Universal Spectral Adversarial Attacks for Deformable Shapes

Arianna Rampini
Sapienza University of Rome
rampini@di.uniroma1.it

Franco Pestarini
Sapienza University of Rome
pestarini.1855627@studenti.uniroma1.it

Luca Cosmo
Sapienza University of Rome
cosmo@di.uniroma1.it

Simone Melzi
Sapienza University of Rome
melzi@di.uniroma1.it

Emanuele Rodolà
Sapienza University of Rome
rodola@di.uniroma1.it

Abstract

Machine learning models are known to be vulnerable to adversarial attacks, namely perturbations of the data that lead to wrong predictions despite being imperceptible. However, the existence of "universal" attacks (i.e., unique perturbations that transfer across different data points) has only been demonstrated for images to date. Part of the reason lies in the lack of a common domain, for geometric data such as graphs, meshes, and point clouds, where a universal perturbation can be defined. In this paper, we offer a change in perspective and demonstrate the existence of universal attacks for geometric data (shapes). We introduce a computational procedure that operates entirely in the spectral domain, where the attacks take the form of small perturbations to short eigenvalue sequences; the resulting geometry is then synthesized via shape-from-spectrum recovery. Our attacks are universal, in that they transfer across different shapes, different representations (meshes and point clouds), and generalize to previously unseen data.

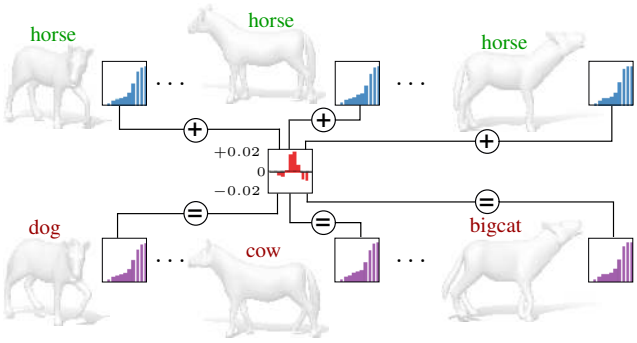


Figure 1: Universal spectral attacks on 3D horses from the SMAL dataset (only 3 out of 10 shapes are visualized). *Top row*: Original shapes with their first ten Laplacian eigenvalues and the correct labels predicted by a state-of-the-art classifier. The shapes undergo pose deformations, and have different scale, orientation and location in 3D space. *Middle row*: A universal perturbation is applied to the eigenvalues. *Bottom row*: The resulting shape embeddings synthesized from the perturbed spectra, which are now assigned wrong labels by the classifier.

1. Introduction

As machine learning methods become more and more pervasive, so their vulnerabilities are becoming more exposed. In recent years, it has been extensively shown that classifiers are susceptible to so-called adversarial attacks, i.e., misclassifications induced by feeding carefully perturbed data (*adversarial examples*) into the trained model. Adversarial examples can be crafted for image, graph, point cloud, and mesh data, as demonstrated by a thriving stream of research output across the computer vision, geometry processing, and machine learning communities.

Remarkably, *universal* perturbations are also known to exist for image data. For a given classifier $\mathcal{C}$ acting on images of size $w \times h$, a universal perturbation $P \in \mathbb{R}^{w \times h}$ is such that $\mathcal{C}(I + P) \neq \mathcal{C}(I)$ for a large number of images I; crucially, P is small under some norm $\|P\|$ so that it is hard to perceive, and is *fixed* for all such images. In other words, P is image-agnostic to some extent. This definition of universal perturbations is made possible by the fact that the operation $I + P$ can be invariably defined for all possible I and P, since all $w \times h$ images share the same grid of 2D coordinates. As soon as one shifts the focus from images to geometric data, the existence of a common space is no longer guaranteed; each individual graph G is a different domain in and by itself, and an operation of the form "$G + P$" can only be defined if G and P share the same topology. Therefore, if universality is desired, one has to define a way to transfer the perturbation P across different graphs, while at the same time ensuring it induces misclas-

sification in all cases.

In this paper, we introduce a new paradigm for universal adversarial attacks on geometric data (specifically, meshes and point clouds), in which the perturbation transfer is carried out implicitly. We do so by identifying a common domain as the space of (truncated) Laplacian spectra. This space is compact, since it only consists of short sequences of eigenvalues; it is invariant to isometric deformations (e.g., changes in pose); it only loosely depends on resolution and connectivity of the source geometry; and it is easy and efficient to compute for any given geometric object. Once a *universal* perturbation is computed in this space, the individual adversarial examples are recovered via a synthesis process that goes from eigenvalues to 3D coordinates.

2. Related work

Adversarial attacks were discovered in the seminal paper by Szegedy et al. [46], and have since been extensively explored in the image domain [16, 28, 51, 43, 23, 57, 19, 54, 5, 2, 28, 42, 35, 8, 25], natural language processing [14, 6, 21], and reinforcement learning [15], to name just a few. In this paper, we focus on *universal* attacks for *geometric* data, hence this section covers relevant prior work addressing the two aspects.

Adversarial attacks on geometric data. Compared to the image domain, the literature on adversarial attacks for geometric or topological data is less crowded, but is growing at a steady pace. Attacks on *graphs* are relatively more explored due to their relevance in tasks of community detection [9], plausibility and link prediction [55, 45], and classification [12, 60] among others. These attacks operate by modifying the graph topology, i.e., by adding, removing, or rewiring edge connections (see the recent survey [52] for an in-depth treatment). Our aim is different; instead of attacking the discrete structure representing the 3D shape, for example by changing its triangle connectivity, we seek for attacks that modify the 3D point coordinates. This brings us closer to attacking the underlying surface itself, independently of its specific representation, which in turn endows us with the ability to concoct attacks for both meshes and point clouds within a unified framework.

