A Survey on Universal Adversarial Attack

Chaoning Zhang[1*] , **Philipp Benz**[1*] , **Chenguo Lin**[2*] , **Adil Karjauv**[1] , **Jing Wu**[3] , **In So Kweon**[1]

[1]Korea Advanced Institute of Science and Technology

[2]Sichuan University

[3]University of Electronic Science and Technology of China

chaoningzhang1990@gmail.com, pbenz@kaist.ac.kr, linchenguo@stu.scu.edu.cn
mikolez@gmail.com, wujing@std.uestc.edu.cn, iskweon77@kaist.ac.kr

Abstract

The intriguing phenomenon of adversarial examples has attracted significant attention in machine learning and what might be more surprising to the community is the existence of universal adversarial perturbations (UAPs), *i.e.* a single perturbation to fool the target DNN for most images. With the focus on UAP against deep classifiers, this survey summarizes the recent progress on universal adversarial attacks, discussing the challenges from both the attack and defense sides, as well as the reason for the existence of UAP. We aim to extend this work as a **dynamic** survey that will **regularly update its content** to follow new works regarding UAP or universal attack in a wide range of domains, such as image, audio, video, text, etc. Relevant updates will be discussed at: https://bit.ly/2SbQlLG. We welcome authors of future works in this field to contact us for including your new findings.

1 Introduction

Deep neural networks (DNNs) have achieved milestone performances in numerous computer vision tasks. However, despite their success, DNNs have been discovered to be vulnerable to adversarial examples [Szegedy *et al.*, 2013], carefully crafted, quasi-imperceptible perturbations, which fool a DNN when added to an image. More interestingly, the existence of image-agnostic (universal) adversarial perturbations has been shown in recent works. A universal adversarial perturbation (UAP) is a single perturbation that is capable to fool a DNN when added to most natural images [Moosavi-Dezfooli *et al.*, 2017a]. The discovery of UAPs led to various explorations of this phenomenon, *e.g.* universal adversarial attack, the defense against UAPs as well as attempts to understand the phenomenon of UAPs. Even though UAPs have initially been studied in the domain of image classification, their exploration has expanded into other domains as well.

Scope of the survey. To this date, the amount of works on adversarial robustness is so large that it is impossible to cover them in a single survey. We refer the readers to [Akhtar and Mian, 2018] for an introduction to general adversarial attack

---

*Equal Contribution

and defense. The focus of this survey is mainly on the advancements on a special type of adversarial attack, *i.e.* universal adversarial attack, in the last few years. It is worth mentioning that image classification is the main application field where researchers design new attack and defense techniques and analyze adversarial perturbations. The core element of universal attack lies in the UAP, which can be generated beforehand and then directly applied with a simple summation operation during the attack stage. In this work, unless specified, we discuss the UAP in the context of image classification. *We highlight that this work will be extended as a dynamic survey that will update its content for including new works in this field and any feedback is welcome.*

Structure. The survey is structured as follows: First, the basic notion and notation of UAPs in the context of image-classification will be introduced. Then universal adversarial attack methods will be covered, followed by defense methods against UAPs. Afterward, an overview will be given about the different perspectives on the understanding of the UAP phenomenon. We further identify data-dependency, black-box attack capabilities, and class-discrimination as three challenges of UAPs and discuss them. Finally, works covering UAPs going beyond image-classification will be discussed.

2 A Short Primer on Image-Dependant Attack Methods

Before we get into the topic of UAP, it is relevant to discuss general image-dependent adversarial attacks since most UAP algorithms are developed based on image-dependent attacks. We categorize the adversarial attack into two groups: (a) minimizing perturbation magnitude given that the image is misclassified; (b) maximizing the attack success rate given a limited perturbation budget. Szegedy *et al.* proposed the first adversarial attack algorithm, box-constrained L-BFGS, to generate perturbations that can fool a network [Szegedy *et al.*, 2013]. This algorithm falls into the group (a). Another popular attack method is the Carlini and Wagner (C&W) attack [Carlini and Wagner, 2017]. In essence, the C&W attack is the same as the L-BFGS attack, but with a different loss, function applied. Carlini and Wagner investigate multiple loss functions and find that the loss that maximizes the gap between the target class logit and highest logit (excluding the target class logit) results in superior performance. Yet an-

other popular attack falling into this group is DeepFool that crafts perturbations iteratively by updating the gradient with respect to the model's decision boundaries. In every iteration, DeepFool chooses the perturbation direction of the minimum magnitude that is orthogonal to the decision hyperplane. With the goal of finding pseudo-minimal perturbation, group (a) has the disadvantage of being cumbersome and slow. Relatively, Group (b) that maximizes the attack success rate given a limited budget is more straightforward. The first algorithm that falls into this group is the Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2015]. FGSM is simple and fast, which comes at the cost of its effectiveness. Iterative FGSM (I-FGSM) [Kurakin *et al.*, 2017], iteratively performs the FGSM attack. In each iteration, only a fraction of the allowed noise limit is added, which contributes to its higher attack effect compared to FGSM. Another widely used white-box attack method is termed PGD introduced in [Madry *et al.*, 2018]. In essence, PGD is the same as I-FGSM and the only difference lies in that the PGD attack initializes the perturbation with random noise while I-FGSM just initializes the perturbation with zero values. This random initialization can help improve the attack success rate, especially when the number of iterations is limited to a relatively small value. Another advantage of the initialization is that it can further help improve the attack success rate with multiple trials.

