning models can be susceptible to different attacks even when the attacker does not have direct access to the model's architecture or parameters. The transferability raises questions about the generalizability and security of machine learning

Table 1: Recent Surveys for Learning-based Attacks.

Survey Topic	Survey Domain	Publish Venue	Publish Year
Adversarial Attack and Defense [55]	Graph	IEEE TKDE	2022
Adversarial Attack and Defense [56]	NLP	Neurocomputing	2022
Adversarial Universal Attack [57]	General	IJCAI	2021
Adversarial Attack and Defense [58]	General	CAAI TIT	2021
Adversarial Attack and Defense [59]	General	IJAC	2020
Adversarial Attack [60]	NLP	ACM TIST	2020
Adversarial Attack and Defense [61]	General	IET	2018
Adversarial Attack [62]	CV	IEEE Access	2018
Backdoor Attack [63]	Large Language Model	ArXiv	2023
Backdoor Attack [64]	Audio	ArXiv	2023
Backdoor Attack and Defense [65]	General	IEEE OJCS	2023
Backdoor Attack [66]	NLP	ArXiv	2023
Backdoor Attack and Defense [67]	Federated Learning	ArXiv	2023
Backdoor Attack and Defense [68]	General	IEEE TPAMI	2022
Backdoor Attack [17]	General	IEEE NNLS	2022
Backdoor Attack and Defense [69]	General	IEEE OJSP	2022
Backdoor Attack and Defense [70]	NLP	IEEE QRS	2022
Backdoor Defense [71]	General	Neurocomputing	2021
Backdoor Attack and Defense [72]	General	ArXiv	2020
Backdoor Attack [73]	General	IEEE ISQED	2020
Data Poisoning Attack [13]	General	ACM CSUR	2023
Data Poisoning Attack [74]	General	ACM TECS	2023
Data Poisoning Attack [75]	Federated Learning	IEEE Access	2023
Data Poisoning Attack and Defense [76]	General	ACM CUSR	2022
Data Poisoning Attack and Defense [77]	General	DCN	2022
Data Poisoning Attack and Defense [78]	General	ArXiv	2022
Data Poisoning Attack and Defense [79]	General	DSC	2022
Data Poisoning Attack and Defense [68]	General	IEEE TPAMI	2022
Data Poisoning Attack and Defense [80]	General	ACM CUSR	2022
Data Poisoning Attack [81]	Gereral	Springer ACeS	2021
Membership Inference Defense [82]	General	ACM CUSR	2023
Membership Inference Attack and Defense [83]	General	CSI	2023
Membership Inference Attack [84]	General	IEEE CM	2022
Membership Inference Attack [30]	General	ACM CUSR	2022
Membership Inference Attack [85]	General	Management	2021
Membership Inference Attack and Defense [86]	General	CSI	2021
Membership Inference Defense [87]	General	AASCS	2020
Model Inversion Attack [88]	General	CSF	2023
Model Inversion Attack and Defense [89]	General	IJCAI	2022
Model Inversion Attack [90]	General	TechXiv	2022
Model Inversion Attack [27]	General	ICDCCN	2022
Model Inversion Attack [91]	General	ACM CUSR	2021
Model Inversion Attack [92]	General	ACM CUSR	2021
Model Inversion Attack [93]	General	IEEE TOSE	2020
Model Inversion Attack [94]	General	ACM CUSR	2020
Model Inversion Attack [95]	General	ArXiv	2017
Model Inversion Attack and Defense [24]	General	ACM CCS	2015
Model Stealing Attack [96]	General	IEEE CSR	2023
Model Stealing Attack and Defense [20]	General	ACM CUSR	2023
Model Stealing Attack and Defense [97]	General	IEEE CM	2020
Model Stealing Attack [93]	General	IEEE TOSE	2020
Transferable Learning-based Attacks (Ours)	General	-	2023

models in real-world applications. Some surveys [56, 62, 64] only focus on a specific application area such as natural language processing (NLP). Rather than limiting to a specific area, we review transferable attacks in different modalities across different applications and systems. We categorize and summarize transferable attack designs by examining them from the perspectives of data, processes, models, and systems. In particular, we review the impact of transferable attacks on typical cyber-physical systems, including autonomous driving, speech recognition, and LLM systems. Overall, this survey offers comprehensive insights into the state-of-the-art transferable attack strategies, enabling both horizontal and vertical comparisons of different attack types across various domains.

2 Background

In this section, we provide preliminary knowledge including transfer learning, optimization methods, and divergence measurement for designing transferable attacks.

2.1 Transfer Learning

Transfer learning is a general concept that employs the similarity in data, task, or model from an existing problem to a new one in order to facilitate the sharing of knowledge. A key concept in transfer learning is *Domain*, denoted as $\mathcal{D}$, which includes two parts: data sample (x, y) and data distribution $P(x, y)$, where x is the input (*e.g.*, feature), and y is the output (*e.g.*, label). Correspondingly, the feature space is denoted by $\mathcal{X}$, and the label space is denoted by $\mathcal{Y}$. Overall, a domain is composed by $D = \{\mathcal{X}, \mathcal{Y}, P(x, y)\}$. Generally, we have two domains for transfer learning: source domain $\mathcal{D}_s$ and target domain $\mathcal{D}_t$. $\mathcal{D}_s$ and $\mathcal{D}_t$ have differences in at least one of the conditions: $\mathcal{X}_s \neq \mathcal{X}_t$ (*e.g.*, different features), $\mathcal{Y}_s \neq \mathcal{Y}_t$ (*e.g.*, different tasks) or $P_s(x, y) \neq P_t(x, y)$ (*e.g.*, different probability distributions). For detailed guidance to transfer learning algorithms and applications, we refer to the related surveys and books [98, 99, 100, 101].

There are three main scenarios that demand the consideration of transfer learning. First, there is insufficient labeled data in the target domain, and the training data from the source domain differs from that of the target domain. Meanwhile, it is important to be aware of the negative transfer, which can deteriorate results in the target domain. Negative transfer typically results from discrepancies between data from the source and target domains, as well as from inadequate algorithms. Second, there are pre-trained foundation models [51], which can be utilized for various target applications with limited computational capability. Third, there are generalization and personalization requirements. For generalization, it requires the model to be generalized to accurately predict unseen samples, applications, or environments. For personalization, it would require the model to be adaptive to users' specific habits or preferences.

When designing transfer learning models, we typically evaluate factors such as the presence of labels in the target domain, the similarity of feature spaces between domains, whether the learning should be conducted online or offline, and the most appropriate learning strategy to utilize. For example, *domain adaptation* occurs under the condition that $P_s(x, y) \neq P_t(x, y)$ when $\mathcal{X}_s = \mathcal{X}_t$ and $\mathcal{Y}_s = \mathcal{Y}_t$, indicating different probability distributions in the source and target domains. Here, we describe a toy example of $P(x, y)$ variation. Suppose we have the Gaussian distribution with different variances: $N_1(0, 2)$, $N_2(0, 3)$, and $N_3(0, 5)$. In the context of domain adaptation, the training data adheres to $N_1(0, 2)$, while the test data may conform to $N_2(0, 3)$ or $N_3(0, 5)$. Note that in real-life problems, the true data distribution is exceedingly intricate, making it challenging to precisely define the probability mass function. Indeed, contemporary machine learning approaches, including deep learning, strive to understand and model data distributions. Domain adaptation has been explored across a range of applications. For example, in smart home security, Glint [102] incorporates supervised domain adaptation where the target domain is fully labeled but the number of samples is insufficient for training a robust threat detection model. In mobile sensing, Facer [103] involves the unsuper-

vised domain adaption where the target domain is unlabeled, so as to generalize the expression recognition model to incoming new users (target domain).

2.2 Optimization Methods

Optimization methods play a pivotal role in a multitude of fields for fine-tuning parameters and optimizing system performance. In this section, we introduce three basic optimization methods: gradient-based method, heuristic-based method, and bi-level optimization method.

2.2.1 Gradient-based Method

The gradient-based optimization method aims to find the optimal parameters that minimize or maximize a given objective function. These methods are commonly used in machine learning models such as deep neural networks, linear regression, and logistic regression. Gradient descent is used to find the minimum of a function. First, we need to initialize the model parameters. We start with an initial estimate for the parameters (weights and biases) of the model, often set randomly or to some default values. Second, we calculate the gradient of the objective function with respect to the parameters. The gradient represents the direction and magnitude of the steepest increase in the function. Third, we adjust the parameters in the opposite direction of the gradient to move towards the minimum. This adjustment is controlled by a hyperparameter α called *learning rate*. The update rule for gradient descent is:

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \cdot \nabla J(\theta_{\text{old}}), \tag{1}$$

where θ_{new} is updated parameters, θ_{old} is old parameters, $\nabla J(\theta_{\text{old}})$ is the gradient of the objective function J with respect to θ_{old}. The process continues until the algorithm converges to a minimum, where the gradient becomes close to zero. Gradient ascent is the counterpart of gradient descent and is used to find the maximum of a function. The process is similar, but instead of moving in the direction of the negative gradient, it moves in the direction of the positive gradient.

2.2.2 Heuristic-based Method

The heuristic-based optimization method is a class of techniques used to find approximate solutions, especially when finding an exact solution is computationally infeasible or time-consuming. These methods use heuristics, which are rules of thumb or strategies that guide the search for the best solution. Generally, heuristic-based methods rely on intelligent search strategies rather than explicit mathematical models.

Heuristic-based methods typically involve four key procedures. First, heuristic-based optimization begins with an initial solution or state. This initial solution can be generated randomly or from domain-specific knowledge. Second, the core of heuristic methods involves exploring the neighborhood of the current solution. This neighborhood search is guided by specific rules or heuristics. Typically, the method generates neighboring solutions by making small, localized changes to the current solution. Third, at each step, an evaluation function is used to assess the quality of the current solution or state. The evaluation function quantifies how close the current solution is to an optimal one. This function helps decide whether to accept or reject a new solution based on certain criteria. Fourth, heuristic methods use acceptance criteria to determine whether a neighboring solution should be accepted as the new current solution. Common criteria include comparing the objective function values of the current and neighboring solutions and considering factors such as minimizing or maximizing a cost or maximizing utility.

Here, we present three examples of heuristic-based methods used for optimization in discrete spaces such as text data space: (i) In the hill climbing algorithm, the search starts at an initial solution and iteratively moves to the neighboring solution with the highest evaluation function value. It stops when no better

neighbors are found. (ii) Simulated annealing, inspired by the annealing process in metallurgy, explores a wider solution space, allowing for occasional moves to solutions with worse objective values to avoid getting stuck in local optima. (iii) Evolutionary algorithms such as genetic programming and natural evolution strategy are widely used in numerical optimization in discrete space or black-box setting. Generally, they use concepts of selection, mutation, and crossover to explore a population of potential solutions.

2.2.3 Bi-level Optimization Method

The bi-level optimization addresses scenarios where optimization itself is nested within another optimization problem. This class of methods is essential for hyperparameter tuning in machine learning, where one optimization task is nested within the broader optimization task of model training. Related to Stackelberg games, bi-level optimization deals with two interconnected optimization problems: the upper-level problem (leader) and the lower-level problem (follower). The solution at the upper-level influences the lower-level problem, while conversely, the solution at the lower-level also impacts the upper-level problem.

For the upper-level problem, the leader determines a set of decision variables x that affect the lower-level problem. The leader aims to optimize an objective function $F(x, y)$, which depends on both the upper-level decision variables x and the lower-level solutions y. For the lower-level problem, the follower selects a set of decision variables y that are typically influenced by the upper-level decisions. The follower aims to optimize an objective function $G(x, y)$ that depends on both the upper-level decisions x and the lower-level variables y. The goal is to find a set of upper-level decisions x and lower-level decisions y that jointly optimize both objective functions while satisfying any constraints.

The optimization process is similar to expectation maximization, which is an iterative solution to maximum likelihood estimation with latent variables. First, we start with initial guesses for both the upper-level decisions x and lower-level decisions y. Second, we solve the lower-level problem while keeping the upper-level decisions x fixed. Third, we update the upper-level decisions x by solving the upper-level problem while keeping the lower-level decisions y fixed. These steps are iterated until convergence is achieved, often guided by stopping criteria. We refer readers to the survey [104] about bi-level optimization for learning.

2.3 Divergence Measurement

In this section, we introduce several divergence measurement methods including maximum mean discrepancy, explicit distance measurement, and implicit measurement.

2.3.1 Maximum Mean Discrepancy

The concept behind Maximum Mean Discrepancy (MMD) involves initially transforming two distributions into another space using a kernel function and subsequently calculating the disparity in their means. MMD can directly assess the divergence between the marginal distributions, expressed as $D(P_s(x), P_t(x))$. The kernel function is similar to the kernel functions used in support vector machines such as polynomial kernel, and Gaussian kernel. Specifically, linear kernel is $k(x_i, x_j) =< x_i, x_j >$, where $< \cdot, \cdot >$ is the inner product operation. Polynomial kernel is $k(x_i, x_j) =< x_i, x_j >^d$, where d is the order of polynomial. Gaussian kernel is $k(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$. Note that how to select the best kernel remains an open problem, but a simple idea is to employ a fusion of multiple kernels to compute a combined value.

Now we present an MMD-based domain adaptation example. Recall that the input-output joint distribution $P_s(x, y) \neq P_t(x, y)$ in domain adaptation, the general distribution divergence D includes both the marginal and conditional distribution divergence as follows [105]:

$$D(\mathcal{D}_s, \mathcal{D}_t) \cong (1 - \mu)D(P_s(x), P_t(x)) + \mu D(P_s(y|x), P_t(y|x)), \tag{2}$$

where μ is a balancing factor. From Eq. (2), we can see that it requires the conditional distribution $P_t(y|x)$ on the target domain. According to the Bayes' theorem, $P(y|x) = \frac{P(y)P(x|y)}{P(x)}$. As $P(x)$ is fixed and if we can ignore $P(y)$, we can use $P(x|y)$ to estimate $P(y|x)$. Sufficient statistics [106] points out that when dealing with a large number of samples, it becomes feasible to select part of statistical factors to serve as approximations for our objective. Still, we do not have the y value. In order to address this issue, we can employ an iterative training approach as follows: (i) we utilize the labeled data (x_s, y_s) to train a classifier and obtain pseudo-labels $\hat{y}_t$ for the unlabeled x_t in the target domain; (ii) these pseudo-labels are used for learning optimization, and after feature transformation, they can be updated or refined in subsequent iterations. For the MMD-based domain adaptation, we have the following distance function:

$$MMD(P_s(y|x), P_t(y|x)) = \sum_{c=1}^{C} || \frac{1}{N_s^c} \sum_{i=1}^{N_s^c} A^T x_i - \frac{1}{N_t^c} \sum_{j=1}^{N_t^c} A^T x_j ||^2, \tag{3}$$

where A denotes the feature transformation matrix, which is the learning objective. N_s^c and N_t^c are the number of samples for the c-th class in source and target domains, and C is number of classes. To solve Eq. (3), we refer readers to the statistical feature transformation methods in [100]. Moreover, deep neural networks (DNNs) have been demonstrated to be effective in implementing feature extraction to estimate the transformation matrix A. For example, DeepClustering [107] and Facer [103] employ a similar approach, utilizing DNN to extract features and applying pseudo labels to optimize their respective objective functions.