The literature on adversarial attacks for irregular *point cloud* data has also witnessed a recent growth. Works such as [26, 49, 56, 17] define the adversarial perturbations as small point shifts in 3D space, or as the addition of outlier points to the cloud to confuse the classifier. Since these sparse displacements can lead to noticeable artifacts, additional regularization terms to promote smooth perturbations were introduced in [48, 47], while in [58] the perturbation is a global rigid isometry applied to the 3D point cloud.

Works targeting *mesh* data are more scarce. In [50], the authors employ a differentiable renderer to define a perceptual loss, and generate attacks on photorealistic renderings by perturbing the shape texture and geometry. More recently, the work [30] introduced band-limited perturbations for mesh and point cloud classifiers, resulting in perturbations that are smooth by construction. Since we are also interested in smooth perturbations, we include the smoothness term of [30] into our construction as well.

None of the aforementioned methods provides a way to seek for universal perturbations, i.e., each perturbation is crafted for a given data sample independently of others, nor can these methods be trivially extended to address the more challenging, universal setting. We will clarify this statement more formally in the sequel.

Universal adversarial attacks for image classifiers were discovered by Moosavi-Dezfooli et al. [32], who introduced an iterative algorithm to compute universal perturbations over a set of input images. Since then, other approaches have been proposed to find universal perturbations using generative models [18, 37, 41], more efficient optimization schemes [44], based on patches rather than individual pixels [4], or applied to other image-based tasks different from classification [20, 33]; we refer to the recent survey [7] for additional examples. On graph structured data, universal attacks were recently considered in [53]; however, in their setting, universality is meant across different signals defined on a fixed graph, therefore their attacks do *not* transfer among different graphs. To the best of our knowledge, to date, no approaches have been proposed to address universal attacks for graphs or other geometric data such as point clouds and meshes.

For the sake of clarity, we mention here the closely related notion of *transferability* of attacks across different architectures, see [27, 34] for examples with image-based classifiers, and [17] for point clouds. This is different than universality, which is instead meant across data samples (the scope of this paper), rather than across learning models.

2.1. Contribution

Our main contributions can be summarized as follows:

- We demonstrate, for the first time, the existence of universal adversarial perturbations for non-rigid 3D geometric data;
- We introduce a computational procedure for finding such perturbations, which operates in the spectral domain, and follows an analysis–synthesis paradigm;
- We show that our attacks are universal in two ways: (i) across different shapes, and (ii) across different representations, such as meshes and point clouds;
- We show the generalization property of our attacks to previously unseen data.

3. Universal spectral perturbations

Following prior work on universal attacks, our framework assumes white-box access to a given classifier, since we backpropagate the error through its parameters (which are held fixed throughout the entire optimization). Further, we focus on *untargeted* attacks; namely, we do not specify a target class for the misclassification, but only require the classifier to change its prediction.

3.1. Problem setting & motivation

Given a pre-trained classifier $\mathcal{C}$ and a set of objects $\{X_i\}$, our objective is to find a perturbation P such that:

1. $\mathcal{C}(X_i + P) \neq \mathcal{C}(X_i)$ for most i and for a proper definition of the '$+$' operation (**universality**);
2. P is small in some sense, since it must remain unnoticed (**noticeability**).

The goals set above generalize those found in [32] to a broader setting. If the objects $\{X_i\}$ are plain images of fixed size as in [32], then the sum operation is well defined pixel-wise, since both P and the image set $\{X_i\}$ belong to the same vector space. However, if each X_i is an instance of non-flat structured data, one faces a number of issues.

Assume for simplicity that each $X_i \in \mathbb{R}^{n\times 3}$ is a 3D point cloud with n points. Following previous adversarial schemes for point clouds [26, 49, 56, 17, 48, 47], a perturbation $P \in \mathbb{R}^{n\times 3}$ can be defined as a displacement field such that $X_i + P$ is a slight modification of the point positions of X_i in 3D space. However, such a P can not be optimized to be universal, since it can not be directly added to a different shape X_j with $j \neq i$. First, the sum $X_j + P$ only makes sense if X_i and X_j have the same point ordering, or equivalently, if a dense point-to-point map is available between them. Second, even if a map is available, this type of attack can not be deformation-invariant: since a per-vertex perturbation is extrinsic by definition, it depends on the specific 3D coordinates to which it is applied (see Figure 2). Therefore, a successful per-vertex attack on mesh X_i will not remain successful if X_i is rotated or isometrically deformed, even if these transformations preserve the mesh topology.

3.2. Our approach

The main issue of the aforementioned approach is that it models perturbations as *extrinsic* quantities, i.e., which depend on the specific way in which the 3D objects are embedded into the ambient Euclidean space.

To address this issue, we propose to shift to an *intrinsic* representation. Let us be given a set of shapes $\mathcal{S} = \{X_i\}$. For each shape $X_i \in \mathcal{S}$ we define its *spectral representation* of length k as the sequence:

$$\sigma(X_i) = (\lambda_1^i, \lambda_2^i, \ldots, \lambda_k^i)\,, \tag{1}$$

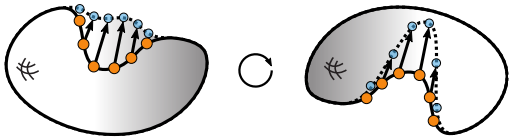


Figure 2: An extrinsic perturbation can not be universal and transformation-invariant at the same time. Even if a map is available for each pair of shapes, using it to estimate the right transformation for the perturbation is possible in the rigid case, but much harder in the general non-rigid case, and not guaranteed to lead to misclassification in both cases.

where the λ's are the first k eigenvalues of the Laplace-Beltrami operator of X_i, ordered increasingly. The Laplacian eigenvalues capture geometric information of the shape and can be computed easily; importantly, they do *not* depend on how the shape is embedded in 3D, but they are an intrinsic quantity that is invariant to non-rigid isometries.