3 Image-Agnostic Adversarial Attacks

3.1 Definition of UAPs in Deep Image Classifiers

The existence of UAPs to fool the deep classifier for most images has first been demonstrated in [Moosavi-Dezfooli *et al.*, 2017a], and we will mainly follow the notation introduced in their work. Given a distribution of images μ in $\mathbb{R}^d$ and a classifier function $\hat{k}$, we denote the output of the classifier given an image $x \in \mathbb{R}^d$ as $y = \hat{k}(x)$. The overall objective is to find a single perturbation vector $\nu \in \mathbb{R}^d$, such that the $\hat{k}$ is fooled for most encountered images. Additionally, ν should be sufficiently small, which is commonly modeled through an upper-bound ϵ on the l_p-norm, commonly denoted as $|| \cdot ||_p$, of the perturbation, *i.e.* $||\nu||_p \leq \epsilon$. More formally, we seek a UAP, *i.e.* ν such that:

$$\hat{k}(x + v) \neq \hat{k}(x) \text{ for most } x \sim \mu \quad \text{s.t.} \quad ||\nu||_p \leq \epsilon. \quad (1)$$

A popular choice is to set $p = \infty$, and to set the value of ϵ to $10/255$, assuming images to be in the range $[0, 1]$ [Moosavi-Dezfooli *et al.*, 2017a; Poursaeed *et al.*, 2018; Zhang *et al.*, 2020b].

3.2 Metrics to Evaluate the UAP Effectiveness

Given the above definition of UAPs, the *fooling ratio* is the most widely adopted metric for evaluating the efficacy of the generated UAP. Specifically, the *fooling ratio* is defined as the percentage of samples whose prediction changes after the UAP is applied, *i.e.* $\underset{x \sim X}{\mathbb{P}}(\hat{k}(x + \nu) \neq \hat{k}(x))$. Some works [Zhang *et al.*, 2020b; Benz *et al.*, 2020] have investigated targeted UAPs whose goal is to flip the prediction of

most samples to a pre-defined target class. The *targeted fooling ratio* is defined as $\underset{x \sim X}{\mathbb{P}}(f(x + \nu) = t)$, where t is the target label.

3.3 Universal Attack Methods

The vanilla universal attack. UAPs were first introduced in [Moosavi-Dezfooli *et al.*, 2017a]. The proposed algorithm accumulates the UAP by iteratively crafting image-dependant perturbations for the data points. Specifically, if the already accumulated perturbation does not send the current data point across the decision boundary, the minimal perturbation $\Delta\nu$ is computed to send the sample over the decision boundary. After every iteration update, the perturbation is projected on the l_p ball of radius ϵ. In the vanilla UAP algorithm the projection operator $\mathcal{P}_{p,\epsilon}$ is defined as: $\mathcal{P}_{p,\epsilon} = \arg\min_{\nu'} ||\nu - \nu'||$ subject to $||\nu'||_p \leq \epsilon$. The accumulation of minimal perturbations is repeated until the fooling rate of ν exceeds a certain threshold. The authors note that the number of encountered data points can be smaller than the number of total training points.

Generating UAPs with singular vectors (SV-UAP). A different algorithm to craft UAPs has been proposed in [Khrulkov and Oseledets, 2018]. Their method is based on the calculation of the singular vectors of the Jacobian matrices of the feature maps to obtain UAPs. The proposed approach shows a good data-efficiency, which can generate UAPs with a fooling rate of more than 60% on the ImageNet validation set by using only 64 images.

Generating UAPs with generative networks. A Network for adversary generation **(NAG)** was first introduced by [Mopuri *et al.*, 2018b]. Inspired by Generative Adversarial Networks (GAN) [Goodfellow *et al.*, 2014], NAG aims to model the distribution of UAPs. Therefore the authors modify the GAN framework by replacing the discriminator with the (frozen) target model and introduce a novel loss to train the generator. The novel loss function is composed of a fooling objective and a diversity objective. As the name suggests, the fooling objective is designed such that the generated perturbation fools the target classifier. Specifically, the loss is formulated to encourage the generator to generate perturbations that decrease the confidence of the original (benign) predictions. The diversity objective encourages the diversity of perturbations by increasing the distance of their feature embeddings predicted by the target classifier. Another variant of generative adversarial perturbations **(GAP)** using a generator to craft UAPs was also explored in [Poursaeed *et al.*, 2018]. The objective is to train a generative network that transforms a random pattern to an image-dependant perturbation or UAP. The scale operation is introduced to guarantee the perturbation lies in a certain range. Concurrent to this, the authors of [Hayes and Danezis, 2018] also explored the idea of generating adversarial perturbations with a generator network.