2.3.2 Explicit Distance Measurement

Many functions can be used to measure the distance between data points, data vectors, or data distributions. First, for data points, the Minkowski distance is the p-norm distance $d = (||x_1 - x_2||^p)^{1/p}$. When $p = 1$, it represents the Manhattan distance, and when $p = 2$, it stands for the Euclidean distance. Besides, the Mahalanobis distance $d = \sqrt{(x_1 - x_2)^T \sum^{-1} (x_1 - x_2)}$, where $\sum$ is the covariance of a distribution. For data sets, we can use the Jaccard index to measure the similarity between sets X and Y as $J = \frac{X \cap Y}{X \cup Y}$. For example, the concept of the intersection of union (IoU) in object detection and image segmentation tasks follows the definition of the Jaccard index.

Second, for data vectors, cosine similarity is widely used to measure the correlation, which is represented as $cos(x_1, x_2) = \frac{x_1 \cdot x_2}{|x_1| \cdot |x_2|}$. $cos(x_1, x_2) = -1$ means that the vectors are opposite, and 0 represents orthogonal vectors, and 1 signifies similar vectors. The Pearson correlation can also measure the similarity, which is defined as $\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, where $Cov(\cdot, \cdot)$ is the covariance and σ is the standard deviation.

Third, for data distributions, we use the mutual information to measure the distance between two distributions X and Y as follows:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}. \tag{4}$$

Another widely used method is the Killback-Leibler (KL) divergence, which measures the divergence between two probability distributions $P(x)$ and $Q(x)$ as follows:

$$D_{KL}(P||Q) = \sum_{i=1} P(x) log \frac{P(x)}{Q(x)}. \tag{5}$$

To achieve symmetry in divergence measurement, Jensen-Shannon (JS) divergence is defined as

$$JSD(P||Q) = \frac{1}{2} D_{KL}(P||R) + \frac{1}{2} D_{KL}(Q||R), \tag{6}$$

where $R = \frac{1}{2}(P + Q)$. JS divergence is often preferred for complicated tasks because its normalized outcomes provide better stability during loss function optimization. The Wasserstein distance (WD) is utilized to quantify the distance between the distributions P and Q, defined as:

$$WD(P||Q) = inf_{\pi \in \Gamma(P,Q)} \int_{\mathcal{R} \times \mathcal{R}} |x - y| d\pi(x, y), \tag{7}$$

where $\Gamma(P, Q)$ is the set of distributions on $\int_{\mathcal{R} \times \mathcal{R}}$ whose marginals are P and Q. The WD can be interpreted as the smallest amount of "effort" needed for the transformation from P to Q.

2.3.3 Implicit Distance Measurement

The Generative Adversarial Networks (GANs) were originally designed for generative tasks. The basic idea is to train two neural networks simultaneously: a generator and a discriminator, which are engaged in a game-like scenario, hence the name "adversarial". The generator network takes random noise as input and produces data samples that are intended to resemble the real data distribution. Formally, GAN can be regarded as a nonlinear parametric mapping $g : Z \to X$, where Z represents feature space and X represents the data space. Ideally, we can learn the latent representation z in feature space $z = H(x)$, where $z \in Z$, so that we can map the latent representation back to the input space as $G(z) = x$. However, in practice, G can only learn an approximation to the true data manifold, which causes the approximation error.

Essentially, GAN constantly improves its ability to generate convincing samples by minimizing the loss function $L_{gen} = -log(D(G(z))$, where $G(\cdot)$ is the generator's output, and $D(\cdot)$ represents the discriminator's output. The discriminator network's role is to distinguish between real data samples and those generated by the generator. the discriminator network learns to classify data as either "real" or "fake", which naturally learns the distribution divergence. The discriminator becomes more accurate in distinguishing between genuine and generated data by minimizing the loss function $L_{disc} = log(D(x)) + log(1 - D(G(z)))$. This adversarial training dynamic is at the core of GANs and leads to insights into the similarities or differences between two data distributions. We can adopt metrics such as MMD and Wasserstein distance to assess how well a GAN has learned to approximate the target data distribution.

3 Transferable Attack from Data Perspective

From the data perspective, we review existing work that boosts the transferability of attacks, including data synthesis, data transformation, and data disentanglement as shown in Table 2.

3.1 Data Synthesis for Transferable Attacks

Data synthesis refers to the creation of new data by combining or generating information. This includes data generated by GANs, data sampled from standard distributions, and data obtained by certain optimization methods. Many black-box attacks rely on substitute models to design attacks toward the black-box system. Synthesized datasets can be used to enhance the transferability of learning-based attacks. Papernot *et al.* [108] first propose to generate a synthetic dataset to attack DNN that was trained on the MNIST dataset. Devil's whisper [33] crafted a customized speech dataset that only focuses on a small portion of commonly used phrases to fine-tune the substitute model. Krishna *et al.* [23] attempt to steal a BERT model by crafting a query dataset. There are two different settings when crafting the query dataset. First, the random generator setting uses a nonsensical sequence of words constructed by randomly sampling a Wikipedia vocabulary from WikiText-103 [109]. Second, in the Wiki setting, the input queries are formed from actual sentences or paragraphs from the WikiText-103 corpus. They found that directly using these two datasets is insufficient for extracting models. Therefore, they implemented two additional strategies for accomplishing different

Table 2: Selected work for boosting the transferability of different attacks from the data and learning-process perspective. The source dataset and model are related to training a surrogate model, and the target dataset and target model are the victims. Unless specified, the transferable type belongs to adversarial attacks.

Methodology	Publication	Source Dataset/Model (Selected)	Target Dataset (Selected)	Target Model (Selected)	Transferable Type
Data Transformation	Diversity [111]	Model ImageNet	ImageNet	Inception V3	Cross-model Attack
	Admix [112]	ImageNet	ImageNet	ResNet 101	Cross-model Attack
	Learning [113]	ImageNet	ImageNet	Inception V3	Cross-model Attack
	Copycat [19]	ImageNet	SVHN	VGG-16	Model Stealing
Data Synthesis	Whisper [33]	Songs	Songs	Commerical API	Cross-model Attack
	MAZE [18]	CIFAR-10	CIFAR-10	ResNet-20	Model Stealing
	Jiasaw [46]	Artifical Image	ImageNet	ResNet-152	Cross-model Attack
Data Disentanglement	StyleLess [34]	ResNet	ImageNet	VGG19	Cross-model Attack
	ColorBackdoor [12]	CIFAR-100	CIFAR-100	DeepSweep	Backdoor Attack
	StyleText [15]	SST-2	SST-2	BERT	Backdoor Attack
	StyleTrigger [36]	COVID	COVID	BERT+LSTM	Backdoor Attack
	Audio [114]	GSC	GSC	CNN	Backdoor Attack
	StyleFool [115]	UCF-101	UCF-101	C3D	Query-based Attack
Generative-based Universal Attack	Perturbation [44]	Painting, Cartoon Images	ImageNet	Inception V3	Cross-domain Attack
	PhoneyTalker [45]	LibriSpeech	VoxCeleb1	DeepSpeaker	Universal Example
	Object [116]	PASCAL VOC	VOC2007	FasterRCNN	Universal Example
Heuristic-based Universal Attack	Patch [40]	Predicted Hard Label	CIFAR-10,SVHN	ResNet18	Universal Patch
	Trigger [41]	GPT-2 117M	Input Text	GPT-2 345M	Cross-model Attack
	LLM [42]	Vicuna Training Dataset	Input Text	ChatGPT, Bard	Cross-model Attack
Gradient-based Universal Attack	Ranking [117]	MSMARCO DEV Dataset	Input Text	Imitate V2	Universal Trigger
	UAP [118]	ImageNet	ImageNet Val	ResNet 152	Universal Example
	Nesterov [119]	Model on ImageNet	ImageNet	Inception V3	Cross-model Attack
	Evading [120]	Model on ImageNet	ImageNet	Inception V3	Cross-model Attack
	Tuning [38]	Model on ImageNet	ImageNet	Inception V3	Cross-model Attack

tasks. For the MNLI task, they randomly replaced three words in the premise with three random words to construct the hypothesis. For the SQuAD task, they additionally prepend a question start work (such as "what") to the question and append a question symbol to the end. Finally, MAZE [18] is a data-free approach that leverages the generative model to generate the samples where there is a significant disagreement between the attacker and the target model. A Wasserstein GAN is trained to generate these samples, and the effectiveness of the attack is demonstrated on various target models. The results show that MAZE achieves an accuracy of 89.85% on the CIFAR-10 dataset, in comparison to the target model's 92.26% accuracy. However, MAZE still uses softmax outputs of the victim model to generate input-output pairs. Later, Sanyal *et al.* [110] propose a new GAN-based framework in hard-label settings to generate synthetic data and steal models. Employing synthesized datasets to develop a surrogate model can serve as an intermediary step in converting black-box systems into white-box systems, thereby facilitating attacks on more complex black-box systems.

3.2 Data Transformation for Transferable Attack

Data transformation refers to the process of converting data from one format or structure into another. For example, data transformation in the image domain includes scaling, rotating, translating, and flipping, filtering. Dong *et al.* [121] propose MI-FGSM, which is designed with random and differentiable transformations to the input images. Xie *et al.* [111] propose to use image transformation such as resizing and padding to improve transferability, while Dong *et al.* [120] consider optimizing a perturbation over an ensemble of translated images to enhance transferability. Lin *et al.* [119] propose to compute gradients using scaled benign samples, and Admix [38] derives iterative gradients by blending benign images with randomly selected

sample images. Different from the pixel-wise attack, Gao *et al.* [122] propose a patch-wise adversarial attack on unknown target models. Wang *et al.* [123] propose to utilize the feature importance computed on a batch of transforms of raw image to guide the search for adverarial examples. Zhang *et al.* [124] propose a feature-level attack by contaminating the intermediate feature outputs of a surrogate model with neuron importance estimations, so as to enhance the transferability. Finally, Fan *et al.* [113] propose to boost the transferability of adversarial attacks from both data augmentation and model augmentation perspectives. They use random resizing, and padding for data augmentation, and modify backpropagation for model augmentation to enhance the transferability across varied tasks.

Data transformation also includes generating labels for existing data samples [21]. In a query process, the adversary feeds the existing data and collects outputs of the target model to be used as labels. With the data and labels, the adversary can train a surrogate model to approach the behaviors of target models. The existing data comprise both original datasets and datasets from similar domains. The original dataset refers to the training dataset used for training the target model. Previous work [125] demonstrated that if the adversary uses the same training data as the target training data, the surrogate model can achieve $> 97\%$ accuracy as the target model. The similar domain dataset represents the dataset related to the original training data. For example, if the target model is trained on ImageNet for image classification, similar domain data can be CIFAR-10. Jagielski *et al.* [126] propose to use 10% and 100% ImageNet as training data, and then query the output of a large-scale model that was trained on 1 billion Instagram images [127]. Then, the data and queried labels are used to train a classifier based on ResNet. Their results demonstrate that the adversary can extract a target model using merely 10% training data. Overall, the data transformation is an effective way to enhance the transferability of learning-based attacks.

3.3 Data Disentanglement for Transferable Attack

Data can reflect both the content and style of things in the physical world. Many deep learning-based tasks essentially rely on the content (task-relevant) features rather than the style (task-irrelevant) features. In this section, we investigate the recent studies that attempt to improve the transferability of adversarial and poisoned examples by disentangling style and content in data.

3.3.1 Disentangling Style and Content in Image

The style refers to the unique characteristics, or aesthetic qualities used to create an image. For example, the style can include different color schemes, brush strokes, or filters. It has been established that the style of an image can be characterized by examining the means and correlations among various feature maps in deep neural networks (DNNs) [128]. The style information can be encapsulated in a Gram matrix as follows:

$$G^l_{cd} = \frac{\sum_{ij} F^l_{ijc}(x) F^l_{ijd}(x)}{IJ}, \tag{8}$$

where I denotes the number of feature maps at layer l and J denotes the length of the vectorized feature map. The content refers to the primary subject or theme of an image that is visually recognized such as the people, things, and places. The content of an image is represented by the values present in the intermediate feature maps in DNNs.

To improve the transferability of adversarial attacks, Liang *et al.* [34] propose StyLess to achieve style-less perturbation. The basic idea involves the disentanglement of the image style and content, while a robust DNN would place greater emphasis on content features rather than style features. Specifically, the authors create a stylized model by adding an adaptive instance normalization (AdaIN) layer [129] to the vanilla model. The AdaIN layer serves to align the mean and variance of the style features with those of the content features. Additionally, they use a graph model to represent style transfer, where vertices represent

content and style in images [130], and the style transfer process involves message passing between these nodes. Compared to stylized models, the vanilla model's adversarial loss increases more rapidly, leading to a greater loss disparity. By utilizing gradients from both stylized surrogate models and the vanilla model, adversarial examples are refined and updated. This strategy significantly enhances transferability, enabling adversarial examples to deceive a broader range of classification models.

