360-Attack: Distortion-Aware Perturbations from Perspective-Views

Yunjian Zhang[1,2] Yanwei Liu[1,*] Jinxia Liu[3] Jingbo Miao[1,2]
Antonios Argyriou[4] Liming Wang[1] Zhen Xu[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]University of Chinese Academic of Sciences, [3]Zhejiang Wanli University, [4]University of Thessaly
{zhangyunjian, liuyanwei}@iie.ac.cn, liujinxia1969@126.com, miaojingbo@iie.ac.cn
{wangliming, xuzhen}@iie.ac.cn

Abstract

The application of deep neural networks (DNNs) on 360-degree images has achieved remarkable progress in the recent years. However, DNNs have been demonstrated to be vulnerable to well-crafted adversarial examples, which may trigger severe safety problems in the real-world applications based on 360-degree images. In this paper, we propose an adversarial attack targeting spherical images, called 360-attactk, that transfers adversarial perturbations from perspective-view (PV) images to a final adversarial spherical image. Given a target spherical image, we first represent it with a set of planar PV images, and then perform 2D attacks on them to obtain adversarial PV images. Considering the issue of the projective distortion between spherical and PV images, we propose a distortion-aware attack to reduce the negative impact of distortion on attack. Moreover, to reconstruct the final adversarial spherical image with high aggressiveness, we calculate the spherical saliency map with a novel spherical spectrum method and next propose a saliency-aware fusion strategy that merges multiple inverse perspective projections for the same position on the spherical image. Extensive experimental results show that 360-attack is effective for disturbing spherical images in the black-box setting. Our attack also proves the presence of adversarial transferability from $\mathbb{Z}^2$ *to* SO(3) groups.

1. Introduction

Previous studies have shown that deep neural networks (DNNs) are vulnerable to carefully crafted adversarial examples [10, 28, 36, 40]. Many attack algorithms have been proposed for various tasks, including image classification [16], video caption [39], 3D mesh classification [33], and point cloud recognition [43]. However, during the investigation of this issue, the security of DNNs applying to spherical images has been largely ignored.

Recently, we have observed an increasing number of computer vision problems requiring spherical signals, for instance, omnidirectional RGB-D images generated from panorama cameras [4, 17], 360-degree videos captured from sensors on self-driving cars [15, 47], and spherical data projected from the 3D domain [12]. Inspired by the remarkable success of DNNs in various tasks, many approaches have been proposed to apply DNNs on spherical images to solve real world problems, including advanced driver assistance systems (ADAS) [24], autonomous navigation [13, 22, 29], and VR/AR applications [35, 45].

DNNs used for spherical images typically operate in two domains: the spherical domain and the panoramic domain. The first type of models, referred to **spherical models**, directly dispose the spherical image in the spherical domain [7, 8, 12], while models on the planar domain, which are called **panoramic models**, operate on the panorama transformed from the spherical image [19, 34, 46].

Due to the extensive applications of spherical images, the vulnerability of DNNs used for applications around them needs further investigation. A straightforward way to generate adversarial spherical images is to attack the spherical or panoramic models in the white-box setting directly. However, these models are difficult to obtain in practice due to their greatly divergent principles, and backpropagation on them is of low efficiency, compared to the same operation on standard planar CNNs [34]. Another intractable problem of attacking the panoramic model is that the panoramas usually suffer from great distortion compared to the raw spherical images, reducing the effect of the added perturbations when the adversarial panoramas are re-projected to the original spherical domain. Therefore, an efficient attack method with less distortion is required. Considering the perspective views of a spherical image are less distorted and can be processed with a simple planar network, in this paper, we propose to generate adversarial spherical exam-

*Yanwei Liu is the corresponding author.

ples by disturbing their planar perspective-view (PV) representations. Specifically, we simultaneously disturb these PV images rendered from different positions on the spherical image, and reconstruct the adversarial spherical image from them with a re-projection and a fusion method. As our attack is implemented by transferring 2D adversarial perturbations to the 3D space without any knowledge about the target model, it can be considered as a black-box attack.