Objective 1: Universality. We define our universal perturbation ρ to be a local solution to the following nonlinear optimization problem:

$$\min_{\substack{\rho\in\mathbb{R}^k \\ \mathcal{P}_i}} \sum_{X_i\in\mathcal{S}} \|\sigma(X_i)(1+\rho) - \sigma(\mathcal{P}_i(X_i))\|_2^2 \tag{2}$$

$$\text{s.t. } \mathcal{C}(\mathcal{P}_i(X_i)) \neq \mathcal{C}(X_i) \quad \forall X_i \in \mathcal{S} \tag{3}$$

Note that $\rho \in \mathbb{R}^k$ is universal and is an element-wise multiplicative perturbation in the *spectral* domain, while $\mathcal{P}_i(X_i)$ are shape-specific extrinsic perturbations for each shape. The constraints of Eq. (3) ensure that all the shapes in the optimization are misclassified. The spectral perturbation is multiplicative, rather than additive, so that it does not depend on the absolute scale of the eigenvalues.

To get a better grasp of what the energy of Eq. (2) enforces, we illustrate its action via the following commutative diagram:

$$\begin{array}{ccc} X_i & \xrightarrow{\sigma} & (\lambda^i) \\ {\scriptstyle \mathcal{P}_i}\downarrow & & \downarrow{\scriptstyle \rho} \\ \tilde{X}_i & \xrightarrow{\sigma} & (\tilde{\lambda}^i) \end{array}$$

The diagram expresses the fact that perturbing the eigenvalues of a given shape X_i (upper path) is equivalent to first perturbing the shape embedding itself, and then computing its eigenvalues (lower path). This is known to be true for small perturbations; a classical result of Bando and Urakawa [3] states that Laplacian eigenvalues change continuously with the surface metric, meaning that a small perturbation of the spectrum corresponds to a small perturbation of the geometry.

Both the spectral perturbation ρ and the spatial perturbations $\mathcal{P}_i(X_i)$ are unknown and must be solved for; however,

the former is shape-agnostic and fixed for all i, while the latter is shape-dependent. We seek for the set of extrinsic modifications to the geometries in $\mathcal{S}$, that simultaneously give rise to the *same* change in the eigenvalues.

Remark. The optimal ρ^* minimizing Eq. (2)-(3) is a well-defined *universal* perturbation, since it applies to all shapes in the optimization set $\mathcal{S}$. In particular, one can discard the shape-dependent $\mathcal{P}_i$'s, and verify misclassification for all $X_i \in \mathcal{S}$ by decoding $\sigma(X_i)(1+\rho^*)$ to a 3D shape (we give an algorithm in Sec. 3.4).

The remark above supports our main claim on the existence of universal adversarial perturbations for non-rigid 3D geometric data. Furthermore, as we empirically show in our experiments, the universal spectral perturbations ρ^* also exhibit *generalization* outside of the optimization set in several cases.

Objective 2: Noticeability. We model the *per-shape* perturbation $\mathcal{P}_i(X_i)$ as an extrinsic displacement $X_i + P_i$. Adversarial attacks on images explicitly impose an upper bound $\|P_i\| < \epsilon$ (see, e.g., [32]) to ensure imperceptible perturbations. Here we appeal instead to the theoretical result, also mentioned previously, that small changes in the geometry correspond to small changes in the eigenvalues [3]. This expectation is encoded in our energy of Eq. (2). Therefore, by minimizing this energy, we are also implicitly bounding the perturbation strength.

Further, we follow the smoothness principle of [30, 47, 48], which aims to ensure that each P_i is as smooth as possible. This corresponds to imposing a bound on the gradient norm $\|\nabla P_i\|$, preventing jittered perturbations. In particular, we adopt subspace parametrization [30] due to its simplicity. Each P_i is expressed as a linear combination of smooth vector fields:

$$\mathcal{P}_i(X_i) = X_i + \Phi_i \alpha_i \,, \tag{4}$$

where Φ_i is a $n \times b$ matrix whose columns are the first b Laplacian eigenfunctions of X_i (with $b \ll n$, where n is the total number of vertices), and α_i is a $b \times 3$ matrix of expansion coefficients. For smaller values of b, one gets a smoother deformation field. This band-limited representation of the displacement only requires solving for $3b \ll 3n$ coefficients per shape; furthermore, it ensures smoothness (bounded gradient) as Laplacian eigenfunctions are optimal for representing smooth functions (see [1, Th. 3.1]).

The complete optimization problem reads:

$$\min_{\substack{\rho\in\mathbb{R}^k \\ \{\alpha_i\}_i}} \sum_{X_i\in\mathcal{S}} \|\sigma(X_i)(1+\rho) - \sigma(X_i+\Phi_i\alpha_i)\|_2^2 \tag{5}$$

$$\text{s.t. } \mathcal{C}(X_i+\Phi_i\alpha_i) \neq \mathcal{C}(X_i) \quad \forall X_i \in \mathcal{S} \tag{6}$$

which involves in total k optimization variables for the spectral perturbation ρ, and $3b \cdot |\mathcal{S}|$ variables for the spatial perturbation coefficients α_i; in our tests, we typically use $b = 20$ and $k = 3b$. These numbers do *not* depend on the number of points of the shapes X_i, hence we can afford optimizing over shapes with varying resolutions.

3.3. Properties of spectral perturbations

Before moving on to the algorithmic details, we list here a few important properties of our formulation. The key idea behind this approach lies in the realization that the space of eigenvalues can serve as a convenient common domain, where different geometric data can be easily represented. The spectral domain carries important invariances that are directly inherited by our perturbations:

- We do not need an input correspondence between the shapes, nor do we have to solve for one. This also goes beyond adversarial perturbation methods for images, where one exploits the correspondence given "for free" by the canonical ordering of the pixel grid;
- Since Laplacian eigenvalues are robust against varying point density and resolution, our optimization does not require the shapes to have the same number of points or same resolution;
- Since Laplacian eigenvalues can be computed both for meshes and point clouds, spectral perturbations do not require a special treatment depending on the geometry representation.