Many backdoor attacks [131, 132, 133, 134, 12, 16, 135] have been proposed to enhance the imperceptibility and transferability of triggers. For example, Zhao *et al.* [131] study the transferability of poisoning attacks across different machine learning models, including SVM and LR. They have developed a Projected Gradient Ascent (PGA) algorithm to implement label contamination attacks on a range of empirical risk minimization models. Certain approaches incorporate distinct image styles, such as reflections [133], Instagram filters [132], weather conditions [134], and color variations [12], as backdoor triggers. For instance, Jiang *et al.* [12] present a color backdoor attack. Specifically, they apply a uniform color space shift to all pixels as the trigger. To identify the optimal trigger, they employ the Particle Swarm Optimization (PSO) algorithm. Their approach involves assessing the effectiveness of a trigger by employing the backdoor loss of a semi-trained model with a surrogate model architecture. Additionally, they incorporate a penalty function that enforces naturalness constraints during the PSO search process. Such trigger is robust to image transformation and preprocessing-based defenses [136, 137, 138]. This method enhances the transferability of backdoor triggers, allowing the poisoned data to be easily transferred to potential training datasets.

3.3.2 Disentangling Style and Content in Text

The text style is defined as the common patterns of lexical choice and syntactic constructions that are independent of semantics [139]. Text style is a feature that is irrelevant to most natural language processing tasks. Text style transfer refers to altering the style of a sentence while preserving its semantics. Many text-style transfer methods have been proposed. For example, style transfer via paraphrasing (STRAP) [35] is an unsupervised text style transfer model based on controlled paraphrase generation.

A number of studies have demonstrated that adversarial perturbation in text can mislead the DNN classification results. For instance, making subtle alterations to harmful phrases can effectively deceive Google's toxic comment detection systems [140]. To generate adversarial examples, the prevailing strategy involves altering the aspects of the data that are irrelevant to the task at hand while preserving features that are pertinent to the task. For example, to attack a sentiment analysis model, an adversarial attacker alters the syntax (task-irrelevant) but preserves the sentiment (task-relevant) of the original samples [141]. Qi *et al.* [15] propose to implement textual adversarial attacks and backdoor attacks based on style transfer. Specifically, for adversarial attacks, their approach involves the transformation of original inputs into multiple text styles as a means of crafting adversarial examples. For backdoor attacks, certain training samples are converted into a chosen trigger style, and these transformed samples are introduced into the victim model during training, effectively implanting a backdoor.

Later, Pan *et al.* [36] present the Linguistic Style-Motivated backdoor attack (LISM), a unique backdoor that leverages text style transfer models to paraphrase a base sentence in a secret attacker-specified linguistic style (trigger style), which links the text style transfer to hidden textual triggers. Unlike word-based triggers, LISM focuses on generating sentences with attacker-specified linguistic styles, preserving the semantics of the base sentence while minimizing abnormality. Such a style-based trigger scheme dynamically and independently paraphrases each base sentence to create semantic-preserving triggers, achieving advanced design goals for hidden triggers. The approach has been shown to pose considerable challenges for detection algorithms and defensive techniques, emphasizing the stealthiness of the poison attack.

3.3.3 Disentangling Style and Content in Audio and Video

The style of audio refers to the distinctive characteristics of a sound such as tone, dynamics, and tempo. For example, different styles can be created by combining effects such as PitchShift, Distortion, Chorus, Reverb, Gain, Ladderfilter, and Phaser [114]. For the backdoor attack, the poisoned samples incorporate specific patterns that lead the victim model to associate them with the target class. For example, Koffas *et al.* [114] use the stylistic properties to define the pattern, so as to poison the training data. They poison up to 1% of the training data and investigate two backdoor settings: clean-label and dirty-label attacks. However, it is worth noting that style transfer can degrade the sample quality. In their experiments, the backdoored model has only a slight drop in clean data accuracy. Their results show that the dirty-label attack achieves better performance across all styles, while the clean-label attack is effective primarily with certain styles.

Similar to the style of images, the style of videos refers to the distinctive visual, auditory, and thematic elements that define their appearance, sound, and narrative. For example, many factors such as color grading, camera movement, lighting, and sound effects can influence the style of a video. To craft adversarial perturbations for videos, Cao *et al.* [115] propose StyleFool, which is a black-box adversarial attack framework designed for video classification systems. StyleFool tries to change the non-semantic information while avoiding confusing human understanding of video content and misleading the classifier. Specifically, in the StyleFool framework, a best-style image is first selected from an image set. Then, the input video is transformed into the style of the selected image by minimizing content loss, style loss, total variation loss, and temporal loss. Finally, the natural evolution strategy is applied to generate adversarial perturbations in a black box setting. The advantage of the adversarial style transfer in StyleFool is the ability to initialize the perturbations, which steers the stylized video closer to the decision boundary. This initialization reduces the number of queries needed for generating adversarial samples.

3.4 Summary

In this section, we have discussed the transferable attacks from the data perspective. We have summarized various data synthesis and transformation methods to boost the transferability of learning-based attacks during the optimization process. Moreover, we review the existing efforts to enhance the transferability of attack via disentangling style and content in data, including image, text, audio, and video. The data-centric methods focus on improving and optimizing the quality of data used in designing transferable attacks. Such augmented data can potentially exploit a wider range of vulnerabilities across different models.

4 Transferable Attack from Process Perspective

From the learning process perspective, we review different learning methods such as gradient-based, and heuristic-based optimization for enhancing the transferability of adversarial attacks as shown in Table 2. The preliminary knowledge of optimization methods is introduced in Section 2.2.

4.1 Gradient-based Learning Process for Transferable Attack

Universal adversarial attacks [142, 2, 43, 121, 143, 144, 145, 146, 40, 147, 148] compromise different black-box models in a universal manner. Transferability has been observed across various data types, architectures, and prediction tasks [149]. Fredrikson *et al.* [24] develop a new class of model inversion attacks that exploits confidence values revealed along with predictions, and discuss the transferability of these methods in decision tree models. Papernot *et al.* [150] explore the transferability of adversarial perturbations across different machine learning models, including SVMs and decision trees. They introduce adversarial sample crafting techniques for non-differentiable machine learning models by generalizing the learning of substitute

models. Their results show that SVM and decision tree classifiers respectively misclassify 91.43% and 87.42% of adversarial samples crafted for a logistic regression model. Demontis *et al.* [151] propose a unified optimization framework for evasion and poisoning attacks. To test transferability, they assess the attack across a range of classifiers, including logistic regression, SVMs, ridge regression, random forests, and others, with varying complexity via different hyperparameters.

The transferability of adversarial examples lies in the observation that various machine learning models tend to learn similar decision boundaries around a given data point [121]. The gradient-based methods are widely applied to obtain the optimized universal adversarial perturbations or patches [43, 152, 153, 117]. For example, Moosavi *et al.* [43] design universal adversarial perturbations capable of deceiving a model across a majority of source dataset images. Dong *et al.* [121] propose the one-step gradient-based approach, exemplified by the fast gradient sign method (FGSM), which excels in crafting adversarial examples that exhibit enhanced transferability. Lin *et al.* [119] introduce two approaches, the Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack methods (SIM), to enhance the transferability of adversarial examples by regarding adversarial example generation as an optimization process. NI-FGSM incorporates the Nesterov accelerated gradient in iterative attacks, while SIM exploits the scale-invariant property of deep models to optimize adversarial perturbations, ensuring more effective and transferable attacks against defense models. Later, Dong *et al.* [120] present a translation-invariant attack method for creating highly transferable adversarial examples. By optimizing perturbations across translated images and utilizing a predefined kernel for gradient convolution at the untranslated image, the method ensures improved transferability and attack efficiency. Testing on ImageNet showcased an average success rate of 82% against eight leading defenses, highlighting the vulnerabilities in current defensive strategies.

More recently, Wang *et al.* [38] introduce variance tuning to enhance the transferability of iterative gradient-based attack methods. By incorporating the previous iteration's gradient variance into current gradient calculations, the method stabilizes update directions and avoids poor local optima. Qin *et al.* [154] introduce a novel technique known as reverse adversarial perturbation (RAP). The primary objective of RAP is to identify an adversarial example situated in a region characterized by uniformly low loss values, achieved by integrating reverse adversarial perturbations at each stage of the optimization process. The process of generating RAP is framed as a min-max bi-level optimization problem. By incorporating RAP into the iterative attack procedure, their proposed methods can discover more transferable adversarial examples. Liu *et al.* [117] then introduce an imitation adversarial attack targeting text ranking models in search engines. To facilitate a highly-transferable attack, adversaries have devised a gradient-based approach to generate adversarial examples. By leveraging a pairwise anchor-based trigger, attackers can manipulate malicious texts with only a few tokens. TransAudio [155] is another attack approach involving word-level adversarial attacks, including word insertion, deletion, and substitution. To enhance the transferability of these attacks, TransAudio incorporates an audio score-matching mechanism into the surrogate model. This feature successfully mitigates overfitting issues associated with certain surrogate models.

4.2 Heuristic-based Learning Process for Transferable Attack

The heuristic-based methods [41, 40, 156, 157, 158] are selected in scenarios where gradient-based optimization methods are not applicable. Particularly, these heuristic-based search methods are often employed in conjunction with gradient information. For example, Tao *et al.* [40] develop a universal adversarial patch attack, specifically tailored for the challenging hard-label black-box setting. In such a setting, the attack is limited to accessing predicted labels without confidence scores. The proposed method utilizes historical data points during the search for an optimal patch trigger. The Markov Chain Monte Carlo (MCMC) methods and Genetic Algorithm (GA) are utilized during the search process. Then, they design a gradient estimation mechanism to induce misclassification of multiple samples, making it particularly effective for hard-label black-box scenarios.

An approach for exploiting transferability in text universal adversarial examples has been introduced in [41]. The universal adversarial triggers are input-agnostic sequences of tokens that cause models to generate specific unintended predictions when appended to any input. Through a gradient-guided search, short trigger sequences have been identified. Specifically, the search process iteratively modifies the token within the trigger sequence, aiming to boost the probability of the desired target prediction for batches of examples. Zhou *et al.* [42] conduct research on universal and transferable adversarial attacks targeting aligned language models to induce objectionable behaviors. Their approach involves identifying a suffix that, when appended to a diverse set of queries, causes an LLM to generate objectionable content, including affirmative responses. These adversarial suffixes are generated through a combination of greedy and gradient-based search techniques. One noteworthy aspect of these attacks is their transferability, meaning that the adversarial prompts created in this manner can be effective even against black-box LLMs, thereby raising concerns about the potential for widespread misuse and the need for robust defenses against such universal adversarial triggers in Natural Language Processing (NLP) systems.

There are many adversarial attacks on the graph neural network model [159, 9, 55]. We take attacks on node classification as an example. The goal of an adversarial attack on node classification is to misclassify node labels. For example, Ma *et al.* [4] investigated black-box adversarial attacks for node classification tasks. The attack has two main steps. First, attackers select a subset of nodes, and their targeting is confined to a small number of nodes. Second, the attackers modify either the attributes of nodes or the edges within the specified budget for each node. In this work [4], they mainly focus on the node selection strategy to enhance a black-box adversarial attack. Specifically, they extend the gradient norm in a white-box attack to a model-agnostic importance score by leveraging the relationship between backward propagation in GNNs and random walks. The strong structural inductive biases of the GNN model are utilized as an important information source for designing the attack. However, attacks based on the importance score significantly increased classification loss without markedly impacting the misclassification rate, indicating a gap between these two metrics. Therefore, the authors introduce a greedy correction procedure for the importance score to account for the diminishing return pattern. Similarly, Zhou *et al.* [8] propose to leverage a saliency map to generate hierarchical adversarial examples, focusing on identifying and minimally perturbing critical feature elements. They then develop a hierarchical node selection algorithm based on a random walk with restart. Overall, the heuristic-based method can help find adversarial examples with better transferability.

4.3 Generative-based Learning Process for Transferable Attack

Instead of focusing on individual samples using generative adversarial training, researchers have studied universal adversarial perturbations within a data distribution [160, 44, 161, 45]. For instance, to enhance transferability and efficiency in adversarial attacks, Wei *et al.* [116] introduce an adversarial example creation technique based on a GAN framework. Rather than directly manipulating the video, the approach modifies the neural network's extracted feature output. To augment transferability, the attackers integrate both high-level class loss and low-level feature loss to produce more universally effective adversarial examples. Furthermore, this attack strategy saves time in processing video samples. Naseer *et al.* [44] explore domain-invariant adversarial examples as a means to achieve universal perturbations in images. The core of their adversarial function lies in a generative network as introduced in Section 2.3.3, trained using a relativistic supervisory signal. This approach ensures the generation of domain-invariant perturbations. Their framework comprises a generator, responsible for generating unbounded perturbations, and a discriminator, which assesses class probabilities. The results have unveiled a common adversarial space that transcends different datasets and models. Recently, Hu *et al.* [162] propose to use adversarial makeup transfer GAN (AMT-GAN) to generate adversarial face images, so as to protect photos from being identified by unauthorized face recognition systems.

PhoneyTalker [45] is designed as an accessible toolkit to generate universal and transferable adversarial

audio adversarial examples. Specifically, PhoneyTalk employs a unique approach by generating universal adversarial perturbations at the phoneme level, facilitated by a generative model. First, they break down the voice signals into phoneme combinations using a forced alignment technique. Subsequently, a generative model is employed to acquire input-independent perturbations at the phone level. Additionally, they incorporate a series of digital signal processing methods to minimize the audibility of perturbations. Furthermore, several strategies are proposed to enhance the generalization and transferable abilities. First, a vast and diverse corpus is used to enrich input diversity. Second, multiple speaker recognition models as substitute classifiers for the target model. Third, a novel loss function with a confidence margin is used to train the adversarial perturbation generator. This model has the capability to map any voice to the desired target user as specified by the adversary. The core of PhoneyTalk revolves around training a generative model on a diverse and extensive corpus. By enhancing the model's generalization capabilities, the devised attack can be successfully transferred from local substitute models to previously unseen target models.

4.4 Summary

In this section, we have provided a comprehensive overview of the learning process associated with transferable attacks. Rather than offering a high-level and abstract review of related work, we have used concrete research examples to illustrate different learning-based attack methodologies. Our exploration encompasses three primary methodologies: gradient-based optimization, heuristic-based optimization, and generative methods in designing universal adversarial attacks. We hope the existing methodologies can inspire more practical design strategies for evaluating the safety and security of cyber-physical AI systems.

5 Transferable Attack from Model Perspective

From the model perspective, the transferable attacks need to be generalized to various target models. Many efforts such as ensembling, model pretraining, and surrogate model optimization have been devoted to boosting the transferability of attacks.