Dominant Feature-UAP (DF-UAP). [Zhang *et al.*, 2020b] treats the UAP as network weights and apply the DNN training techniques, such as Adam optimizer and batch training, to maximize feature content of a target class. In both non-targeted and targeted setting, the resultant UAP has domi-

Method	AlexNet	GoogleNet	VGG16	VGG19	ResNet152
UAP [Moosavi-Dezfooli *et al.*, 2017a]	93.3	78.9	78.3	77.8	84.0
SV-UAP [Khrulkov and Oseledets, 2018]	—	—	52.0	60.0	—
GAP [Poursaeed *et al.*, 2018]	-	82.7	83.7	80.1	-
NAG [Mopuri *et al.*, 2018b]	96.44	90.37	77.57	83.78	87.24
DF-UAP [Zhang *et al.*, 2020b]	96.17	88.94	94.30	94.98	90.08
Cos-UAP [Zhang *et al.*, 2021a]	96.5	90.5	97.4	96.4	90.2
FFF [Mopuri *et al.*, 2017]	80.92	56.44	47.10	43.62	-
AAA [Mopuri *et al.*, 2018c]	89.04	75.28	71.59	72.84	60.72
GD-UAP [Mopuri *et al.*, 2018a]	87.02	71.44	63.08	64.67	37.3
PD-UA [Liu *et al.*, 2019]	—	67.12	53.09	48.95	53.51
DF-UAP (COCO) [Zhang *et al.*, 2020b]	89.9	76.8	92.2	91.6	79.9
Cos-UAP (Jigsaw) [Zhang *et al.*, 2021a]	91.07	87.57	89.48	86.81	65.35

Table 1: Fooling ratio (%) of different UAP generation methods in the white-box attack scenario. The results are divided into universal attacks with access to the original ImageNet training data (upper) and data-free methods (lower).

nant features (DF). Zhang *et al.* investigate various loss functions in the context of targeted UAP generation. For the non-targeted setting, Zhang *et al.* further propose a cosine similarity based loss for alleviating the need of ground-truth labels.

A comparison between the different algorithms. The vanilla algorithm [Moosavi-Dezfooli *et al.*, 2017a] attacks a single image at once, scaling the number of iterations linearly with the number of processed images, leading to slow convergence. Moreover, their algorithm is based on the image-dependant DeepFool attack [Moosavi-Dezfooli *et al.*, 2016], which is overall found to be one of the slower attack techniques. Dai and Shu identify that minimum perturbation resulted from the DeepFool is not optimal for the efficient UAP generation. At each iteration, instead of choosing the minimal perturbation vector, they proposed to choose the perturbation that has a similar orientation to the former one. Their empirical results demonstrate that this technique can help boost both the convergence and performance, leading to an increase of 9% in fooling rate over the vanilla UAP attack. The generative networks-based approaches somewhat alleviate the rather cumbersome and slow procedure of the vanilla UAP algorithm. Adopting generative networks has the benefit that conventional training techniques can be applied to obtain a powerful UAP generation network, which overall showed superior performance over the early UAP generation methods. However, the requirement of a generative model itself is a drawback of these UAP generation approaches. The simple methods which directly update the perturbation with the calculated gradient proposed in [Zhang *et al.*, 2020b; Shafahi *et al.*, 2020; Zhang *et al.*, 2021a] demonstrate that a direct optimization of the UAP does not only remove the requirement to train a separate generative network but can also achieve superior performance. We provide an overview of different UAP generation methods in the white-box attack scenario in Table 1 supporting the here presented discussion with quantitative results.

3.4 Defending Against UAP

To mitigate the effect of adversarial perturbations, numerous works have attempted to either detect or defend through various techniques. To our knowledge, adversarial learning is the only defense method that has not been broken by strong white-box attacks [Madry *et al.*, 2018], thus it

has become the de-facto most widely used defense technique. A wide range of works [Goodfellow *et al.*, 2015; Madry *et al.*, 2018; Shafahi *et al.*, 2019b; Zhang *et al.*, 2019; Wong *et al.*, 2020] have investigated adversarial training, but the scope of these techniques is often limited to image-dependent attacks. Here, we summarize relevant advancements on defending against UAPs. One straightforward approach to extend adversarial training to the field of universal attack is to replace the image-dependent adversarial examples with the samples perturbed by the UAP during network training. The main challenge lies in the fact that an effective UAP often takes many iterations to converge, thus adversarial training against universal attacks is challenging in practice due to constraints in computation resources. Note that it can be (N+1) time slower than normal training, where N is the required number of attack iterations. To address this concern, [Moosavi-Dezfooli *et al.*, 2017a] proposes to fine-tune the model parameters with the images perturbed by pre-computed UAPs. Unfortunately, this only leads to marginal robustness enhancements against UAPs, which is somewhat reasonable because the pre-computed fixed UAP is unlike the dynamically generated perturbation for normal (image-dependent) adversarial training. Thus, the model would be expected to be only robust to the fixed perturbations. To alleviate such concern, Mummadi *et al.* have proposed to generate UAPs on-the-fly through shared adversarial training [Mummadi *et al.*, 2019]. However, it still takes 20 times more computation resources than the normal training because the UAP generation process resembles the multi-step PGD adversarial training [Madry *et al.*, 2018]. Universal adversarial training (UAT) [Shafahi *et al.*, 2020] elegantly handles this issue by concurrently updating the networks and perturbation, resulting in fast adversarial training [Wong *et al.*, 2020]. Identifying that the UAP does not attack all classes equally, a recent work [Benz *et al.*, 2021] extends the UAT with classwise perturbations, enhancing the robustness against the attack of UAP by a large margin. Moreover, it also leads to a more balanced class-wise robustness against UAP. The adversarial training on UAP has been perceived as a two-player zero-sum game [Perolat *et al.*, 2018]. Beyond adversarial training, a defense against UAPs has also been applied on the feature-level, through selective feature generation in [Borkar *et al.*, 2020]. Another framework for defending against UAP is proposed in [Akhtar *et al.*, 2018] which has two components: (a) Perturbation Rectifying Network (PRN) used as a rectifier to de-noise the UAP in the adversarial examples; (b) a binary classifier that detects adversarial examples perturbed through UAPs.