Laplacian spectra are isometry-invariant, hence their perturbation is expected to have similar effects on isometric shapes. Similarly, since we use multiplicative perturbations, we are also invariant to scale changes of the eigenvalues, and in turn, to scale changes of the 3D shapes. We empirically confirm these properties in the experimental section.

3.4. Algorithm

We follow the general approach of Carlini and Wagner [5] to minimize problem (5), and pass to the unconstrained minimization:

$$\min_{\substack{\rho\in\mathbb{R}^k \\ \{\alpha_i\}_i}} \sum_{X_i\in\mathcal{S}} \|\sigma(X_i)(1+\rho) - \sigma(X_i+\Phi_i\alpha_i)\|_2^2 + c\mathcal{A}(X_i,\alpha_i) \tag{7}$$

where $\mathcal{A}$ is an adversarial penalty relaxing the constraints of Eq. (6), and defined as follows:

$$\mathcal{A}(X_i,\alpha_i) = \mu(Z(X_i,\alpha_i)_{\mathcal{C}(X_i)} - \max\{Z(X_i,\alpha_i)_j : j \neq \mathcal{C}(X_i)\}) \tag{8}$$

Here $Z(X_i,\alpha_i)$ is the unnormalized log-probability vector predicted by classifier $\mathcal{C}$ for the shape $(X_i+\Phi_i\alpha_i)$, and

$\mu(x) = \max(x, -m)$ is a function sending the penalty to zero once a given misclassification margin m is hit. The contribution of the adversarial penalty to the minimization problem is weighted by the trade-off parameter c.

Solving this unconstrained problem does not guarantee that all the shapes are misclassified and, in general, such a perturbation is not guaranteed to always exist. Nevertheless, in practice the optimized perturbation leads to misclassification for most of the shapes, as we show in our experiments.

Optimization. For each shape X_i, we discretize its Laplace-Beltrami operator as a positive semi-definite matrix using the classical cotangent scheme [36], whose eigenvalues and eigenfunctions can be computed with standard sparse eigensolvers. The optimization variables of Eq. (7) are optimized for with the Adam optimizer [24], which is robust to local minima. This involves computing the quantities $\sigma(X_i + \Phi_i \alpha_i)$ at each iteration, i.e., the eigenvalues of the deformed shapes, as well as their derivatives with respect to the deformation coefficients α_i. For the eigenvalue derivatives, we use the closed form expressions of Magnus [29]. Each iteration takes approximately 1s on an i7 9700k CPU (dominated by eigenvalue decomposition); for an average number of 500 iterations per optimization, the average runtime to find a universal perturbation on a set of 15 shapes is $\sim$1h.

Generalization to new samples. Once a spectral perturbation ρ is estimated for a small set of shapes, it can be applied to new shapes and still fool the classifier. This was also observed for the image-based universal attacks of [32]. However, in the image domain, applying a universal perturbation to a new data sample is a simple addition of two images. In our case, the perturbation is not additive in the spatial domain, but multiplicative in the spectral domain.

Given ρ and a new shape Y, we follow the paradigm:

$$Y \mapsto (\lambda_i)_{i=1}^k \mapsto (\lambda_i + \lambda_i \rho_i)_{i=1}^k \mapsto \tilde{Y}\,, \tag{9}$$

The first two steps are straightforward and can be easily computed. The last step requires resynthesizing the geometry from the perturbed spectrum – an inverse problem known in mathematical physics as 'hearing the shape of the drum' [22], and recently tackled by 'isospectralization' techniques in [11, 31].

To address this, we simply run the optimization of Eq. (7) without the adversarial term and with fixed ρ, which is now given. This way, we optimize only for the coefficients α, which define a smooth transformation for the geometry of Y. Optimizing over smooth geometric perturbations has a regularization effect as it greatly reduces the space of possible embeddings, making this 'isospectralization' problem easier to solve than in [11, 31]; in these works, a solution is sought from scratch over all possible point configurations in 3D, making the optimization much harder and prone to poor reconstructions, as shown with a comparison in Figure 3.

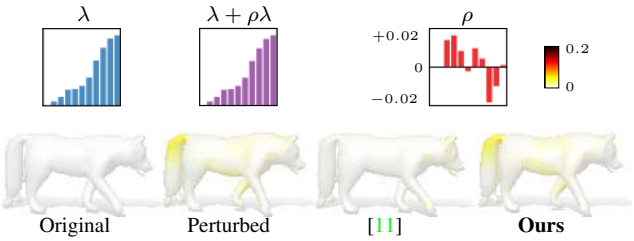


Figure 3: For a source shape $X \in \mathcal{S}$ with spectrum λ (left), we compute a perturbation ρ by optimizing Eq. (7) over $\mathcal{S}$, and obtain the Perturbed shape (second from the left, observe the foreleg movement). If we now discard the Perturbed embedding and try to recover it using [11] with $\lambda + \rho\lambda$ as a target, we get a wrong solution (third from the left) which aligns correctly the eigenvalues, but fails to recover the correct deformation. Our approach based on optimizing for a smooth deformation recovers the Perturbed shape almost exactly (rightmost).

4. Experimental evaluation

In this section we report quantitative results and show qualitative examples of universal spectral perturbations, demonstrating their efficacy and empirically confirming their main properties.

Datasets. We tested with two recent and extensive datasets of non-rigid 3D shapes: the **SMAL** dataset of 3D animals [59], and the **CoMA** dataset of human face expressions [39]. The former is composed by 600 meshes of 5 animal species in different poses, generated via a parametric model. For fair comparisons, we used the same shapes and experimental setup proposed in [30], using 480 shapes for training the classifiers, and the remaining 120 for test. The classification task assigns each shape to a specific animal category. CoMA is a 4D dataset containing sequences of 3D shapes of 13 different people performing 13 different facial expressions. We used the same train/test split proposed in [40] to train the classifiers, where the task is to classify on subject identity.