Overall, our contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose a black-box attack towards spherical models, called 360-atack, by generating adversarial spherical images from their corresponding PV images. 360-attack is performed directly on the planar domain, and eventually the perturbations are transferred to the spherical images.
- To obtain highly transferable adversarial PV images toward attacking the spherical model, we proposed a novel Distortion-Aware Iterative Fast Gradient Sign Method (DAI-FGSM) with considering the perturbation degradation caused by plane-to-sphere projection distortion. Accordingly, the negative effect of the projective distortion on the attack is alleviated.
- We propose a novel spherical-spectrum-based saliency detection method, and then propose a saliency-aware fusion strategy to merge multiple inverse perspective projections for the same position for generating the final adversarial spherical images.
- Extensive experiments on the synthetic and real-world datasets demonstrate the effectiveness of our 360-attack on DNNs designed for spherical images, and it also proves that the adversarial perturbations can be transferred from $\mathbb{Z}^2$ to SO(3) groups. For allowing result reproduction, we have submitted our source code in the supplementary materials.

2. Related Work

2.1. Adversarial Attacks

Since Szegedy et al. [36] first reported the existence of adversarial examples in DNNs, various attacks have been proposed. The first type of attacks are white-box attacks, which generate adversarial examples with full knowledge of the target model. Goodfellow et al. [16] proposed the Fast Gradient Sign Method (FGSM) that directly generates adversarial examples by calculating the gradients of the loss function for the input image. After that, multiple iterative versions of FGSM are proposed, including the Basic Iterative Method (BIM) [23], Projected Gradient Descent (PGD) [26], Momentum Iterative FGSM (MI-FGSM) [11], and Nesterov Accelerate Gradient Method (NI-FGSM) [25].

Different from the white-box attacks, black-box methods aim to attack DNNs without any knowledge regarding their inner workings. Generally, black-box attacks can be divided into three categories: score-based, decision-based, and transfer-based attacks. Score-based attacks assume that the attackers can query the prediction probability of the target model. These methods usually rely on sampling methods to approximate the gradients to generate adversarial examples [5, 21]. In decision-based attacks, the only allowed operation for the attackers is to query the output labels of given samples from the target network. The Boundary attack [2] and its variants [6, 9] are feasible in this setting. Instead of directly attacking the target model, transfer-based attacks make use of the fact that adversarial examples have high transferability across different models [20, 37, 38, 41]. Specifically, they generate adversarial perturbations on a white-box model and then transfer them to the unknown target network. Moreover, it can overcome the gradient mask defenses [1, 30] deployed on the target network. In this paper, we further investigate the transferability of adversarial examples, and the results demonstrate that this property also exists across different representation spaces, such as the planar space and the spherical space.

2.2. DNNs for Spherical Images

Planar DNNs are difficult to be applied directly on spherical images because the underlying projection models and data formats of planar and spherical images are different. To address this discrepancy, there are two types of methods. In the first, the spherical images are projected to panoramas in the planar $\mathbb{Z}^2$ space, and then planar DNN models are applied to them [13, 19, 34, 46]. However, this projection introduces significant distortion, making the convolution results inaccurate. More recently, spherical CNNs that directly handle the spherical images have been presented [8, 12, 22]. In these schemes, the domain space is transformed to the three-dimensional rotated group (SO(3)), and the rotation-equivariant convolution is implemented by using a generalized Fast Fourier Transform algorithm.