5.1 Model Ensembling for Transferable Attack

The learning-based attacks tend to overfit to a single surrogate model. Take the adversarial attack as an example, when targeting a black-box system, a common approach is to utilize a surrogate model to replace the black-box model during the generation of adversarial perturbations [150]. However, existing research [47, 151, 163] shows that the generated adversarial perturbations tend to overfit to the surrogate model, which limits the capability to effectively launch transfer attacks on various target models. The transferability of attacks depends on the intrinsic adversarial vulnerability of the target model and the complexity and diversity of the surrogate model [151].

Ensemble-based methods [2, 47, 163] are proposed to enhance the complexity of surrogate models. The idea of an ensemble method is to combine multiple models to improve accuracy, reliability, and performance. Table 3 shows a list of representative research for improving the transferability of attacks. For example, Liu *et al.* [47] devise an ensemble-based approach for crafting adversarial examples targeted at multiple models. The basic idea of ensemble learning is to combine the predictions of multiple models to solve a computational intelligence problem. Formally, suppose we have k white-box surrogate models with softmax outputs denoted as $J_1, \cdots, J_k$, when dealing with a targeted attack where y^* represents the target label, along with an original input x and its true label y, the ensemble-based method addresses the optimization problem as follows:

$$\arg\min_{x^*} -log((\sum_{i=1}^{k} \alpha_i J_i(x^*)) \cdot \mathbf{1}_{y^*}) + \lambda d(x, x^*), \tag{9}$$

Table 3: The representative work for enhancing the transferability of attacks.

Methodology	Publication	Source Model (Selected)	Target Model (Selected)	Design Keyword
Ensembling-based Adversarial Attack	Liu [47]	ResNet-152	VGG-16	Model-level
	Li [163]	ResNet-50 Ghost	ResNet-152	Feature-level
	Wang [38]	Inception V3	ResNet-101	Gradient Variance
	Qin [154]	VGG-16	Inc-Res-v2	Reverse Perturbation
Pretraining-based Foundation Model	Zhou [52]	CLIP	CLIP-ResNet	Downstream-agnostic Adversarial Patch
	Jia [54]	CLIP	DNN Classifier	Backdoor Attack
	Liu [11]	ResNet18+SimCLR	DNN Classifiier	Data Poisoning
	Liu [28]	None	CLIP	Membership Inference
	Liu [22]	Self-defined Encoder	CLIP	Model Stealing

where $\sum_{i=1}^{k} \alpha_i J_i(x^*)$ is the ensemble model and α_i is the weights. They find that the non-targeted adversarial attacks that mislead a model to arbitrary wrong labels have greater transferability. Nevertheless, the proposed ensemble-based method can boost the transferability of targeted adversarial examples. However, acquiring a diverse family of models for ensemble-based methods is computationally expensive. Thus, Li *et al.* [163] propose Ghost Networks to boost the transferability of adversarial examples. The core idea is to introduce feature-level perturbations to an existing model, so as to generate an array of diverse models. The feature-level perturbation involves densely applying dropout layers to each block within the foundational network. Besides, perturbations are introduced to the skip connection layers to create multiple ghost networks. The experimental findings substantiate the significance of the number of networks in enhancing the transferability of adversarial examples. More recently, Long *et al.* [164] propose the frequency domain model augmentation by applying a spectrum transformation to the input.

5.2 Model Pretraining for Transferable Attack

A pre-trained large encoder model is trained on a substantial volume of unlabeled data based on self-supervised learning, enabling downstream users to focus on fine-tuning tasks [51]. Many efforts have been devoted to investigating attacks towards pertained encoders, such as the adversarial attack [52], backdoor attack [53, 54], poisoning attack [11], and membership inference attacks [165, 28]. The attacks on the pre-trained models can transfer to the downstream fine-tuned models, thereby introducing security vulnerabilities in these downstream tasks. Chen *et al.* [166] propose the data poisoning attack on the cross-domain recommendation system, where people try to improve predictive accuracy in the target domain by transferring knowledge from the source domain with a pretrained model. For the pre-trained NLP models, Chen *et al.* [167] introduce task-agnostic backdoor attacks, which are not constrained by prior knowledge of downstream tasks. The key idea involves implanting a backdoor in the pre-trained model, which, when inherited by downstream models, can be triggered without prior information about the specific task.

To explore security issues of pre-trained encoders, AdvEncoder [52] is designed to generate universal adversarial examples that are agnostic to downstream tasks, leveraging the inherent weaknesses of the pre-trained encoder. The aim is to create perturbations or patches capable of deceiving all downstream tasks relying on the susceptible pre-trained encoder. The primary challenge is that attackers have no prior knowledge of the downstream tasks, including the specific attack type, the pre-trained dataset, and the downstream dataset, or whether the entire model is subject to fine-tuning. This lack of information complicates the crafting of effective adversarial perturbations. Therefore, a generative attack framework is proposed to construct adversarial perturbations. Notably, the authors find that altering the texture information such as the high-frequency components (HFC) of an image is likely to influence model decisions. Consequently,

Table 4: The selected work for surrogate model optimization with model stealing strategies.

Paper	Target model	Backbone model	Original data	Training data	Num. queries
[168]	MLP, SVM, Decision Tree	RBF kernel	Adult dataset	Adult dataset	11k for MLP
[49]	Naive Bayes, SVM	DNN	Reuters-21578	Reuters-21578	Not reported
[23]	BERT	BERT	Unknown	craft data	9k∼392k
[21]	ResNet-34	ResNet, VGG	Caltech-256, BUBS-200, Indoor-Scenes	OpenImages, Crafted dataset	10k
[125]	LSTM	RNN	MINST	MINST	Not reported
[33]	Speech recognition APIs	Kaldi ASpIRE Chain	Unknown	craft phrases	1.5k
[126]	WSL ImageNet	ResNet	1 billion Instagram images	ImageNet	4k
[19]	VGG-16, AlexNet	VGG, AlexNet	UCF101,MINIST,etc.	original data, similar domain data	100k
[18]	LeNet, ResNet-20	WideResNet	FashionMNIST	craft data	100k
[22]	Encoders	ResNet, Vgg, DenseNet, MobileNet, ShuffleNet	Unknown	STL10, Food101, CIFAR10	2.5k∼25k

they propose using a high-frequency component filter to extract HFC from benign and adversarial samples and maximize the Euclidean distance between them to influence the model's decision.

5.3 Model Optimization for Transferable Attack

Transferability is highly impacted by the alignment between the surrogate model and the target model [151]. The model stealing attacks [168, 125, 33, 19, 169, 22, 170, 171] aim to build the surrogate model, which aligns with the target model as much as possible. The basic idea of model stealing is to retrain a substitute model by querying the target model. The architecture of the substitute model can be the same or different from the target model. Table 4 characterizes the transferable model stealing attacks according to the target model, backbone model, and number of queries. An adversary aims to steal the model by obtaining the training parameters, architecture, and learned weights, and re-produce the target model.

Formally, for target model f, the attacker builds a model $\hat{f}$, which can solve the same task of f and produce the consistent result as f. The coincident prediction can be represented as $f(x) = \hat{f}(x)$, which means if there exists a perturbation θ that satisfies $f(x+\theta) = y_t$, the perturbation will work on the stolen model $\hat{f}$. The attackers are assumed to only have black-box access to the target model, indicating that the attacker can only utilize the output (*e.g.*, confidence scores, score distribution, or hard label) of the model. For the training data of the target model, the attackers are assumed to have different levels of access depending on the attack types.

In the transferable attack context, the core idea of the substitute model is to transfer a base backbone model to the target model. The fine-tuning process mainly relies on the query output of the target model, aiming to drive the transferred model to produce consistent results of the target model. Typically, the first step of the model stealing is to choose a backbone model architecture to serve as the initial substitute model. By fine-tuning the backbone model parameters using the query output, the backbone model closely mimics the behaviors of the target model, ultimately generating output that is consistent with that of the target model. For example, the first transferable model stealing attack is proposed by [168], and a substitute model is trained by leveraging the Radial Basis Function (RBF) kernel to steal different models such as neural networks, decision trees, and logistic regression. They assume the attacker knows the soft labels, and the attacker does not obtain the training data of the target model. Later, Takemura *et al.* [125] launch the model stealing attack using Recurrent Neural Networks (RNNs) to steal an LSTM network and achieve an accuracy of 97.3% relative to the target model. Devil's whisper [33] steals the commercial speech recognition model by configuring a KaldiAspire Chain (LSTM-based) model as the backbone model; the Knockoff nets [21] steals an image classification model by using ResNet or VGG as the backbone model. Krishna *et al.* [23] propose to steal a well-trained BERT model by selecting the same BERT architecture as a substitute model. In StolenEncoder [22], the authors have stolen multiple encoders including those trained by themselves

and pre-trained by Google, OpenAI, and Clarifai. In their experiments, they use a ResNet34 model as a backbone model to steal different encoders such as ResNet34, Vgg19, ResNet18, DenseNet121, MobileNet, and ShuffleNet. They find that if the backbone model has a similar architecture as the target model, the stolen model could achieve better accuracy.

To steal a proprietary model, previous attacks typically assume access to the complete probability predictions of the victim model. However, in real-world scenarios, APIs often return only the top-k predictions or the top-1 prediction. In response to this limitation, Wang *et al.* [172] introduce an approach called "black-box dissector". Specifically, they propose the class activation map-driven erasing strategy, which aims to maximize the information capacity hidden within the hard labels generated by the victim model. It achieves this by erasing regions of importance on images based on the top-1 class activation maps obtained from a substitute model. After erasing these important areas, the system queries the victim model for a new prediction. If the new prediction changes, it indicates that the substitute model was focusing on the wrong areas of the image. This process helps align the attention of the substitute and the victim model by learning from clean samples and their corresponding erased counterparts. Moreover, Beetham *et al.* [173] propose a method to explore the transferability between different models by stealing a target model without the need for access to either the training data or the target model itself. This approach, known as data-free model stealing, distinguishes itself from traditional black-box model stealing methods, which typically require real-world data samples to train a student network. In data-free model stealing, a generator is employed to produce synthetic samples for training a student model to align its outputs with those of the target model. The key innovation lies in training two student models that provide the generator with criteria to generate samples on which the two students disagree. This strategy enables the stolen model to be used for downstream adversarial attacks. One student model is optimized to minimize the discrepancy between its outputs and those of the target model. This approach facilitates the transfer of knowledge between models, allowing for the potential use of the stolen model in various adversarial scenarios.

5.4 Summary

In this section, we have reviewed the transferability-related research from the model perspective. To enhance the attack transferability, it is crucial to maximize the similarity between the surrogate and target models. Drawing from ensemble learning principles, employing diverse model architectures concurrently can ensure that the devised attacks are compatible with various potential models. Moreover, attacking the pre-trained foundational models can influence downstream fine-tuned models. Finally, we have examined different strategies for optimizing the surrogate model in the context of model-stealing attacks, enhancing its resemblance to the target model. In this way, we can enhance the transferable attack success rate.

6 Transferable Attack from System Perspective

In this section, we focus on practical black-box attacks toward real-world cyber-physical AI systems. The transferable attacks manipulate systems designed for real-world applications, highlighting the critical need to address security concerns beyond the digital realm.

6.1 Transferable Attack on Computer Vision System

Transferable attacks in the CV domain have brought new security issues, especially in crucial domains such as autonomous driving, drone cruise, surveillance, and face classification. Attacking object detection systems poses a unique challenge because the adversary must not only determine the presence of an object but also mislead the system in its label prediction, all within dynamic environments.

6.1.1 Hiding and Appearing Attack

The security issues of object detection have been widely studied. Song *et al.* [174] propose to generate physically adversarial perturbation on objects, forcing the object detection model to disregard the object (hiding attack) or to misclassify the object (appearing attack). They extend the physical perturbations in [175] that are sampled from a distribution to mimic the physical perturbations of an object. Then, they design two loss functions to generate physical adversarial perturbations. The loss function of hiding attacks is a function that deducts the probability of an object class from a tensor. The adversary can directly minimize that probability until it falls below the detection threshold of the model. On the other hand, the loss function of appearing attacks is a function that computes the probability of confidence that the box contains any object. They configure the threshold of the box confidence and class probability to derive the optimized attack pattern. To further improve the attack success rate with long distances, wide angles, and various real scenarios, Zhao *et al.* [176] propose a feature-interference reinforcement method and enhanced realistic constraints generation model to generate robust adversarial examples against object detectors. The proposed techniques leverage the manipulation of the hidden layers in DNN and the semantics of the target object. The designed adversarial example combines two distinct adversarial perturbations, each specifically targeting a different distance-related sub-task of the attack. They train the adversarial examples on YOLO V3 and faster-RCNN, which have been demonstrated to be highly transferable to four black-box models.

Table 5: The representative work for transferable attacks in cyber-physical systems.

System	Attack	Publication	Attack Vector	Attack Target
Computer Vision	Hiding and Appearing	[174]	Adversarial Sticker	Object Detection
		[176]	Adversarial Example	Object Detection
	Physical Patch	[177]	3D Mesh	Visual Recognition
		[178] [179]	Adversarial Patch	Object Detection
		[59]	Adversarial T-shirt	Object Detection
		[180] [181] [182]	Transformation Pattern	Object Dection
		[183]	3D-printed Object	Autonomous Driving
	Projected Patch	[184]	Phantom from Projector	Objection Detection
		[185]	Len flare Auto-exposure Control	Image Classification
		[186][187]	Projector	Object Detection
		[188]	Adversarial Patch	Depth Estimation
	Side Channel	[189]	Camera	Keystroke Inference Attack
		[190] [191]	Acoustic	Object Detection
		[192]	Electromagnetic	Visual Recognition
Smart Audio	Data Preprocessing	[153]	Processing pipelines	Speech/Speaker Recognition
	Speech Recognition	[193]	Sensors	Eavesdropping Attack
	Speaker Recognition	[152]	Psychoacoustic	Adversarial Attack
		[194]	Live Stream Perturbation	Speaker Verification
		[195]	Phone Powerline	Speaker Verification
Large Language Model System	NLP Tasks	[196]	Encoding Injection	Commerical Text Systems
		[197]	Adversarial Prompt	ChatGPT
		[198]	Backdoor Attack	BERT
		[199]	Adversarial Attack	Large Multimodal Model

6.1.2 Physical Patch-based Attack

To enhance the effectiveness of adversarial examples in the physical world, existing work [200, 201, 202] proposes to attach printable 2D patches or painting patterns onto surfaces. However, the 2D adversarial patches can be detectable due to the preservation of 3D shape information. Therefore, Xiao *et al.* [177] propose to craft adversarial 3D meshes based on objects characterized by abundant shape features while having minimal textural variations. To manipulate either the shape or texture of objects, they employ a differentiable renderer capable of calculating precise shading effects on the shape and propagating gradients. Their research shows the notable transferability of the proposed method when applied to a black-box non-differentiable renderer with unknown parameters.