Classifiers. We perform our attacks on two different state-of-the-art classifiers for 3D shapes:

1. A convolutional mesh classifier with the architecture of [39], where the convolution is based on fast Chebyshev filters [13]. This learning model is powerful, but needs consistent meshing and correspondence at training time. We refer to this classifier as **ChebyNet**;

2. A PointNet based classifier [38]. This architecture is more general, as it is able to handle unorganized point clouds, possibly with different numbers of points. We refer to this classifier as **PointNet**.

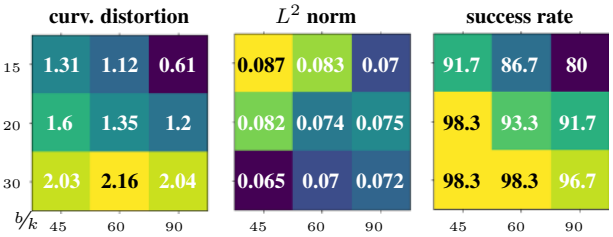


Figure 4: Sensitivity to parameters (SMAL dataset, PointNet classifier). We evaluate each quantitative measure at increasing values of $b = (15, 20, 30)$ and number of eigenvalues $k = (45, 60, 90)$. Large numbers for curvature distortion imply more noticeable perturbations. From the success rate we observe a trend: the spectral bandwidth k should not be too large, and the deformation not too smooth. However, lack of smoothness also leads to larger noticeability.

Both classifiers are trained to classify the subject identity for CoMA data, and the animal species for SMAL data, irrespective of pose. During the training phase, we augmented each dataset by randomly rotating and translating the shapes, and jittering the vertex positions.

4.1. Sensitivity to parameters

We expose two parameters: the spectral bandwidth k, that is the number of eigenvalues that undergo the spectral perturbation, and the spatial bandwidth b, that is the number of eigenfunctions to represent the spatial perturbation. Both affect the noticeability and success rate of the attack.

A small value for b leads to smoother and less noticeable perturbations, but makes it harder to find universal ones. On the other hand, large values allow for stronger deformations, but which are also more universal. This is typical of universal perturbations, which have been observed to be more noticeable than per-instance perturbations also in the image domain [32]. Parameter k expresses the degree of universality that we require from the attack, since each of the k dimensions of ρ encodes a constraint for the perturbation. A small k leads to more global deformations, leaving the attack free to apply local shape-dependent corrections. Increasing k makes it harder to obtain a successful universal attack, since a longer perturbation vector ρ imposes more geometric constraints on the attack.

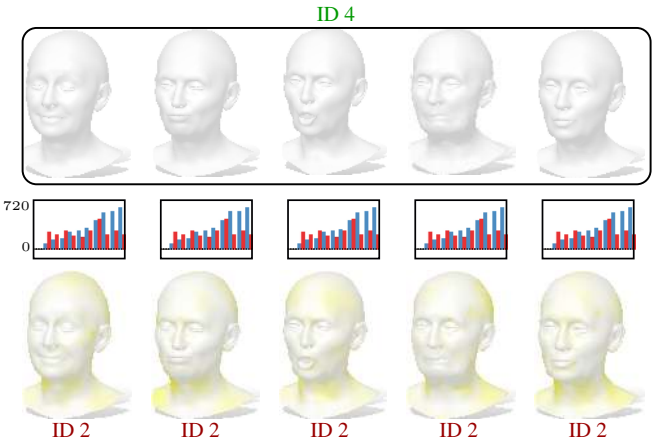


Figure 5: Examples of a universal adversarial attack on ChebyNet from a class of the CoMA dataset. Top to bottom: original shapes, barplot with their spectra in blue and their perturbation ρ in red (ρ is scaled by a factor 10^4 for visualization purposes), deformed shapes.

Table 1: Comparison between our method and non-universal approaches. We observe that a slight drop in accuracy and an increase of the deformation strength is needed, in order to gain a universal encoding of the attacks.

SMAL	success rate	curv. dist.	L^2-norm
SHAPE-DEPENDENT			
[30] (ChebyNet)	100%	2.51	3.6e-2
ChebyNet	100%	2.91	7.7e-2
UNIVERSAL			
ChebyNet	100%	2.49	9.0e-2
PointNet	93.3%	1.31	5.9e-2
CoMA	success rate	curv. dist.	L^2-norm
SHAPE-DEPENDENT			
[30] (ChebyNet)	94.3%	3.30	8.5e-3
ChebyNet	91.7%	1.61	1.7e-3
UNIVERSAL			
ChebyNet	91.7%	2.30	2.8e-3
PointNet	100%	5.71	6.9e-3

We performed a systematic study of b and k, quantifying noticeability through two deformation measures: *curvature distortion*, defined as the average absolute difference between the mean curvature at corresponding vertices in the original and perturbed shape; and L^2-norm, defined as the average Euclidean distance between corresponding vertices in the original and perturbed shape. The *success rate* is the percentage of attacks that give rise to misclassification.

Quantitative results are reported in Figure 4, revealing a trade-off between the amount of curvature distortion and the success rate. On the contrary, the L^2-norm decreases with b; this is consistent with what was shown in [30], since smoother deformations force the attack to move a bigger proportion of the shape, as the classifier can not be fooled with small local perturbations. Based on this, we claim that L^2-norm is probably not a good metric for capturing noticeability on deformable shapes, since localized deformations are usually more disturbing to the human observer. For all our experiments, we use $b = 20$ and $k = 60$ as a good trade-off between noticeability and success rate; in our plots, we only show $k = 10$ eigenvalues for visualization purposes.