4 On the Existence of UAP

The fundamental reason that adversarial examples are intriguing to the community is that a well-trained deep classifier can be fooled by a small imperceptible perturbation. It is counter-intuitive that human invisible adversarial perturbation can fool the target model, which motivates numerous works attempting to explain its existence from a wide range of perspectives, such as the local linearity of DNNs [Goodfellow *et al.*, 2015], input high-dimensionality [Shafahi *et al.*,

2019a] and over-fitting [Schmidt *et al.*, 2018], and noise disturbance [Fawzi *et al.*, 2016]. Those explanations are limited to explain only image-dependent perturbations, in other words, they can not be easily extended to explain the image-agnostic properties of UAPs. The investigation on the existence of UAP is still in its infancy and in the following, we summarize the works in the literature on the existence of UAP. Specifically, we find that those explanations can be divided into two categories: (a) geometric perspective; (b) feature perspective.

Understanding UAPs from a geometric perspective. Moosavi-Dezfooli *et al.* have attributed the existence of UAPs to redundancies in the geometry of the decision boundaries, which was partially supported by their singular value analysis [Moosavi-Dezfooli *et al.*, 2017a]. Overall, the authors conclude that there exists a subspace of low dimension in the high-dimensional input space that contains a perturbation vector being somewhat normal to the decision boundary for most images and that the UAP algorithm exploits this to generate the UAP in that subspace. In [Moosavi-Dezfooli *et al.*, 2017b], UAPs are further interpreted from a geometry perspective in a more fine-grained manner. Specifically, the authors established two models of decision boundaries: (a) flat decision boundaries and (b) curved decision boundaries. The authors showed that *positively curved* decision boundaries provide a better explanation of the existence of UAPs, which is supported by both theoretical analysis and empirical results. Based on this understanding from the geometric perspective, the analysis in [Jetley *et al.*, 2018] has found that the predictive power and adversarial vulnerability of the studied deep classifier are intertwined, suggesting any gain in robustness must come at the cost of accuracy.

Understanding UAPs from a feature perspective. Recently, the existence of adversarial examples has been attributed to the non-robust features [Ilyas *et al.*, 2019]. Inspired by this phenomenon, Zhang *et al.* showed that the UAPs have semantically human-aligned features as shown in Figure 1. The chosen target class is "sea lion", and the human observer can identify the pattern that looks like a sea lion. Specifically, the authors have analyzed the mutual influence of images and perturbations through a Pearson Correlation Coefficient (PCC) analysis and found that UAPs dominate over the images for the model prediction, which is not the case for image-dependent perturbation. As a result, the UAPs can be seen to have independent semantic features and the images behave like noise to them. This is somewhat contrary to the popular belief to only perceive the perturbation as noise to the images. This feature perspective inspires a much more simple yet effective algorithm as discussed in Sec 3.3. Zhang *et al.* have further analyzed the reason why UAP of small magnitude can dominate over images through an investigation of the universal adversarial attack and universal deep hiding [Zhang *et al.*, 2020c]. Frequency is a key factor for the success of both tasks and the reason can be attributed to DNNs being highly sensitive to high-frequency features [Zhang *et al.*, 2021b].
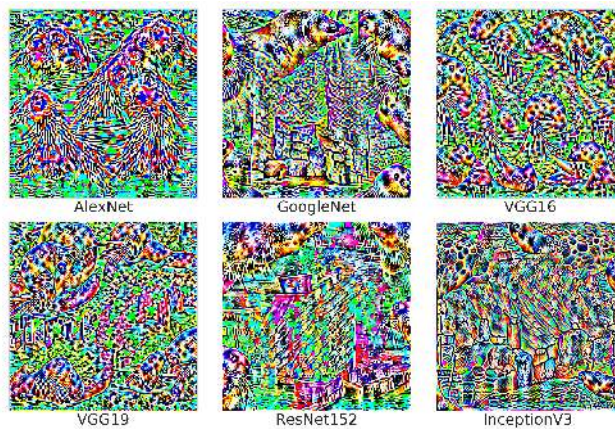

Figure 1: Targeted universal perturbations (target class "sea lion") generated with DF-UAP for different network architectures.