To investigate the real-life adversarial attacks on 3D objects, Wu *et al.* [178] offer a comprehensive examination of the transferability of adversarial attacks applied to object detectors, encompassing different object detectors, object categories, and diverse datasets. In the real-world physical context, the generated adversarial perturbations can be influenced by numerous factors, including variations in cameras, resolutions, lighting conditions, distances, and viewing angles. For example, the attack pattern is distorted on a moving object, such as a person walking when wearing a T-shirt with the printed attack pattern. During the evaluation, it is observed that T-shirts with YOLOv2 adversarial patterns achieve approximately a 50% success rate. However, these patterns do not exhibit strong transferability when applied to other object detectors [178]. Meanwhile, Xu *et al.* [59] propose to model the effect of deformation of an adversarial T-shirt caused by pose changes of a moving person. By taking non-rigid transformation into consideration, the attack success rate can be improved. However, they also find that the transferability of adversarial attacks drops from 74% in digital attack to 57% in the physical attack scenario.

To render physical adversarial attacks inconspicuous, Zhang *et al.* [180] introduce a 3D camouflage pattern designed to conceal vehicles from the detection of object detectors. Besides, they conduct transferability experiments, which demonstrate that the camouflage strategy targeting Mask R-CNN is also partially effective in undermining YOLOv3. Similarly, Huang *et al.* [181] and Wang *et al.* [182] have devised camouflage patterns designed to effectively disrupt object detectors, demonstrating strong transferability across various models. Alon *et al.* [179] propose a contactless translucent physical patch containing a constructed pattern, which is placed on the lens of cameras, to fool object detectors. They demonstrate the transferability of the attack when the patch is generated using a surrogate model and then applied to a different model. Cao *et al.* [183] investigate the potential of launching simultaneous attacks on all fusion sources within an autonomous driving system. Their approach involves optimizing the attack as an optimization problem to create an adversarial 3D-printed object by vertex positions. This object is designed to deceive the autonomous driving system, leading to its failure to detect the object. Such a transferable attack has a simultaneous and consistent impact on both camera images and LiDAR point clouds.

6.1.3 Projected Patch-based Attack

To explore other feasible attack vectors, Ben *et al.* [184] introduce a tactic known as the "split-second phantom attack". A phantom refers to a visual object without depth that can deceive Advanced Driver Assistance Systems (ADAS) and influence how the system perceives the object. These phantoms are generated using projectors or digital screens, such as billboards, to manipulate the perception of ADAS. The designed split-second phantom attack involves projecting actual object images, such as a stop sign or a person, onto various surfaces to manipulate computer vision perception. For instance, an attacker could implant phantom road signs onto a digital billboard, potentially causing Tesla's autopilot system to halt the vehicle unexpectedly in the middle of the road. The attack is a black-box attack that can be transferred to different object detection models. On the other hand, instead of projecting the attack image to the physical world, Man *et al.* [185] introduce a method where the attack image is projected directly onto the camera sensor using lens flare ef-

fects. This remote adversarial pattern projection is accomplished by leveraging lens flare and auto-exposure control. Different from the direct projection of the attack image, Lovisotto *et al.* [186] project adversarial examples onto the surface of the target object by establishing a projection model. Specifically, they develop a model that accounts for the impact of projections under specific environmental conditions. They assess the absolute changes in pixel colors captured by an RGB camera. The proposed approach involves the utilization of a differentiable model, through which the derivatives of the projection are propagated during the crafting phase of AEs. These AEs are trained on one model but exhibit transferability, effectively working across different models.

Cheng *et al.* [188] propose to attack monocular depth estimation through the deployment of an optimized patch in the physical environment. They strike a balance between the effectiveness and stealthiness of the attack by employing an object-oriented adversarial design, pinpointing sensitive regions, and employing natural-style camouflage. This approach demonstrates robust transferability across various objects and networks. In contrast, to execute a physical attack on stereo depth estimation, Zhou *et al.* [203] leverage the disparities in stereo correspondence within the stereo depth estimation algorithm. They orchestrate a black-box attack by utilizing two projectors mounted on a flying drone, and this attack method exhibits transferability across different scenarios and various state-of-the-art stereo depth estimation algorithms.

6.1.4 Side Channel-based Attack

The side channel refers to any indirect and unintended channels to be exploited to compromise the system, which can be regarded as a type of cross-modality attack. In this context, transferability refers to the target information being transferable via different channels. For example, keystroke inputs contain a significant amount of users' private data, such as usernames, passwords, and personal information. One typical keystroke inference attack is the vision-based method, which employs a camera to record the hand movements on the keyboard and deduce the actual keyboard input. However, existing attacks require prior knowledge about the keyboard's model, size, location, and layout [204]. Consequently, attackers must adjust their inference models to suit different victim users and devices, limiting the attack's adaptability. To address these vulnerabilities and enhance attack effectiveness across diverse victims and scenarios, Yang *et al.* [189] propose a general keystroke inference attack model. On one hand, they use a hand modeling technique to reconstruct the complete hand movements while typing on the keyboard. On the other hand, they develop a two-layer self-supervised model for keystroke recognition. This approach involves using a Hidden Markov Model to deduce linguistic information, complemented by 3D-CNN models trained to precisely predict keystroke inputs.

There are physical interactions between different media in different channels, which can be used as an attack vector. Ji *et al.* [190] propose to inject acoustic signals into the inertial sensors, which affects the camera stabilization results to blur the image to fool the object detectors. They formulate the attack process and use Bayesian Optimization to launch a black-box attack with the attack transferability on different detector models. Similarly to [190], Zhu *et al.* [191] utilize acoustic signals injection towards cameras to launch a content-based camouflage attack. It remains benign under normal circumstances but can be triggered when the acoustic signals are injected. To make the attack more stealthy, Zhou *et al.* [187] introduce an approach involving the placement of small bulbs on a board to render infrared pedestrian detectors ineffective in detecting pedestrians in real-world scenarios. For instance, they achieve a significant drop of 34.48% in average precision by constructing a physical board to target the YOLOv3 model. They employ model ensemble techniques to simultaneously reduce the maximum objectness score of each detector, thereby enhancing the transferability of the attack. Jiang *et al.* [192] use intentional electromagnetic interference (IEMI) to actively induce controlled glitch images of a camera at various positions, widths, and numbers. The attack is in a black-box matter, which has the transferability to various cameras, object detection models.

Current attacks towards image classification and object detection systems are difficult to attack models

of object tracking in video because the tracking algorithms can handle sequential information across video frames. Besides, the categories of targets tracked are normally unknown in advance. Therefore, Chen *et al.* [205] propose an attack toward object tracking systems by generating robust attack patches with transferability to different categories, made possible by the extensive prior knowledge gained through substantial amounts of offline training. To attack lane detection models, Takami *et al.* [206] design dirty road patches as a domain-specific attack vector for physical-world adversarial attacks on Automated Lane Centering systems. They design a lane-bending objective function as a differentiable surrogate function. With the proposed method, they achieve a 63% transfer success rate when applying the patch generated by one model to the other two models.

6.1.5 Image and Video Forgery Detection

Obtaining the facial image of an authentic user poses a security threat, e.g., face presentation attacks in a video replay. Researchers have attempted to address these face presentation attacks by investigating inherent differences between genuine and fraudulent faces, such as variations in color texture, image distortion, temporal fluctuations, and deep semantic characteristics [207]. Nonetheless, the effectiveness of these approaches may diminish when applied in novel application scenarios. Wang *et al.* [208] propose a system designed for face presentation attack detection (PAD). More specifically, they develop disentangled representation learning, a method that utilizes a generative model to segregate PAD-relevant features from subject-specific ones. These disentangled characteristics originating from various domains are subsequently input into a multi-domain learning neural network to extract domain-independent features, ultimately serving the purpose of cross-domain face PAD.

Surveillance videos play a critical role in monitoring real-world scenarios for potential attacks. However, these live surveillance videos are susceptible to forgery attempts such as deepfake. When malicious intrusions manipulate keyframes in fake videos, illicit activities can remain undetected by the surveillance system. Traditional methods for detecting video forgery heavily rely on extensive spatial-temporal analysis to pinpoint artificial alterations within surveillance videos. Although highly effective, they exhibit low efficiency and require substantial computational resources, leading to high computation costs. Meanwhile, leveraging wireless signals offers the potential for multi-modal video verification, which relaxes the requirement for computational resources. Nonetheless, this method struggles to uncover imperfections in each separate frame. Secure-Pose [209] introduces a cross-modal video forgery detection system. By deploying coexisting cameras and WiFi technology, Secure-Pose deciphers human semantic information through the comparison of wireless signal data and video frames. When adversaries manipulate videos, discrepancies arise between the video content and WiFi sensing data, which offers an opportunity for detecting malicious video manipulations. Essentially, Secure-Pose establishes a relationship between video recordings and WiFi CSI sensing, irrespective of camera models or physical settings. This advancement greatly improves the adaptability of defense mechanisms against unauthorized surveillance video forgery, reinforcing security in real-world surveillance scenarios.

6.2 Transferable Attack on Smart Audio System

6.2.1 Attacks on Data Processing Module

From a system perspective, there is typically a standard data pre-processing procedure that occurs before inputting data into the model. Several research [210, 153] proposes to implement the physical attack toward the data pre-processing pipeline that precedes the deep learning model. For example, camouflage attacks [210] aim at compromising the image scaling algorithm within the image processing pipeline. Consequently, this manipulation leads to significant modifications in the visual semantics right after the scaling

process. In this way, such attack methods are transferable to multiple systems. A similar attack philosophy has been proposed in [153] for speech recognition and speaker verification systems. The attacks aim to achieve mistranscription and misidentification in voice control systems with minimal impact on human comprehension. Rather than targeting the model directly, the authors focus on attacking the data processing pipeline, which is similar across various systems. The pipeline includes signal preprocessing and feature extraction steps, with the resulting outputs being fed into a machine learning-based model.

6.2.2 Attacks on Speech Recognition Module

The voice eavesdropping attacks can be regarded as a type of cross-modality attack by transferring information between different modalities. There are significant challenges to implementing learning-based eavesdropping attacks. For example, gathering a large amount of sensor data, particularly from a range of different mobile device models, is infeasible for attackers. Besides, the low sampling rate of built-in sensors, such as accelerometers and gyroscopes, severely limits eavesdropping accuracy. VoiceListener [193] represents an eavesdropping attack designed to intercept audio signals emitted by smartphone loudspeakers. Different from existing methodologies, VoiceListener employs multiple sensors for a multi-domain eavesdropping model, to enable precise audio signal recovery. To further enhance audio signal fidelity, VoiceListener incorporates pitch estimation and aliasing correction techniques, facilitating the recovery of high-sampling-rate audio data from built-in sensors. Compared to previous attack techniques, VoiceListener is a training-free method capable of effortless adaptation to a wide array of domains, devices, and speaker characteristics, highlighting its transferability.

6.2.3 Attacks on Speaker Recognition Module

Voice assistants, while offering convenience, can be susceptible to audio signal injection attacks. These attacks illegally trigger voice commands on the targeted device. Among these attacks, noteworthy examples include those that introduce voice commands through mediums such as ultrasound [211], guided waves [212], or laser [213]. However, the existing methods require the voice of legitimate users to mimic the wake-up word, so as to activate voice assistant and inject subsequent malicious commands.

A **replay attack** can fool the voice communication system by replaying the pre-collected voice samples of legitimate users. Researchers exploit the nonlinearity of the microphone circuit and the defect of deep learning algorithms to design various attacks. The distortion during sound transmission in the physical world is taken into account, and the impulse response is used to simulate the distortion over the air. However, with black-box ASV, there are two main challenges in designing black-box attacks. **First**, the model architecture design and parameter selection of most ASVs are proprietary, so the model setup is unknown and the system works in a black-box manner. **Second**, many ASVs incorporate random challenges in the authentication process, rendering replay or synthesizing attacks less practical in smart home environments. To attack the speaker verification, inspired by adversarial examples, VMask [152] is proposed to generate attack audio that sounds like the source speaker but would be recognized as the target speaker by the ASV. In particular, psychoacoustic masking is employed to manipulate the adversarial perturbations under the human perception threshold, thereby keeping the victim unaware of the ongoing attacks.

PhyTalker [194] proposes a physical domain adversarial attack targeting speaker verification models. In contrast to conventional attacks, PhyTalker leverages a live stream adversarial perturbation technique in conjunction with live speech, thereby posing a unique threat. To enhance its effectiveness, PhyTalker injects subphoneme-level perturbations into live human speech, strategically undermining the target speaker verification model. Notably, PhyTalker places a strong emphasis on enhancing transferability, which involves attackers collecting data from multiple speaker verification models. Through the application of ensemble learning techniques, they optimize the adversarial perturbations. Moreover, PhyTalker offers a high degree

of flexibility by permitting the adjustment of perturbation weights, allowing for seamless adaptation to a variety of target speaker verification models. In the end, the multifaceted approach can enable a general and transferable attack method for different target speaker verification models.

To address the constraints of attack transferability and indiscriminately bypass the speaker verification process, GhostTalk [195] is a new attack approach that injects inaudible voice commands through a powerline. While users charge their devices with what appear to be regular cables, hidden circuits within these malicious cables can take control of the mobile device's audio system. Once connected, the victim's device is essentially linked to a rogue earphone, the audio input of which can be remotely controlled by an attacker. By using a switch to connect the ground and the microphone, the attacker simulates the action of a user pressing the earphone button. This allows GhostTalk to activate the voice assistant, even without any prior knowledge of the victim, enabling a more transferable attack. Additionally, because the injected audio signal is transmitted over the line, external noise does not degrade the audio quality, ensuring GhostTalk remains robust under noisy environments or liveness detection methods.