4.2. Universality

We optimize problem (7) over 15 random shapes of the same class from the test set, obtaining a set of spatial coef-

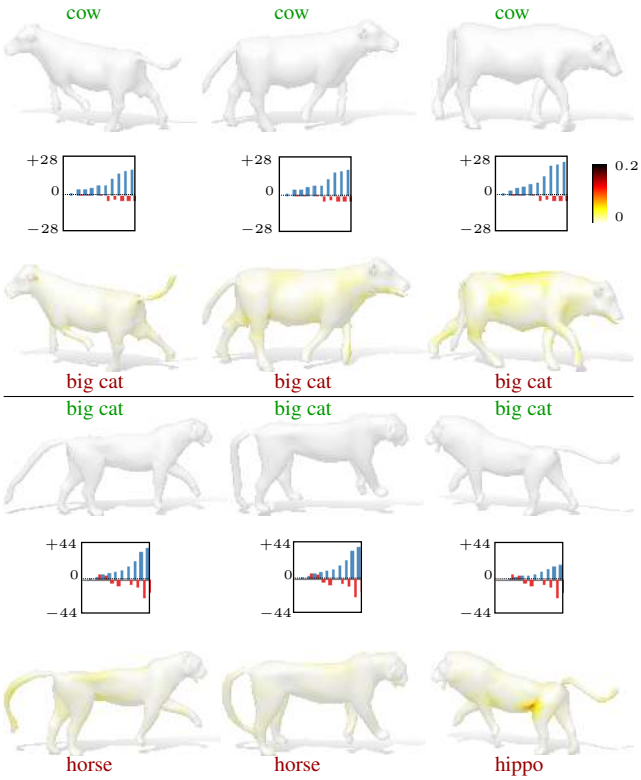


Figure 6: Examples of universal adversarial attacks on PointNet from 2 classes of the SMAL dataset (top: cows, bottom: big cats). The heatmap encodes curvature distortion, growing from white to dark red. Even if the original shapes are not isometric, as can be noted also from their spectra (blue bars), a universal spectral perturbation ρ (red bars, scaled by a factor 10^3) leads to misclassification.

ficients α_i for each shape X_i, and a universal perturbation ρ common to all the shapes. We use the same optimization parameters for all the datasets and classifiers, with $c = 5e-2$. In Table 1 we evaluate the quality of our adversarial attacks in terms of success rate and deformation strength. As we can see, the strength of the attack is inversely proportional to the success rate, further confirming the conclusions drawn in our sensitivity analysis. Table 1 also includes results obtained with our method *without* optimizing for the universal perturbation ρ in Eq. (7), i.e., we optimize for the coefficients α_i independently for each shape. As expected, this leads to a slight increase of the success rate and a less noticeable deformation. For completeness, we also report results from the state-of-the-art method [30], which uses a geometric regularizer to bound the distortion. Several qualitative examples are shown in Figures 5, 6 and 8.

4.3. Generalization

As described in Section 3.4, once we optimize for a universal perturbation ρ on a set of shapes, this can be used to transfer the deformation to a new shape.

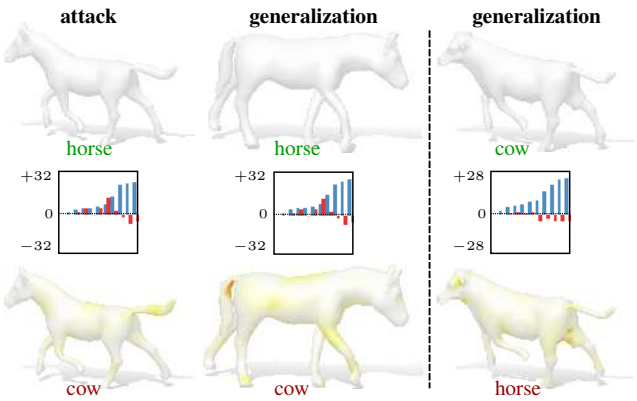
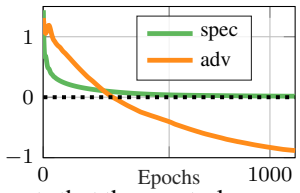


Figure 7: Generalization results on two different classes from SMAL. Left: the spectral perturbation ρ estimated for a set of 15 horses (1 shown in the first column) is applied to an unseen shape of the same class (second column) as described in the text. The resulting deformed shape (second column, bottom shape) is incorrectly classified. Right: another example from a different class. In the bar plots ρ (red) is scaled by a factor 10^3.

We do this by smoothly deforming the new shape as to make its spectrum match the target perturbation ρ. In the inset figure, we show that minimizing only the spectral term in Eq. (7) induces by itself a minimization of the adversarial penalty, bringing in turn the classifier to a wrong prediction. This suggests that the spectral energy is a strong prior for finding adversarial perturbations.



We performed a quantitative evaluation of the generalization capability on the CoMA dataset, obtaining successful adversarial attacks on unseen shapes on 80.8% of the cases for the PointNet classifier, and 49.2% with ChebyNet. Finally, in Figures 7 and 8 we show some qualitative examples on both the CoMA and SMAL datasets, showing how we are able to produce a similar deformation on new samples without requiring any correspondence.

4.4. Meshes and point clouds

One of the main advantages of working in a spectral domain is that it is agnostic to the surface representation, as long as a good approximation of the Laplacian operator can be computed on it. This property allows us to handle seamlessly shapes with different tessellation, resolution, and even surfaces represented by unorganized point clouds. In fact, the only limitation is posed by the classifier under attack, which might be representation-specific.