5 Challenges of UAP Attack

The above discussed UAP algorithms deal with the basic task of universal attacks without taking some underlying challenges into account. Here, we identify three challenges of universal attacks. First, it is not reasonable to get access to the original training dataset that is used for training the target model, thus it is desirable to generate UAPs without the dependence on the original training dataset, *i.e.* data-free UAPs. Second, in practice, access to the target model weights might not be available. Instead, a substitute model that is trained on the same or similar training dataset or the ability to only query the target model might be possible. Thus, black-box universal attacks play also an important role in the area of universal attacks. Lastly, the UAP attacks all images without discrimination, thus causing a serious threat to security-sensitive applications, such as autonomous driving. However, in practice, such an attack also easily catches the attention of the users because samples from all classes are misclassified. Thus, a UAP that can attack the samples in a class-discriminative manner can be more stealthy and might present a more dangerous threat. In the following, we will discuss each of these directions.

5.1 Data-Free UAPs

Despite early works observing that only a subset of the initial training dataset is sufficient to craft UAPs, a data dependence still persists. The issue of crafting data-free UAPs has been addressed by several works. The earliest data-free method was Fast Feature Fool (FFF) [Mopuri *et al.*, 2017], which generates data-free UAPs by introducing a loss function that maximizes the activations at each layer. This over-firing of the neurons in order to deteriorate the extracted features has been further explored in an extension of their work. The authors demonstrate that the generalizable data-free objective for UAPs (GD-UAP) [Mopuri *et al.*, 2018a] can generalize across multiple vision tasks, in particular, image recognition, image segmentation, and depth estimation. Similar to FFF, the work by [Sam *et al.*, 2019] also aims to maximize a certain activation. Specifically, the authors introduce the dilate loss,

which maximizes the Euclidean norm of the activation vector before the ReLU layer. We term this approach UAP with dilated loss (UAP-DL). Mopuri *et al.* craft data-free UAPs by leveraging class-impressions [Mopuri *et al.*, 2018c]. They generate data-free UAPs in a two-stage process. First, class impressions are generated as a substitute for the actual data samples. Class-impressions are generic representations of an object category learned by a model in the input space. Starting from a noisy image, class-impressions are obtained by updating the input noise with the objective to maximize the confidence of the desired class category. In the second stage, a generator is trained to craft UAPs which fool the classifier when added to the class-impression images. During inference, the generated UAP can then be applied to the real images. Instead of generating class impressions, the authors in [Zhang *et al.*, 2020b] leverage a random proxy dataset, different from the original training dataset, to craft UAPs. The authors motivated the use of proxy datasets by their finding that images only behave like noise to the perturbation. To alleviate the need of real images, [Zhang *et al.*, 2021a] further extends by applying jigsaw images to replace proxy dataset.

5.2 Black-Box UAPs

One property of adversarial examples is their transferability, meaning that a perturbation crafted for a source model is also capable of attacking another, unseen model. This is also called a black-box attack since no knowledge about the target model is assumed. The transferability property emphasizes the threat of adversarial examples for the application of Deep Neural Networks in security-critical applications. Transferability is a very active research field for image-dependant attacks. Very few works on UAPs solely focus on the exploration of the transferability properties of UAPs. However, a great portion of the works on UAP report the transferability capabilities of their generated UAPs. We summarize the black-box attack capabilities of a few works in Table 2. Overall Table 2 shows that in the context of UAPs it is a good rule of thumb that a higher white-box attack rate correlates with a higher black-box capability. Further, we discuss a few works that specifically aim to increase the transferability capabilities of UAPs. The authors of [Li *et al.*, 2020b] investigate the regional homogeneity of UAPs. Their finding suggests that perturbations crafted for models, which are optimized to defend against adversarial examples, show more homogeneous patterns than those crafted for naturally trained models. Therefore, the authors propose regionally homogeneous perturbations (RHP) and showcase their effectiveness against defense models in the transfer setting. To achieve more transferable UAPs (T-UAP) Hashemi *et al.* introduce a new loss that focuses on the adversarial energy in the first layer of source models to work together with the widely used cross-entropy loss to improve its transferability on the target models [Hashemi *et al.*, 2020]. Naseer *et al.* take transferability a step further and shows the existence of domain-invariant adversaries [Naseer *et al.*, 2019]. The authors show that adversaries learned on Paintings, Cartoons or Medical Images can successfully perturb ImageNet samples to fool the target classifier. A Decision-based UAP was introduced by [Wu *et al.*, 2020]. Their decision-based universal adversarial attack

(DUAttack) has no access to the internal information of the target models. DUAttack only has access to the hard-label returned by the target models. It utilizes the final inferred label to guide the direction of the perturbation. Specifically, to craft a perturbation with a stripe texture, they apply the orthogonal matrix and iteratively switch the rows of the matrix to determine the location of where the alteration should be applied. Besides, to avoid the altered pixels to offset each other, DUAttack extends its approach with a momentum term, which is commonly used in deep learning, helping to reduce the number of queries. The majority of these works discuss the transferability of UAPs in the non-targeted context, meaning that the attack is considered successful if misclassification on the black-box model is achieved. The targeted black-box attack, in which the samples have to be misclassified toward a specific target class is a much more challenging attack scenario and is rarely discussed. Due to the problem setup of UAPs, targeted UAPs can also only be discussed in a more limited scenario, where only one target class can be chosen for all samples since it is unlikely that a single perturbation will be capable to misclassify different samples toward different target classes. The work by Zhang *et al.* also considers the data-free targeted UAP case. With relatively low targeted fooling ratios for most networks, but interestingly higher targeted fooling ratios for models from the same architecture families, it emphasizes the difficulty of this attack scenario. Since their attack method can be categorized as a data-free universal attack method, it is considered as the first work to achieve data-free targeted UAPs [Zhang *et al.*, 2020b]. Improving its performance in the black-box scenario would be an interesting future direction.