6.2.4 Audio Replay Detection

Various defenses have been proposed for replay audio detection [214, 215, 216]. However, a common challenge lies in the high dependence on training datasets, making it difficult to generalize the defenses for a wide range of application scenarios. Wang *et al.* [217] have pointed out that domain mismatch is a potential reason for the limited transferability of existing defenses. To combat this issue, they employ adversarial training techniques to boost defense transferability. They introduce a domain classifier to achieve unsupervised domain adaptation, enabling adaptation to unseen replay attack examples and strengthening the defense transferability.

Nevertheless, a single-domain detector may not provide sufficient robustness, as sophisticated adversaries can manipulate cross-domain data alignments to deceive anti-spoofing models. There is a continued need for the development of more robust transferable anti-spoofing defenses against adaptive attackers. To further enhance the performance of the method in [217], Wang *et al.* [218] introduce a new defense framework based on dual-adversarial domain adaptation. In contrast to a single cross-domain adversarial training approach, this framework employs two separate cross-domain discriminators—one for aligning spoofing data and another for real data. Consequently, this dual-adversarial adaptation method demonstrates superior robustness and effectiveness in detecting replay audio attacks.

6.3 Transferable Attack on Large Language Model System

6.3.1 Transferable Attack on Large Language Model System

Large language models (LLMs) are grounded in transformer architectures, which leverage self-attention mechanisms to capture complex patterns and relationships in textual data, enabling the models to generate coherent and contextually relevant text. Many efforts [196, 219, 197, 198, 199] have been devoted to exploring the vulnerability of LLMs.

Boucher *et al.* [196] delve into the realm of adversarial examples that can target text-based models in a black-box setting. Their approach involves employing encoding-specific perturbations, including techniques such as introducing invisible characters, homoglyphs, reordering, or deletion. These perturbations are crafted to be imperceptible to human observers while effectively manipulating the outputs of NLP systems. These imperceptible attacks represent a significant challenge for contemporary language processing systems, underscoring the importance of developing robust defense mechanisms and detection strategies to safeguard against such subtle but potentially impactful manipulations of text data. In the case of ChatGPT, adversarial prompts have the potential to lead the model to generate harmful or false information, posing significant security risks, as discussed in the work [197]. To mitigate these risks, the authors introduce two

solutions. First, the training-free prefix prompt mechanism is designed to detect and prevent the generation of toxic or harmful texts without the need for additional training. It involves implementing a prefix prompt mechanism that can filter out potentially problematic responses generated by the model. Second, the RoBERTa-based detection mechanism utilizes a RoBERTa-based model to identify manipulative or misleading input texts. This mechanism incorporates an external detection model to assess the nature of the input and determine if it contains deceptive or harmful elements, thus helping to safeguard against the generation of inappropriate or false responses. These strategies are aimed at enhancing the security and trustworthiness of ChatGPT by proactively identifying and mitigating the generation of harmful or misleading content.

Prompt-based learning, a popular NLP approach, is vulnerable to backdoor attacks. Different from backdoor attacks in computer vision that use different patterns such as invisible patterns, one-pixel patterns, or leveraging natural semantic objects, established trigger designs in NLP often revolve around specific words or phrases as trojans. For example, Mei *et al.* [198] propose transferable backdoor attacks against prompt-based NLP models, named NOTABLE. These attacks are independent of downstream tasks and prompting strategies, achieving superior attack performance. NOTABLE [198] injects backdoors into model encoders and utilizes adaptive verbalizers to identify target anchors. Finally, the proposed method successfully compromises various NLP tasks with high attack success rates. This approach binds triggers directly to target anchors in the encoder, distinguishing it from existing attacks that inject backdoors into embedding layers or word embedding vectors. By building direct shortcut connections between triggers and target anchors, NOTABLE achieves transferability to various prompt-based tasks and models, signifying a critical concern in the design and deployment of prompt-based systems.

In general, LLMs are fine-tuned to align with the desired goals and values of their creators, which is often referred to as alignment. Carlini *et al.* [199] have conducted a study to investigate the extent to which these models remain aligned when subjected to adversarial examples. They found that existing optimization attacks in the context of NLP are not potent enough to successfully attack aligned text models. These models appear to maintain their alignment even when exposed to adversarial perturbations. On the other hand, multimodal-based LLMs, which can process both text and images, are more susceptible to attacks. Adversarial perturbations applied to input images can lead these models to exhibit arbitrary unaligned behavior, potentially deviating from the intended alignment and generating unexpected or harmful outputs.

6.3.2 Machine Generated Text Detection

The advent of LLMs such as ChatGPT and Bard, marks a significant leap in AI capabilities. LLMs demonstrate remarkable proficiency in creating documents and generating executable code [220]. However, this surge in LLM usage has raised substantial concerns regarding potential misuse, ranging from orchestrating social engineering and election manipulation campaigns via automated bots on social media to creating misleading content and enabling academic dishonesty using AI [221, 222]. To understand the characteristics of machine-generated text, Pu *et al.* [223] conduct a measurement study to gather synthetic text from real-world sources, presenting four new datasets. These datasets encompass text generated by state-of-the-art Transformer-based models from prominent text-generation platforms and a GPT3-powered bot on Reddit. The authors subsequently evaluate six state-of-the-art defenses, unveiling a concerning trend - many defenses experience significant performance degradation compared to their originally claimed efficacy. This highlights the evolving challenges in creating a universally effective defense tool with high transferability.

Beyond detection, the traceability and provenance of text are vital considerations. Recent research [224] introduces watermarking techniques to trace the origin of the text. Watermarking is achieved through models such as the Adversarial Watermarking Transformer (AWT), designed to embed binary messages unobtrusively into input text using adversarial training and a joint encoder-decoder architecture. AWT emphasizes minimizing semantic impact, ensuring the encoded data remains effectively hidden, and demonstrating robustness against various attacks. AWT is successful in preserving text utility while hiding encoded data

from adversaries. The model exhibits robustness against various attacks, making it a promising solution for watermarking text. Another work [225] proposes a watermarking framework for proprietary language models, aiming to embed invisible signals into generated text that can be algorithmically detected. The watermarking process selects a randomized set of "green" tokens before text generation and promotes their use during sampling. The watermark can be detected using an efficient open-source algorithm without access to the language model API or parameters. The watermark remains detectable when only a contiguous portion of the generated text is used to create a larger document, ensuring its robustness. Research in deepfake text detection and watermarking continues to evolve, aiming towards creating a more secure and trustworthy digital communication environment.

6.4 Summary

In this section, we have reviewed the concept of transferable attacks across multiple AI-based systems, including computer vision systems, speech recognition systems, and LLM systems. The different features of AI-based systems introduce a significant level of complexity when designing transferable attacks. Each stage such as data collection, data processing, model design, model training, and model application is vulnerable to potential transferable attacks. For instance, during the data collection stage, vulnerabilities may arise that can be exploited through data poisoning attacks. In the model training phase, attackers may inject backdoors to compromise the integrity of the model. As for the model application stage, the introduction of adversarial perturbations can potentially mislead AI-based systems. These transferable attack strategies, which transcend domains and modalities, significantly increase the vulnerability of AI-based systems.

7 Future Directions

Enhancing the transferability of learning-based attacks in the cyber-physical world necessitates sustained efforts. The attack designs based on the white-box or grey-box scenarios provide us with a great foundation for the black-box attack design. However, the new attack design should be built upon a more practical threat model. We discuss five main research directions toward transferable learning-based attack design.

Attack Practicality. The assumption of the attacker's capability should be more realistic. For instance, for an adversarial attack, it is not reasonable to obtain the training data of the target model. Thus, the attack design should avoid relying on the original training dataset. Many attack designs aim to achieve data-free attacks [226, 173, 227, 228] with generative models. However, the synthesized data suffers from low data quality and data distribution shift problems. Thus, more advanced algorithms need to be designed to accommodate the limited capability of attackers to find effective perturbations.

Attack Stealthiness. The attack scenarios should be more stealthy. There is a balance between stealthiness and transferability in the attack design. Take an adversarial audio attack as an example, the perceptibility of adversarial audio attacks constrains the feasibility of attacks. For physical attacks especially, weak perturbations are often lost during airborne transmission. However, the perturbation can lead to noticeable noise, potentially alerting users to the presence of an attack. Recent work has employed ultrasound to inject inaudible perturbations to mitigate this issue [211, 212, 229]. Still, either short-range access or prior knowledge about the target model is required. Another example is the adversarial patch, which shows transferability in the physical world. However, additional constraints are typically imposed to avoid being noticed in the physical world, which tends to diminish the attack's transferability. Therefore, an important future direction is to enhance the imperceptibility of the attack while maintaining the transferability.

Physical Environments. The new attack design should consider more unexpected physical factors. For example, for the audio attack design, most of the existing attacks necessitate an uninterrupted scenario. Adversaries must play the malicious audio without any external interference. If the adversarial audio signal

intertwines with the user's benign voice command, the attack will likely fail. Regarding the physical projection attacks, either the attacks use visible light or invisible light, they usually work in the condition of low illumination and short attack distance. Their attack success rate would be significantly reduced if the specific condition is not satisfied. Moreover, even though many attacks have high transferability for different victim models in the digital world, they are still very difficult to transfer to the physical domain due to dynamic physical environments.

Computation Costs. The attack design should consider the computation cost. Many methods rely on querying target models such as commercial APIs, which can be time-consuming and incur financial costs. Moreover, the generation of transferable adversarial examples requires a prolonged time. For example, the universal perturbations require extensive gradient updates in many iterations, limiting its efficiency for adversarial perturbation generation. Nevertheless, increasing the transferability of attacks can reduce the cost of attacks since it can be applied in different domains even modalities. For example, once the speech recognition model is updated, previously generated transferable adversarial examples should remain valid. Therefore, the new attacks should enhance the efficiency and transferability to reduce the attack costs.

Multi-modality. Apart from the image, text, audio, and video, there are many other signals (*e.g.*, accelerometer, gyroscopes, Lidar signals) that can sense the environment and people [230, 231, 232]. These signals are also vulnerable to various attacks. We need to consider the specific features and requirements when designing attacks on different types of data. Besides, multimodal systems have gained popularity in the cyber-physical world. A transferable attack on a multimodal model involves crafting attacks to be effective to multimodal data and models. These attacks can exploit vulnerabilities in the cross-modal fusion mechanisms, underscoring the importance of robustness against various data types in multimodal systems. In the future, we need to explore the comprehensive defense strategies that account for potential vulnerabilities in multimodal systems.

8 Conclusion

With a focus on transferability, in this survey, we have reviewed and analyzed various attacks in different domains. We first provide background knowledge about the design of learning-based attacks. Then, from data, process, model, and system perspectives, we have summarized the state-of-the-art research that enhances the transferability of different attacks. Specifically, we have reviewed various attack methods to enhance transferability based on data augmentation. We have summarized the gradient, heuristic, and generative-based methods to build cross-domain universal adversarial attacks. We have explored methods from the model perspective to enhance transferability, including ensemble-based approaches, optimization techniques, and the use of pre-trained foundation models. Furthermore, we have reviewed the transferable attack applications in different areas such as computer vision systems, speech recognition systems, and large language model systems. In conclusion, we have identified various challenges that future research should address. Investigating transferable attacks can assist in developing countermeasures, enhancing system defense, and contributing to the development of more secure and reliable machine learning models.

References

[1] OpenAI, "Gtp-4." `https://openai.com/research/gpt-4`, 2023.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] W. Jin, Y. Li, H. Xu, Y. Wang, and J. Tang, "Adversarial attacks and defenses on graphs: A review and empirical study," *arXiv preprint arXiv:2003.00653*, vol. 10, no. 3447556.3447566, 2020.

[4] J. Ma, S. Ding, and Q. Mei, "Towards more practical adversarial attacks on graph neural networks," *Advances in neural information processing systems*, vol. 33, pp. 4756–4766, 2020.

[5] G. Wang, J.-H. Lai, W. Liang, and G. Wang, "Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10568–10577, 2020.

[6] L. Jiang, X. Ma, S. Chen, J. Bailey, and Y.-G. Jiang, "Black-box adversarial attacks on video recognition models," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 864–872, 2019.

[7] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, "Skip connections matter: On the transferability of adversarial examples generated with resnets," *arXiv preprint arXiv:2002.05990*, 2020.

[8] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, I. Kevin, and K. Wang, "Hierarchical adversarial attacks against graph-neural-network-based iot network intrusion detection system," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9310–9319, 2021.

[9] H. Dai, H. Li, T. Tian, X. Huang, L. Wang, J. Zhu, and L. Song, "Adversarial attack on graph structured data," in *International conference on machine learning*, pp. 1115–1124, PMLR, 2018.

[10] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1962–1966, IEEE, 2018.

[11] H. Liu, J. Jia, and N. Z. Gong, "{PoisonedEncoder}: Poisoning the unlabeled pre-training data in contrastive learning," in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 3629–3645, 2022.

[12] W. Jiang, H. Li, G. Xu, and T. Zhang, "Color backdoor: A robust poisoning attack in color space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8133–8142, 2023.

[13] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli, "Wild patterns reloaded: A survey of machine learning security against training data poisoning," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.

[14] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, "Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation," in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 3611–3628, 2022.

[15] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, and M. Sun, "Mind the style of text! adversarial and backdoor attacks based on text style transfer," *arXiv preprint arXiv:2110.07139*, 2021.

[16] Y. Liu, Z. Li, M. Backes, Y. Shen, and Y. Zhang, "Backdoor attacks against dataset distillation," *arXiv preprint arXiv:2301.01197*, 2023.

[17] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[18] S. Kariyappa, A. Prakash, and M. K. Qureshi, "Maze: Data-free model stealing attack using zeroth-order gradient estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13814–13823, 2021.

[19] J. R. Correia-Silva, R. F. Berriel, C. Badue, A. F. De Souza, and T. Oliveira-Santos, "Copycat cnn: Are random non-labeled data enough to steal knowledge from black-box models?," *Pattern Recognition*, vol. 113, p. 107830, 2021.

[20] D. Oliynyk, R. Mayer, and A. Rauber, "I know what you trained last summer: A survey on stealing machine learning models and defences," *ACM Computing Surveys*, 2023.

[21] T. Orekondy, B. Schiele, and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4954–4963, 2019.