Triangle meshes. To prove the robustness of our attack to different tessellations, we independently remeshed each shape of the SMAL dataset to random number of vertices

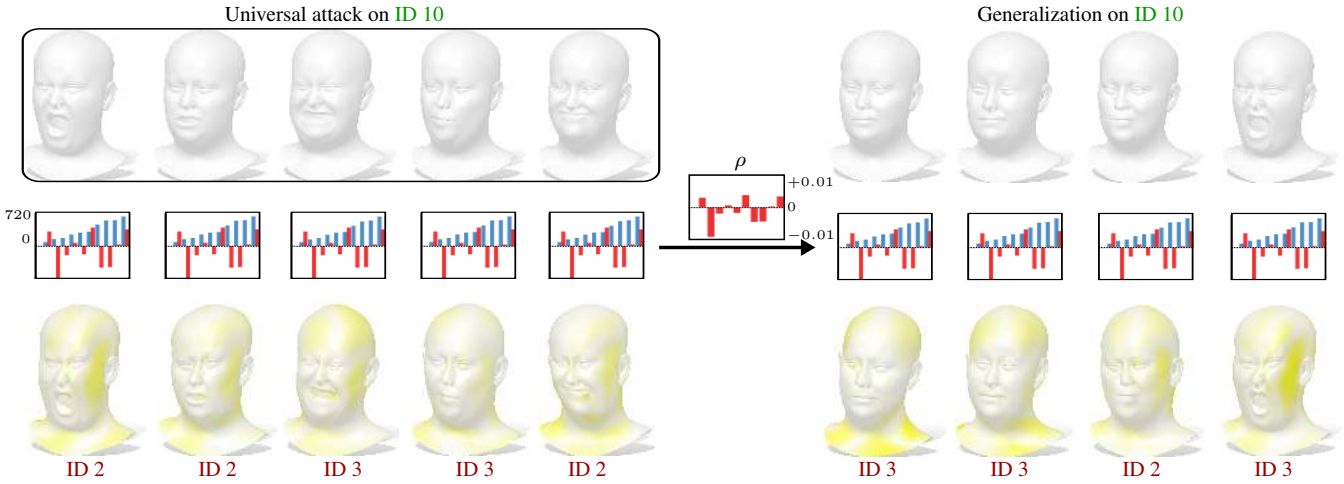


Figure 8: On the left, examples of a universal adversarial attack on 15 shapes (only 5 shown) classified with PointNet. The resulting universal spectral perturbation (middle) is used to generalize the attack to 4 new shapes of the same subject (right).

within 30% to 50% of the original ones. We then used these shapes to perform a universal adversarial attack on the PointNet classifier. Not surprisingly, we noted an increase of performance for the adversarial attack, obtaining up to 94% of success rate with an average curvature distortion of 0.80. This improvement is explainable by the reduced number of points given as input to PointNet for classification, making the adversarial attack easier to perform. Qualitative examples are shown in Figure 9.

Point clouds. In Figure 10 we show a qualitative example of generalization to point clouds. We optimized for a universal perturbation to the PointNet classifier, using 15 meshes from the CoMA test set. We then applied our generalization procedure to a point cloud derived from a new pose of the same subject. To estimate a Laplace operator for the point cloud, we used the method described in [10].

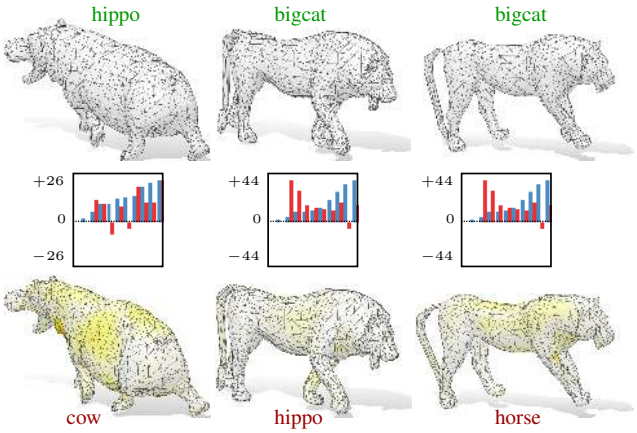


Figure 9: Results on remeshed shapes from SMAL. The first column shows the result of the universal attack on the hippo category, while the remaining two are on the bigcat category. Both are conducted on a set of 15 shapes.

5. Conclusion

We introduced a method to compute universal adversarial perturbations on 3D geometric data. The key idea lies in the adoption of the Laplacian spectrum as an intermediate shared domain for multiple shapes, where perturbations can be computed and then resynthesized into the geometry via shape-from-spectrum recovery. Operating with eigenvalues endows our attacks with robustness to deformation, sampling, and shape representation, leading in turn to generalization outside of the optimization set. Currently, the main **limitation** of this approach is its limited applicability to shapes belonging to very different classes; for example, we were not able to find successful universal perturbations for faces *and* animals simultaneously. This is due to the fact that different classes may have very different spectra; looking for an alternative, perhaps learned, representation might be a potential solution to explore in the future.

Acknowledgments

We gratefully acknowledge Luca Moschella for the technical help. This work is supported by the ERC Grant No. 802554 (SPECGEO) and the MIUR under grant "Dipartimenti di eccellenza 2018-2022".

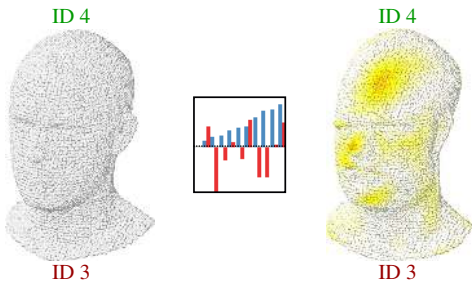


Figure 10: Example of generalization to a point cloud in CoMA using PointNet classifier.

References

[1] Yonathan Aflalo, Haim Brezis, and Ron Kimmel. On the optimality of shape and data representation in the spectral domain. *SIAM Journal on Imaging Sciences*, 8(2):1141–1160, 2015.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proc. ICML*, volume 80, pages 274–283, 2018.