5.3 Class-Discriminative UAPs (CD-UAPs)

All of the previously introduced attack methods attack samples from all classes. The authors of [Zhang *et al.*, 2020a] argue that this obvious misbehavior caused by UAPs might be suspicious to an observer. The authors investigate whether such a UAP exists that only attacks samples from a few classes while limiting the adversarial influence on the remaining classes. Such CD-UAPs would then raise less suspicion since the system under attack would only misbehave when a specific sample from a targeted class would be encountered. By combining separated loss terms for the samples from the non-targeted and targeted samples the authors successfully demonstrate a CD-UAP achieving class discrimination. In an extension to this work, the same group of authors further extends the CD-UAP to a targeted version. The objective of the introduced Double Targeted Attack (DTA) [Benz *et al.*, 2020] is to craft a single perturbation to fool samples of a specific class toward a pre-defined targeted class. Class-wise UAPs have also been explored in [Gupta *et al.*, 2019]. Gupta *et al.* propose a data-independent approach to craft CD-UAPs, by exploiting the linearity of the decision boundaries of deep neural networks.

6 Universal Attack Beyond Classification

The universal attack against deep classifier has been extended from the image domain to video domain. Li *et al.* introduce

Method	VGG-F	CaffeNet	GoogleNet	VGG-16	VGG-19	ResNet152
UAP [Moosavi-Dezfooli *et al.*, 2017a]	64.0	57.2	53.6	73.5	77.8	58.0
NAG [Mopuri *et al.*, 2018b]	67.8	67.6	74.5	80.6	83.8	65.4
T-UAP [Hashemi *et al.*, 2020]	-	-	-	84.7	94.0	36.4
FFF [Mopuri *et al.*, 2017]	39.9	38.0	30.7	38.2	43.6	26.3
GD-UAP [Mopuri *et al.*, 2018a]	49.1	53.5	40.9	55.7	64.7	35.8
UAP-DL [Sam *et al.*, 2019]	-	-	33.7	47.5	52.0	30.4
AAA [Mopuri *et al.*, 2018c]	62.5	59.6	68.8	69.5	72.8	51.7
DF-UAP [Zhang *et al.*, 2020b]	-	53.7	39.8	83.4	92.5	35.4

Table 2: Fooling ratio (%) of various transfer-based attack methods with VGG-19 as the source model. The results are divided into universal attacks with access to the original ImageNet training data (upper) and data-free methods (lower).

the first UAP against video recognition [Li *et al.*, 2018]. Chen *et al.* introduce a new variant of a universal attack on videos by appending multiple dummy frames to an arbitrary video clip [Chen *et al.*, 2019]. We briefly summarize the universal attack in applications beyond image (video) classification.

6.1 Beyond Classification in the Image Domain

Hendrik Metzen *et al.* explore how to exploit universal adversarial perturbations against semantic segmentation [Hendrik Metzen *et al.*, 2017]. The authors proposed two successful methods for the attack: the first method is to teach a network to output a desired target segmentation map regardless of the input image; the second method aims to remove target classes from the resulting segmentation map leaving other parts unchanged. Mopuri *et al.* show that their proposed data-free GD-UAP also attacks semantic segmentation effectively [Mopuri *et al.*, 2018a]. Additionally, the success of GD-UAP for attacking depth estimation is demonstrated. Li *et al.* proposed a universal adversarial perturbation technique against the image-retrieval task [Li *et al.*, 2019]. The main idea is to attack the point-wise, pair-wise, and list-wise neighborhood relationships. In addition, a coarse-to-fine distillation strategy is also introduced for the black-box attack. Despite good performance on standard benchmarks, the method also extends to real-world systems such as Google Images.

6.2 Text Classification

Wallace *et al.* have introduced universal adversarial triggers for attacking and analyzing natural language processing (NLP) [Wallace *et al.*, 2019]. Universal adversarial triggers are defined in [Wallace *et al.*, 2019] as input-agnostic sequences of tokens that can be concatenated to any input from a dataset and consequently result in a specific prediction. Behjati *et al.* introduce for UAP against text classifier. The UAP for text is defined as a sequence of words that can be added to any input sentence in order and leads to a significant accuracy drop for the text classifier[Behjati *et al.*, 2019]. The existence of a universal yet small perturbation vector in the embedding space that causes natural text to be misclassified is discovered in [Gao and Oates, 2019]. Unlike images for which a UAP of fixed size can be found, the length of the text can change. Thus, the "universality" has been defined as "token-agnostic". Specifically, they apply a single perturbation to each token, resulting in different perturbations of flexible sizes at the sequence level. The methods introduced in [Wallace *et al.*, 2019; Behjati *et al.*, 2019;

Gao and Oates, 2019] for attacking text classifier are successful. However, the generated sequence of words do not carry semantic meaning thus can be easily detected by the human. To overcome this drawback, Song *et al.* leverage an adversarially regularized autoencoder (ARAE) for generating natural English phrases that can confuse the text classifier.