[22] Y. Liu, J. Jia, H. Liu, and N. Z. Gong, "Stolenencoder: stealing pre-trained encoders in self-supervised learning," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2115–2128, 2022.

[23] K. Krishna, G. S. Tomar, A. P. Parikh, N. Papernot, and M. Iyyer, "Thieves on sesame street! model extraction of bert-based apis," *arXiv preprint arXiv:1910.12366*, 2019.

[24] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

[25] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 253–261, 2020.

[26] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 148–162, 2019.

[27] J. Song and D. Namiot, "A survey of the implementations of model inversion attacks," in *International Conference on Distributed Computer and Communication Networks*, pp. 3–16, Springer, 2022.

[28] H. Liu, J. Jia, W. Qu, and N. Z. Gong, "Encodermi: Membership inference against pre-trained encoders in contrastive learning," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2081–2095, 2021.

[29] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18, IEEE, 2017.

[30] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[31] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot, "Label-only membership inference attacks," in *International conference on machine learning*, pp. 1964–1974, PMLR, 2021.

[32] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, IEEE, 2022.

[33] Y. Chen, X. Yuan, J. Zhang, Y. Zhao, S. Zhang, K. Chen, and X. Wang, "{Devil's} whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 2667–2684, 2020.

[34] K. Liang and B. Xiao, "Styless: Boosting the transferability of adversarial examples," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8163–8172, 2023.

[35] K. Krishna, J. Wieting, and M. Iyyer, "Reformulating unsupervised style transfer as paraphrase generation," *arXiv preprint arXiv:2010.05700*, 2020.

[36] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, "Hidden trigger backdoor attack on nlp models via linguistic style manipulation," in *USENIX Security Symposium*, 2022.

[37] X. Liu, Y. Zhong, Y. Zhang, L. Qin, and W. Deng, "Enhancing generalization of universal adversarial perturbation through gradient aggregation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4435–4444, 2023.

[38] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

[39] V. Khrulkov and I. Oseledets, "Art of singular vectors and universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8562–8570, 2018.

[40] G. Tao, S. An, S. Cheng, G. Shen, and X. Zhang, "Hard-label black-box universal adversarial patch attack," in *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 697–714, 2023.

[41] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing nlp," in *Conference on Empirical Methods in Natural Language Processing*, 2019.

[42] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *ArXiv*, vol. abs/2307.15043, 2023.

[43] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

[44] M. M. Naseer, S. H. Khan, M. H. Khan, F. Shahbaz Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[45] M. Chen, L. Lu, Z. Ba, and K. Ren, "Phoneytalker: An out-of-the-box toolkit for adversarial example attack on speaker recognition," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1419–1428, IEEE, 2022.

[46] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon, "Data-free universal adversarial perturbation and black-box attack," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7868–7877, 2021.

[47] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[48] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[49] Y. Shi, Y. Sagduyu, and A. Grushin, "How to steal a machine learning classifier with deep learning," in *2017 IEEE International symposium on technologies for homeland security (HST)*, pp. 1–5, IEEE, 2017.

[50] S. Liang, A. Liu, J. Liang, L. Li, Y. Bai, and X. Cao, "Imitated detectors: Stealing knowledge of black-box object detectors," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4839–4847, 2022.

[51] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.

[52] Z. Zhou, S. Hu, R. Zhao, Q. Wang, L. Y. Zhang, J. Hou, and H. Jin, "Downstream-agnostic adversarial examples," *arXiv preprint arXiv:2307.12280*, 2023.

[53] S. Hu, Z. Zhou, Y. Zhang, L. Y. Zhang, Y. Zheng, Y. He, and H. Jin, "Badhash: Invisible backdoor attacks against deep hashing with clean label," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 678–686, 2022.

[54] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059, IEEE, 2022.

[55] L. Sun, Y. Dou, C. Yang, K. Zhang, J. Wang, S. Y. Philip, L. He, and B. Li, "Adversarial attack and defense on graph data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[56] S. Qiu, Q. Liu, S. Zhou, and W. Huang, "Adversarial attack and defense technologies in natural language processing: A survey," *Neurocomputing*, vol. 492, pp. 278–307, 2022.

[57] C. Zhang, P. Benz, C. Lin, A. Karjauv, J. Wu, and I. S. Kweon, "A survey on universal adversarial attack," *arXiv preprint arXiv:2103.01498*, 2021.

[58] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, 2021.

[59] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, pp. 151–178, 2020.

[60] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[61] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.

[62] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *Ieee Access*, vol. 6, pp. 14410–14430, 2018.

[63] H. Yang, K. Xiang, H. Li, and R. Lu, "A comprehensive overview of backdoor attacks in large language models within communication networks," *arXiv preprint arXiv:2308.14367*, 2023.

[64] B. Yan, J. Lan, and Z. Yan, "Backdoor attacks against voice recognition systems: A survey," *arXiv preprint arXiv:2307.13643*, 2023.

[65] Y. Li, S. Zhang, W. Wang, and H. Song, "Backdoor attacks to deep learning models and countermeasures: A survey," *IEEE Open Journal of the Computer Society*, 2023.

[66] M. Omar, "Backdoor learning for nlp: Recent advances, challenges, and future research directions," *arXiv preprint arXiv:2302.06801*, 2023.

[67] T. D. Nguyen, T. Nguyen, P. L. Nguyen, H. H. Pham, K. Doan, and K.-S. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *arXiv preprint arXiv:2303.02213*, 2023.

[68] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Mądry, B. Li, and T. Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, 2022.

[69] W. Guo, B. Tondi, and M. Barni, "An overview of backdoor attacks against deep neural networks and possible defences," *IEEE Open Journal of Signal Processing*, 2022.

[70] X. Sheng, Z. Han, P. Li, and X. Chang, "A survey on backdoor attack and defense in natural language processing," in *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*, pp. 809–820, IEEE, 2022.

[71] S. Kaviani and I. Sohn, "Defense against neural trojan attacks: A survey," *Neurocomputing*, vol. 423, pp. 651–667, 2021.

[72] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *arXiv preprint arXiv:2007.10760*, 2020.

[73] Y. Liu, A. Mondal, A. Chakraborty, M. Zuzak, N. Jacobsen, D. Xing, and A. Srivastava, "A survey on neural trojans," in *2020 21st International Symposium on Quality Electronic Design (ISQED)*, pp. 33–39, IEEE, 2020.

[74] H. Chen and F. Koushanfar, "Tutorial: Toward robust deep learning against poisoning attacks," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 3, pp. 1–15, 2023.

[75] G. Xia, J. Chen, C. Yu, and J. Ma, "Poisoning attacks in federated learning: A survey," *IEEE Access*, vol. 11, pp. 10708–10722, 2023.

[76] Z. Tian, L. Cui, J. Liang, and S. Yu, "A comprehensive survey on poisoning attacks and countermeasures in machine learning," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–35, 2022.

[77] C. Wang, J. Chen, Y. Yang, X. Ma, and J. Liu, "Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects," *Digital Communications and Networks*, vol. 8, no. 2, pp. 225–234, 2022.

[78] M. A. Ramirez, S.-K. Kim, H. A. Hamadi, E. Damiani, Y.-J. Byon, T.-Y. Kim, C.-S. Cho, and C. Y. Yeun, "Poisoning attacks and defenses on artificial intelligence: A survey," *arXiv preprint arXiv:2202.10276*, 2022.

[79] J. Fan, Q. Yan, M. Li, G. Qu, and Y. Xiao, "A survey on data poisoning attacks and defenses," in *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pp. 48–55, IEEE, 2022.

[80] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, and K. Ren, "Threats to training: A survey of poisoning attacks and defenses on machine learning systems," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–36, 2022.

[81] I. M. Ahmed and M. Y. Kashmoola, "Threats on machine learning technique by data poisoning attack: A survey," in *Advances in Cyber Security: Third International Conference, ACeS 2021, Penang, Malaysia, August 24–25, 2021, Revised Selected Papers 3*, pp. 586–600, Springer, 2021.

[82] L. Hu, A. Yan, H. Yan, J. Li, T. Huang, Y. Zhang, C. Dong, and C. Yang, "Defenses to membership inference attacks: A survey," *ACM Computing Surveys*, 2023.

[83] X. Zhang, C. Chen, Y. Xie, X. Chen, J. Zhang, and Y. Xiang, "A survey on privacy inference attacks and defenses in cloud-based deep neural network," *Computer Standards & Interfaces*, vol. 83, p. 103672, 2023.

[84] X. Gong, Y. Chen, Q. Wang, M. Wang, and S. Li, "Private data inference attacks against cloud: Model, technologies, and research directions," *IEEE Communications Magazine*, vol. 60, no. 9, pp. 46–52, 2022.

[85] Y. Bai, T. Chen, and M. Fan, "A survey on membership inference attacks against machine learning," *management*, vol. 6, p. 14, 2021.

[86] X. Zhang, C. Chen, Y. Xie, X. Chen, J. Zhang, and Y. Xiang, "Privacy inference attacks and defenses in cloud-based deep neural network: A survey," *arXiv preprint arXiv:2105.06300*, 2021.

[87] J. Jia and N. Z. Gong, "Defending against machine learning based inference attacks via adversarial examples: Opportunities and challenges," *Adaptive autonomous secure cyber systems*, pp. 23–40, 2020.

[88] S. V. Dibbo, "Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap," in *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pp. 439–456, IEEE, 2023.

[89] R. Zhang, S. Guo, J. Wang, X. Xie, and D. Tao, "A survey on gradient inversion: Attacks, defenses and future directions," *arXiv preprint arXiv:2206.07284*, 2022.

[90] Z. Li, L. Wang, G. Chen, M. Shafq, *et al.*, "A survey of image gradient inversion against federated learning," 2022.

[91] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[92] Y. Miao, C. Chen, L. Pan, Q.-L. Han, J. Zhang, and Y. Xiang, "Machine learning–based cyber attacks targeting on controlled information: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1–36, 2021.

[93] Y. He, G. Meng, K. Chen, X. Hu, and J. He, "Towards security threats of deep learning systems: A survey," *IEEE Transactions on Software Engineering*, vol. 48, no. 5, pp. 1743–1770, 2020.

[94] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Computing Surveys*, 2020.

[95] A. Kumar and S. Mehta, "A survey on resilient machine learning," *arXiv preprint arXiv:1707.03184*, 2017.

[96] D. Genç, M. Özuysal, and E. Tomur, "A taxonomic survey of model extraction attacks," in *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 200–205, IEEE, 2023.

[97] X. Gong, Q. Wang, Y. Chen, W. Yang, and X. Jiang, "Model extraction attacks and defenses on cloud-based machine learning models," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 83–89, 2020.

[98] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[99] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[100] J. Wang and Y. Chen, *Introduction to Transfer Learning: Algorithms and Practice*. Springer Nature, 2023.

[101] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer learning in deep reinforcement learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[102] G. Wang, N. Ivanov, B. Chen, Q. Wang, T. Nguyen, and Q. Yan, "Graph learning for interactive threat detection in heterogeneous smart home rule data," *Proceedings of the ACM on Management of Data*, vol. 1, no. 1, pp. 1–27, 2023.

[103] G. Wang, Q. Yan, S. Patrarungrong, J. Wang, and H. Zeng, "Facer: Contrastive attention based expression recognition via smartphone earpiece speaker," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pp. 1–10, IEEE, 2023.

[104] R. Liu, J. Gao, J. Zhang, D. Meng, and Z. Lin, "Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 10045–10067, 2021.

[105] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 402–410, 2018.

[106] T. Fritz, "A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics," *Advances in Mathematics*, vol. 370, p. 107239, 2020.

[107] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

[108] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

[109] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," *arXiv preprint arXiv:1609.07843*, 2016.

[110] S. Sanyal, S. Addepalli, and R. V. Babu, "Towards data-free model stealing in a hard label setting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15284–15293, 2022.

[111] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2730–2739, 2019.

[112] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16158–16167, 2021.

[113] S. Fang, J. Li, X. Lin, and R. Ji, "Learning to learn transferable attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 571–579, 2022.

[114] S. Koffas, L. Pajola, S. Picek, and M. Conti, "Going in style: Audio backdoors through stylistic transformations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[115] Y. Cao, X. Xiao, R. Sun, D. Wang, M. Xue, and S. Wen, "Stylefool: Fooling video classification systems via style transfer," in *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1631–1648, IEEE, 2023.

[116] X. Wei, S. Liang, N. Chen, and X. Cao, "Transferable adversarial attacks for image and video object detection," *arXiv preprint arXiv:1811.12641*, 2018.

[117] J. Liu, Y. Kang, D. Tang, K. Song, C. Sun, X. Wang, W. Lu, and X. Liu, "Order-disorder: Imitation adversarial attacks for black-box neural ranking models," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2025–2039, 2022.

[118] M. Li, Y. Yang, K. Wei, X. Yang, and H. Huang, "Learning universal adversarial perturbation by adversarial example," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 1350–1358, 2022.

[119] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," *arXiv preprint arXiv:1908.06281*, 2019.

[120] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

[121] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

[122] L. Gao, Q. Zhang, J. Song, X. Liu, and H. T. Shen, "Patch-wise attack for fooling deep neural network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp. 307–322, Springer, 2020.

[123] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7639–7648, 2021.

[124] J. Zhang, W. Wu, J.-t. Huang, Y. Huang, W. Wang, Y. Su, and M. R. Lyu, "Improving adversarial transferability via neuron attribution-based attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14993–15002, 2022.

[125] T. Takemura, N. Yanai, and T. Fujiwara, "Model extraction attacks against recurrent neural networks," *arXiv preprint arXiv:2002.00123*, 2020.

[126] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot, "High accuracy and high fidelity extraction of neural networks," in *29th USENIX security symposium (USENIX Security 20)*, pp. 1345–1362, 2020.

[127] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 181–196, 2018.

[128] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2414–2423, 2016.

[129] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.

[130] Y. Jing, Y. Mao, Y. Yang, Y. Zhan, M. Song, X. Wang, and D. Tao, "Learning graph neural networks for image style transfer," in *European Conference on Computer Vision*, pp. 111–128, Springer, 2022.

[131] M. Zhao, B. An, W. Gao, and T. Zhang, "Efficient label contamination attacks against black-box learning models," in *IJCAI*, pp. 3945–3951, 2017.