[3] Shigetoshi Bando and Hajime Urakawa. Generic properties of the eigenvalue of the laplacian for compact riemannian manifolds. *Tohoku Mathematical Journal, Second Series*, 35(2):155–172, 1983.

[4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[5] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[6] Akshay Chaturvedi, Abijith KP, and Utpal Garain. Exploring the robustness of nmt systems to nonsensical inputs. *arXiv preprint arXiv:1908.01165*, 2019.

[7] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020.

[8] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, pages 15–26, New York, NY, USA, 2017. ACM.

[9] Yizheng Chen, Yacin Nadji, Athanasios Kountouras, Fabian Monrose, Roberto Perdisci, Manos Antonakakis, and Nikolaos Vasiloglou. Practical attacks against graph-based clustering. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1125–1142, 2017.

[10] Ulrich Clarenz, Martin Rumpf, and Alexandru Telea. Finite elements on point based surfaces. In *Proceedings of the First Eurographics conference on Point-Based Graphics*, pages 201–211. Eurographics Association, 2004.

[11] Luca Cosmo, Mikhail Panine, Arianna Rampini, Maks Ovsjanikov, Michael M Bronstein, and Emanuele Rodolà. Isospectralization, or how to hear shape, style, and correspondence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7529–7538, 2019.

[12] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. *arXiv preprint arXiv:1806.02371*, 2018.

[13] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3844–3852, Red Hook, NY, USA, 2016. Curran Associates Inc.

[14] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

[15] Adam Gleave, Michael Dennis, Neel Kant, Cody Wild, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.

[16] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[17] Abdullah Hamdi, Sara Rojas, Ali Thabet, and Bernard Ghanem. Advpc: Transferable adversarial perturbations on 3d point clouds. *arXiv preprint arXiv:1912.00461*, 2019.

[18] Jamie Hayes and George Danezis. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 43–49. IEEE, 2018.

[19] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2755–2764, 2017.

[21] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*, 2019.

[22] Mark Kac. Can one hear the shape of a drum? *The American Mathematical Monthly*, 73(4):1–23, 1966.

[23] Marc Khoury and Dylan Hadfield-Menell. Adversarial training with voronoi constraints. *arXiv preprint arXiv:1905.01019*, 2019.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3866–3876, 2019.

[26] Daniel Liu, Ronald Yu, and Hao Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2279–2283. IEEE, 2019.

[27] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

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

[29] Jan R Magnus. On differentiating eigenvalues and eigenvectors. *Econometric Theory*, pages 179–191, 1985.

[30] Giorgio Mariani, Luca Cosmo, Alex Bronstein, and Emanuele Rodolà. Generating adversarial surfaces via band-limited perturbations. *Computer Graphics Forum*, 39(5):253–264, 2020.

[31] Riccardo Marin, Arianna Rampini, Umberto Castellani, Emanuele Rodolà, Maks Ovsjanikov, and Simone Melzi. Instant recovery of shape from spectrum via latent space connections. In *International Conference on 3D Vision (3DV)*, 2020.

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[33] Konda Reddy Mopuri, Aditya Ganeshan, and R Venkatesh Babu. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018.

[34] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[35] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 506–519, New York, NY, USA, 2017. ACM.

[36] Ulrich Pinkall and Konrad Polthier. Computing discrete minimal surfaces and their conjugates. *Experimental mathematics*, 2(1):15–36, 1993.

[37] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018.

[38] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[39] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, 2018.

[40] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11207, pages 725–741. Springer, Cham, Sept. 2018.

[41] Konda Reddy Mopuri, Utkarsh Ojha, Utsav Garg, and R Venkatesh Babu. Nag: Network for adversary generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 742–751, 2018.

[42] Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[43] Anindya Sarkar, Nikhil Kumar Gupta, and Raghu Iyengar. Enforcing linearity in dnn succours robustness and adversarial image generation. *arXiv preprint arXiv:1910.08108*, 2019.

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

[45] Mingjie Sun, Jian Tang, Huichen Li, Bo Li, Chaowei Xiao, Yao Chen, and Dawn Song. Data poisoning attack against unsupervised node embedding methods. *arXiv preprint arXiv:1810.12881*, 2018.

[46] 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.

[47] Tzungyu Tsai, Kaichen Yang, Tsung-Yi Ho, and Yier Jin. Robust adversarial objects against deep learning models. In *Proc. AAAI*, 2020.

[48] Yuxin Wen, Jiehong Lin, Ke Chen, and Kui Jia. Geometry-aware generation of adversarial and cooperative point clouds. *arXiv preprint arXiv:1912.11171*, 2019.

[49] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2019.

[50] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, 2019.

[51] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 501–509, June 2019.

[52] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil Jain. Adversarial attacks and defenses in images, graphs and text: A review. *arXiv preprint arXiv:1909.08072*, 2019.

[53] Xiao Zang, Yi Xie, Jie Chen, and Bo Yuan. Graph universal adversarial attacks: A few bad actors ruin graph learning models. *arXiv preprint arXiv:2002.04784*, 2020.

[54] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[55] Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, and Kui Ren. Data poisoning attack against knowledge graph embedding. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4853–4859. AAAI Press, 2019.

[56] Yu Zhang, Gongbo Liang, Tawfiq Salem, and Nathan Jacobs. Defense-pointnet: Protecting pointnet against adversarial attacks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 5654–5660. IEEE, 2019.

[57] Yuchen Zhang and Percy Liang. Defending against whitebox adversarial attacks via randomized discretization. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89, pages 684–693, 2019.

[58] Yue Zhao, Yuwei Wu, Caihua Chen, and Andrew Lim. On isometry robustness of deep 3d point cloud models under adversarial attacks. *arXiv preprint arXiv:2002.12222*, 2020.

[59] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[60] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2847–2856, 2018.