6.3 Audio Classification

The existence of UAPs that can fool audio classification architectures for tasks such as speech commands, has been demonstrated in some co-occurring works [Vadillo and Santana, 2019; Neekhara *et al.*, 2019]. The algorithms adopted in [Vadillo and Santana, 2019; Neekhara *et al.*, 2019] resemble each other and are inspired by the DeepFool based vanilla UAP algorithm [Moosavi-Dezfooli *et al.*, 2017a]. Due to the reasons discussed abovr, such algorithms are often cumbersome and slow. In [Xie *et al.*, 2020a; Li *et al.*, 2020a] UAPs are generated for audio classifier based on generative networks. For example, Xie *et al.* adopt a Wave-U-Net based fast audio adversarial perturbation generator (FAPG). To improve the robustness of the generated UAP against audio, Xie *et al.* propose to adopt an acoustic room simulator to estimate the sound distortions [Xie *et al.*, 2020b]. Their results show that the proposed acoustic room simulator significantly improves the performance of the UAP. The efficacy of their approach has been demonstrated on a public dataset of 109 speakers. Overall, we find that the research in the audio domain is highly influenced by the algorithms developed in the image domain, which is expected because most of the early researches on UAP is exclusively done on the image domain.

7 Conclusion

With a focus on image classification, this survey discusses the recent progress of UAPs for both attack and defense as well as the reason for the existence of UAPs. Additionally, this survey identifies data-dependency, black-box attack, and class-discrimination as three challenges for UAPs and discusses them. This survey also summarizes universal adversarial attacks in a wide range of applications beyond image classification. Overall, the topic of UAP is a fast-evolving field, and our survey can serve as a solid basis for future researches in this field. We believe a joint investigation with data hiding as done in [Zhang *et al.*, 2021b] might be an interesting future direction for providing deeper insight.

References

[Akhtar and Mian, 2018] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 2018.

[Akhtar *et al.*, 2018] Naveed Akhtar, Jian Liu, and Ajmal Mian. Defense against universal adversarial perturbations. In *CVPR*, 2018.

[Behjati *et al.*, 2019] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. Universal adversarial attacks on text classifiers. In *ICASSP*. IEEE, 2019.

[Benz *et al.*, 2020] Philipp Benz, Chaoning Zhang, Tooba Imtiaz, and In So Kweon. Double targeted universal adversarial perturbations. In *ACCV*, 2020.

[Benz *et al.*, 2021] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Universal adversarial training with class-wise perturbations. *ICME*, 2021.

[Borkar *et al.*, 2020] Tejas Borkar, Felix Heide, and Lina Karam. Defending against universal attacks through selective feature regeneration. In *CVPR*, 2020.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

[Chen *et al.*, 2019] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Qi Tian. Appending adversarial frames for universal video attack. *arXiv preprint arXiv:1912.04538*, 2019.

[Dai and Shu, 2019] Jiazhu Dai and Le Shu. Fast-uap: Algorithm for speeding up universal adversarial perturbation generation with orientation of perturbation vectors. *arXiv preprint arXiv:1911.01172*, 2019.

[Fawzi *et al.*, 2016] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *NeurIPS*, 2016.

[Gao and Oates, 2019] Hang Gao and Tim Oates. Universal adversarial perturbation for text classification. *arXiv preprint arXiv:1910.04618*, 2019.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[Goodfellow *et al.*, 2015] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[Gupta *et al.*, 2019] Tejus Gupta, Abhishek Sinha, Nupur Kumari, Mayank Singh, and Balaji Krishnamurthy. A method for computing class-wise universal adversarial perturbations. *arXiv preprint arXiv:1912.00466*, 2019.

[Hashemi *et al.*, 2020] Atiye Sadat Hashemi, Andreas Bär, Saeed Mozaffari, and Tim Fingscheidt. Transferable universal adversarial perturbations using generative models. *arXiv preprint arXiv:2010.14919*, 2020.

[Hayes and Danezis, 2018] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *IEEE Security and Privacy Workshops (SPW)*, 2018.

[Hendrik Metzen *et al.*, 2017] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, 2017.

[Ilyas *et al.*, 2019] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *NeurIPS*, 2019.

[Jetley *et al.*, 2018] Saumya Jetley, Nicholas Lord, and Philip Torr. With friends like these, who needs adversaries? In *NeurIPS*, 2018.

[Khrulkov and Oseledets, 2018] Valentin Khrulkov and Ivan Oseledets. Art of singular vectors and universal adversarial perturbations. In *CVPR*, 2018.

[Kurakin *et al.*, 2017] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

[Li *et al.*, 2018] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*, 2018.

[Li *et al.*, 2019] Jie Li, Rongrong Ji, Hong Liu, Xiaopeng Hong, Yue Gao, and Qi Tian. Universal perturbation attack against image retrieval. In *ICCV*, 2019.