[132] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. 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*, pp. 1265–1282, 2019.

[133] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 182–199, Springer, 2020.

[134] S. Cheng, Y. Liu, S. Ma, and X. Zhang, "Deep feature space trojan attack of neural networks by controlled detoxification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1148–1156, 2021.

[135] L. Yang, Z. Chen, J. Cortellazzi, F. Pendlebury, K. Tu, F. Pierazzi, L. Cavallaro, and G. Wang, "Jigsaw puzzle: Selective backdoor attack to subvert malware classifiers," in *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 719–736, IEEE, 2023.

[136] H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. 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*, pp. 363–377, 2021.

[137] M. Xue, X. Wang, S. Sun, Y. Zhang, J. Wang, and W. Liu, "Compression-resistant backdoor attack against deep neural networks," *Applied Intelligence*, pp. 1–16, 2023.

[138] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," *arXiv preprint arXiv:2004.04692*, 2020.

[139] E. Hovy, "Generating natural language under pragmatic constraints," *Journal of Pragmatics*, vol. 11, no. 6, pp. 689–719, 1987.

[140] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective api built for detecting toxic comments," *arXiv preprint arXiv:1702.08138*, 2017.

[141] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," *arXiv preprint arXiv:1804.06059*, 2018.

[142] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[143] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1738–1742, IEEE, 2020.

[144] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1121–1134, 2020.

[145] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Universal adversarial perturbations generative network for speaker recognition," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2020.

[146] W. Zhang, S. Zhao, L. Liu, J. Li, X. Cheng, T. F. Zheng, and X. Hu, "Attack on practical speaker verification system using universal adversarial perturbations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2575–2579, IEEE, 2021.

[147] H. Guo, G. Wang, Y. Wang, B. Chen, Q. Yan, and L. Xiao, "Phantomsound: Black-box, query-efficient audio adversarial attack via split-second phoneme injection," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, pp. 366–380, 2023.

[148] Y. Wang, H. Guo, G. Wang, B. Chen, and Q. Yan, "Vsmask: Defending against voice synthesis attack via real-time predictive perturbation," in *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, p. 239–250, 2023.

[149] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013.

[150] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[151] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli, "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks," in *28th USENIX security symposium (USENIX security 19)*, pp. 321–338, 2019.

[152] L. Zhang, Y. Meng, J. Yu, C. Xiang, B. Falk, and H. Zhu, "Voiceprint mimicry attack towards speaker verification system in smart home," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 377–386, IEEE, 2020.

[153] H. Abdullah, M. S. Rahman, W. Garcia, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor, "Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 712–729, IEEE, 2021.

[154] Z. Qin, Y. Fan, Y. Liu, L. Shen, Y. Zhang, J. Wang, and B. Wu, "Boosting the transferability of adversarial attacks with reverse adversarial perturbation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29845–29858, 2022.

[155] G. Qi, Y. Chen, Y. Zhu, B. Hui, X. Li, X. Mao, R. Zhang, and H. Xue, "Transaudio: Towards the transferable adversarial audio attack via learning contextualized perturbations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[156] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 694–711, IEEE, 2021.

[157] R. Taori, A. Kamsetty, B. Chu, and N. Vemuri, "Targeted adversarial examples for black box audio systems," in *2019 IEEE security and privacy workshops (SPW)*, pp. 15–20, IEEE, 2019.

[158] S. Khare, R. Aralikatte, and S. Mani, "Adversarial black-box attacks on automatic speech recognition systems using multi-objective evolutionary optimization," *arXiv preprint arXiv:1811.01312*, 2018.

[159] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2847–2856, 2018.

[160] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4422–4431, 2018.

[161] H. Zhu, Y. Zhu, H. Zheng, Y. Ren, and W. Jiang, "Ligaa: Generative adversarial attack method based on low-frequency information," *Computers & Security*, vol. 125, p. 103057, 2023.

[162] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15014–15023, 2022.

[163] Y. Li, S. Bai, Y. Zhou, C. Xie, Z. Zhang, and A. Yuille, "Learning transferable adversarial examples via ghost networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 11458–11465, 2020.

[164] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song, "Frequency domain model augmentation for adversarial attack," in *European Conference on Computer Vision*, pp. 549–566, Springer, 2022.

[165] X. He and Y. Zhang, "Quantifying and mitigating privacy risks of contrastive learning," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 845–863, 2021.

[166] H. Chen and J. Li, "Data poisoning attacks on cross-domain recommendation," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2177–2180, 2019.

[167] K. Chen, Y. Meng, X. Sun, S. Guo, T. Zhang, J. Li, and C. Fan, "Badpre: Task-agnostic backdoor attacks to pre-trained nlp foundation models," *ArXiv*, vol. abs/2110.02467, 2021.

[168] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

[169] A. Dziedzic, N. Dhawan, M. A. Kaleem, J. Guan, and N. Papernot, "On the difficulty of defending self-supervised learning against model extraction," in *International Conference on Machine Learning*, pp. 5757–5776, PMLR, 2022.

[170] Z. Sha, X. He, N. Yu, M. Backes, and Y. Zhang, "Can't steal? cont-steal! contrastive stealing attacks against image encoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16373–16383, 2023.

[171] T. Cong, X. He, and Y. Zhang, "Sslguard: A watermarking scheme for self-supervised learning pre-trained encoders," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 579–593, 2022.

[172] Y. Wang, J. Li, H. Liu, Y. Wang, Y. Wu, F. Huang, and R. Ji, "Black-box dissector: Towards erasing-based hard-label model stealing attack," in *European Conference on Computer Vision*, pp. 192–208, Springer, 2022.

[173] J. Beetham, N. Kardan, A. S. Mian, and M. Shah, "Dual student networks for data-free model stealing," in *The Eleventh International Conference on Learning Representations*, 2022.

[174] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno, "Physical adversarial examples for object detectors," in *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018.

[175] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1625–1634, 2018.

[176] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pp. 1989–2004, 2019.

[177] C. Xiao, D. Yang, B. Li, J. Deng, and M. Liu, "Meshadv: Adversarial meshes for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6898–6907, 2019.

[178] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, "Making an invisibility cloak: Real world adversarial attacks on object detectors," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 1–17, Springer, 2020.

[179] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15232–15241, 2021.

[180] Y. Zhang, P. H. Foroosh, and B. Gong, "Camou: Learning a vehicle camouflage for physical adversarial attack on object detections in the wild," *ICLR*, 2019.

[181] L. Huang, C. Gao, Y. Zhou, C. Xie, A. L. Yuille, C. Zou, and N. Liu, "Universal physical camouflage attacks on object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 720–729, 2020.

[182] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8565–8574, 2021.

[183] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 176–194, IEEE, 2021.

[184] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, "Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks," in *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 293–308, 2020.

[185] Y. Man, M. Li, and R. Gerdes, "{GhostImage}: Remote perception attacks against camera-based image classification systems," in *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020)*, pp. 317–332, 2020.

[186] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic, "{SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1865–1882, 2021.

[187] X. Zhu, X. Li, J. Li, Z. Wang, and X. Hu, "Fooling thermal infrared pedestrian detectors in real world using small bulbs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 3616–3624, 2021.

[188] Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, and X. Zhang, "Physical attack on monocular depth estimation with optimal adversarial patches," in *European Conference on Computer Vision*, pp. 514–532, Springer, 2022.

[189] Z. Yang, Y. Chen, Z. Sarwar, H. Schwartz, B. Y. Zhao, and H. Zheng, "Towards a general video-based keystroke inference attack," in *Proceedings of the 2023 32nd USENIX Security Symposium, Anaheim, CA, USA*, pp. 9–11, 2023.

[190] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 160–175, IEEE, 2021.

[191] W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, "Tpatch: A triggered physical adversarial patch,"

[192] Q. Jiang, X. Ji, C. Yan, Z. Xie, H. Lou, and W. Xu, "Glitchhiker: Uncovering vulnerabilities of image signal transmission with iemi," in *USENIX Security*, vol. 23, 2023.

[193] L. Wang, M. Chen, L. Lu, Z. Ba, F. Lin, and K. Ren, "Voicelistener: A training-free and universal eavesdropping attack on built-in speakers of mobile devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–22, 2023.

[194] Q. Chen, M. Chen, L. Lu, J. Yu, Y. Chen, Z. Wang, Z. Ba, F. Lin, and K. Ren, "Push the limit of adversarial example attack on speaker recognition in physical domain," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pp. 710–724, 2022.

[195] Y. Wang, H. Guo, and Q. Yan, "Ghosttalk: Interactive attack on smartphone voice system through power line," in *Network and Distributed Systems Security (NDSS) Symposium*, 2022.

[196] N. P. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad characters: Imperceptible nlp attacks," *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1987–2004, 2021.

[197] B. Liu, B. Xiao, X. Jiang, S. Cen, X. He, W. Dou, *et al.*, "Adversarial attacks on large language model-based system and mitigating strategies: A case study on chatgpt," *Security and Communication Networks*, vol. 2023, 2023.

[198] K. Mei, Z. Li, Z. Wang, Y. Zhang, and S. Ma, "Notable: Transferable backdoor attacks against prompt-based nlp models," in *Annual Meeting of the Association for Computational Linguistics*, 2023.

[199] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, A. Awadalla, P. W. Koh, D. Ippolito, K. Lee, F. Tramèr, and L. Schmidt, "Are aligned neural networks adversarially aligned?," *ArXiv*, vol. abs/2306.15447, 2023.

[200] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International conference on machine learning*, pp. 284–293, PMLR, 2018.

[201] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *arXiv preprint arXiv:1707.08945*, vol. 2, no. 3, p. 4, 2017.

[202] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, pp. 99–112, Chapman and Hall/CRC, 2018.

[203] C. Zhou, Q. Yan, Y. Shi, and L. Sun, "{DoubleStar}:{Long-Range} attack towards depth estimation based obstacle avoidance in autonomous systems," in *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1885–1902, 2022.

[204] J. Lim, T. Price, F. Monrose, and J.-M. Frahm, "Revisiting the threat space for vision-based keystroke inference attacks," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 449–461, Springer, 2020.

[205] X. Chen, C. Fu, F. Zheng, Y. Zhao, H. Li, P. Luo, and G.-J. Qi, "A unified multi-scenario attacking network for visual object tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1097–1104, 2021.

[206] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack," in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 3309–3326, 2021.

[207] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

[208] G. Wang, H. Han, S. Shan, and X. Chen, "Cross-domain face presentation attack detection via multi-domain disentangled representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6678–6687, 2020.

[209] Y. Huang, X. Li, W. Wang, T. Jiang, and Q. Zhang, "Towards cross-modal forgery detection and localization on live surveillance videos," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10, IEEE, 2021.

[210] Q. Xiao, Y. Chen, C. Shen, Y. Chen, and K. Li, "Seeing is not believing: Camouflage attacks on image scaling algorithms," in *28th USENIX Security Symposium (USENIX Security 19)*, pp. 443–460, 2019.

[211] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, "Dolphinattack: Inaudible voice commands," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 103–117, 2017.

[212] Q. Yan, K. Liu, Q. Zhou, H. Guo, and N. Zhang, "Surfingattack: Interactive hidden attack on voice assistants using ultrasonic guided waves," in *Network and Distributed Systems Security (NDSS) Symposium*, 2020.

[213] T. Sugawara, B. Cyr, S. Rampazzi, D. Genkin, and K. Fu, "Light commands:{Laser-Based} audio injection attacks on {Voice-Controllable} systems," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 2631–2648, 2020.

[214] M. E. Ahmed, I.-Y. Kwak, J. H. Huh, I. Kim, T. Oh, and H. Kim, "Void: A fast and light voice liveness detection system," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 2685–2702, 2020.

[215] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[216] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks.," in *Interspeech*, pp. 82–86, 2017.

[217] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Cross-domain replay spoofing attack detection using domain adversarial training.," in *Interspeech*, pp. 2938–2942, 2019.

[218] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Dual-adversarial domain adaptation for generalized replay attack detection.," in *Interspeech*, pp. 1086–1090, 2020.

[219] B. Chen, G. Wang, H. Guo, Y. Wang, and Q. Yan, "Understanding multi-turn toxic behaviors in open-domain chatbots," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, pp. 282–296, 2023.

[220] OpenAI, "Chatgpt," 2023. Accessed on September 6, 2023.

[221] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, *et al.*, ""so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," *International Journal of Information Management*, vol. 71, p. 102642, 2023.

[222] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath, "Deepfake text detection: Limitations and opportunities," in *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1613–1630, IEEE, 2023.

[223] J. Pu, Z. Sarwar, S. M. Abdullah, A. ur Rehman, Y. Kim, P. Bhattacharya, M. Javed, and B. Viswanath, "Deepfake text detection: Limitations and opportunities," *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 1613–1630, 2022.

[224] S. Abdelnabi and M. Fritz, "Adversarial watermarking transformer: Towards tracing text provenance with data hiding," *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121–140, 2020.

[225] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *International Conference on Machine Learning*, 2023.

[226] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu, "Dast: Data-free substitute training for adversarial attacks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 234–243, 2020.

[227] J. Zhang, B. Li, J. Xu, S. Wu, S. Ding, L. Zhang, and C. Wu, "Towards efficient data free black-box adversarial attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15115–15125, 2022.

[228] M. Shao, L. Meng, Y. Qiao, L. Zhang, and W. Zuo, "Data-free black-box attack based on diffusion model," *arXiv preprint arXiv:2307.12872*, 2023.

[229] J. Ze, X. Li, Y. Cheng, X. Ji, and W. Xu, "Ultrabd: Backdoor attack against automatic speaker verification systems via adversarial ultrasound," in *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 193–200, IEEE, 2023.

[230] G. Wang, L. Zhang, Z. Yang, and X.-Y. Li, "Socialite: Social activity mining and friend auto-labeling," in *2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC)*, pp. 1–8, IEEE, 2018.

[231] J. Wang, G. Wang, X. Zhang, L. Liu, H. Zeng, L. Xiao, Z. Cao, L. Gu, and T. Li, "Patch: A plug-in framework of non-blocking inference for distributed multimodal system," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–24, 2023.

[232] C. Li, Z. Cao, and Y. Liu, "Deep ai enabled ubiquitous wireless sensing: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–35, 2021.