[Li *et al.*, 2020a] Jiguo Li, Xinfeng Zhang, Chuanmin Jia, Jizheng Xu, Li Zhang, Yue Wang, Siwei Ma, and Wen Gao. Universal adversarial perturbations generative network for speaker recognition. In *ICME*, 2020.

[Li *et al.*, 2020b] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan L Yuille. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses. In *ECCV*, 2020.

[Liu *et al.*, 2019] Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, and Feiyue Huang. Universal adversarial perturbation via prior driven uncertainty approximation. In *ICCV*, 2019.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[Moosavi-Dezfooli *et al.*, 2016] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[Moosavi-Dezfooli *et al.*, 2017a] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017.

[Moosavi-Dezfooli *et al.*, 2017b] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard, and Stefano Soatto. Analysis of universal adversarial perturbations. *arXiv preprint arXiv:1705.09554*, 2017.

[Mopuri *et al.*, 2017] Konda Reddy Mopuri, Utsav Garg, and R. Venkatesh Babu. Fast feature fool: A data independent approach to universal adversarial perturbations. In *BMVC*, 2017.

[Mopuri *et al.*, 2018a] Konda Reddy Mopuri, Aditya Ganeshan, and Venkatesh Babu Radhakrishnan. Generalizable data-free objective for crafting universal adversarial perturbations. *TPAMI*, 2018.

[Mopuri *et al.*, 2018b] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R. Venkatesh Babu. Nag: Network for adversary generation. In *CVPR*, 2018.

[Mopuri *et al.*, 2018c] Konda Reddy Mopuri, Phani Krishna Uppala, and R. Venkatesh Babu. Ask, acquire, and attack: Data-free uap generation using class impressions. In *ECCV*, 2018.

[Mummadi *et al.*, 2019] Chaithanya Kumar Mummadi, Thomas Brox, and Jan Hendrik Metzen. Defending against universal perturbations with shared adversarial training. In *ICCV*, 2019.

[Naseer *et al.*, 2019] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. In *NeurIPS*, 2019.

[Neekhara *et al.*, 2019] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. Universal adversarial perturbations for speech recognition systems. *arXiv preprint arXiv:1905.03828*, 2019.

[Perolat *et al.*, 2018] Julien Perolat, Mateusz Malinowski, Bilal Piot, and Olivier Pietquin. Playing the game of universal adversarial perturbations. *arXiv preprint arXiv:1809.07802*, 2018.

[Poursaeed *et al.*, 2018] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *CVPR*, 2018.

[Sam *et al.*, 2019] Deepak Babu Sam, KA Sudharsan, Venkatesh Babu Radhakrishnan, et al. Crafting data-free universal adversaries with dilate loss. 2019.

[Schmidt *et al.*, 2018] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *NeurIPS*, 2018.

[Shafahi *et al.*, 2019a] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *ICLR*, 2019.

[Shafahi *et al.*, 2019b] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *NeurIPS*, 2019.

[Shafahi *et al.*, 2020] Ali Shafahi, Mahyar Najibi, Zheng Xu, John P Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. In *AAAI*, 2020.

[Song *et al.*, 2020] Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. Universal adversarial attacks with natural triggers for text classification. *arXiv preprint arXiv:2005.00174*, 2020.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Vadillo and Santana, 2019] Jon Vadillo and Roberto Santana. Universal adversarial examples in speech command classification. *arXiv preprint arXiv:1911.10182*, 2019.

[Wallace *et al.*, 2019] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*, 2019.

[Wong *et al.*, 2020] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *ICLR*, 2020.

[Wu *et al.*, 2020] Jing Wu, Mingyi Zhou, Shuaicheng Liu, Yipeng Liu, and Ce Zhu. Decision-based universal adversarial attack. *arXiv preprint arXiv:2009.07024*, 2020.

[Xie *et al.*, 2020a] Yi Xie, Zhuohang Li, Cong Shi, Jian Liu, Yingying Chen, and Bo Yuan. Enabling fast and universal audio adversarial attack using generative model. *arXiv preprint arXiv:2004.12261*, 2020.

[Xie *et al.*, 2020b] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan. Real-time, universal, and robust adversarial attacks against speaker recognition systems. In *ICASSP*. IEEE, 2020.

[Zhang *et al.*, 2019] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, 2019.

[Zhang *et al.*, 2020a] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Cd-uap: Class discriminative universal adversarial perturbation. In *AAAI*, 2020.

[Zhang *et al.*, 2020b] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In-So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *CVPR*, 2020.

[Zhang *et al.*, 2020c] Chaoning Zhang, Philipp Benz, Adil Karjauv, Geng Sun, and In Kweon. Udh: Universal deep hiding for steganography, watermarking, and light field messaging. *NeurIPS*, 2020.

[Zhang *et al.*, 2021a] Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, and In So Kweon. Towards data-free universal adversarial perturbations with artificial images. *RobustML workshop at ICLR2021*, 2021.

[Zhang *et al.*, 2021b] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Universal adversarial perturbations through the lens of deep steganography: Towards a fourier perspective. *AAAI*, 2021.