3. The Proposed 360-Attack

3.1. Overview of the Framework

The pipeline of 360-attack is shown in Fig. 1. First, multiple PV images are rendered from different positions on the sphere. Next, we attack a planar DNN to obtain adversarial PV images. Note the projection between spherical and PV images introduces distortion [48], and interferes with the effect of the perturbations in PV images. Therefore, traditional 2D attack methods are not very efficient in this scenario because they do not consider the characteristic of the projection distortion and treat all pixels equally on the PV image to be non-distorted corresponding to the spherical counterpart. In 360-attack, adversarial PV images are generated by a novel distortion-aware method that conquers the projection distortion. After that, the adversarial

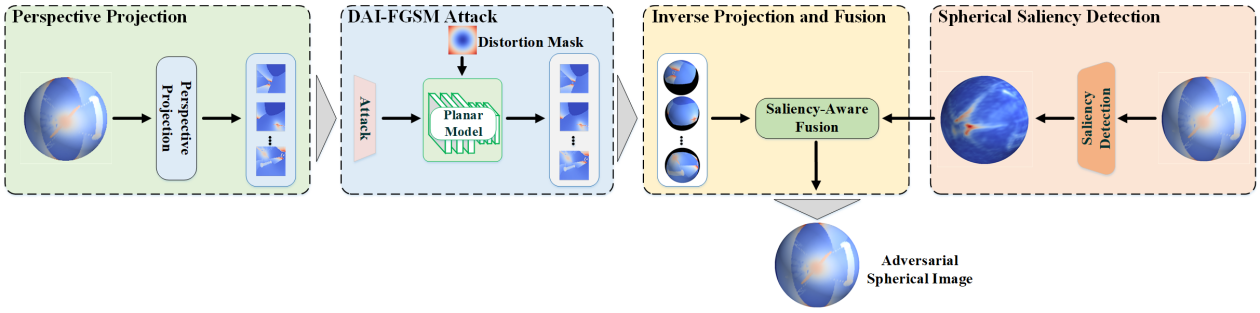


Figure 1. The pipeline of the proposed 360-attack.

PV images are projected to the sphere. As the inversely projected spherical areas for multiple PV images have overlaps with each other, we merge them ultimately to an adversarial spherical image with a saliency-aware fusion method.

3.2. Perspective Projection

For a given position P in spherical coordinates (θ_P, ϕ_P) on the sphere, where θ_P and ϕ_P stand respectively for latitude and longitude. If the field of view $f_h \times f_w$ and desired perspective resolution $h \times w$ are set, the PV image can be generated by a rectilinear projection that maps a position (u, v) on the PV image to a 3D position at (X, Y, Z). The mapping relation between 2D and 3D coordinates is formulated by

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \frac{R}{\sqrt{x^2+y^2+z^2}} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{1}$$

$$\begin{aligned} x &= 2\tan(f_h/2) \cdot (u+0.5)/2 - \tan(f_h/2), \\ y &= \tan(f_w/2) - 2\tan(f_w/2) \cdot (v+0.5)/h, \\ z &= 1.0, \end{aligned} \tag{2}$$

where R is a rotation matrix.

3.3. Distortion-Aware Iterative Fast Gradient Sign Attack

Given a spherical image x_s labeled as y_o and the spherical model C_s, the problem of generating adversarial spherical image x_s^{adv} can be formulated as:

$$\begin{aligned} &\min ||x_s^{adv} - x_s||_\infty \\ &s.t.\ C_s(x_s^{adv}) \neq y_o. \end{aligned} \tag{3}$$

As 360-attack is implemented from the PV domain, the adversarial PV image x_p^{adv} is also required to sastify $C_p(x_p^{adv}) \neq y_o$, where $C_p(\cdot)$ is the planar classifier. It can be proved that the magnitude of the perturbations added on the PV image will be decreased after the projection. For a position ρ^o on the spherical image, the perturbation added on it (denoted as ξ^o) is calculated by the perturbations of several positions on the PV image, that is

$$|\xi^o| = |\sum_i \omega_i \xi_i^p| \leq \sum_i |\omega_i \xi_i^p|, \tag{4}$$

where ξ_i^p is the perturbation of the i-th pixel that related to ρ^o on the PV image, ω_i is its weight during the interpolation, and $\sum_i \omega_i = 1$. As $|\omega_i \xi_i^p| \leq |\xi_i^p|$, then $|\xi^o| \leq |\xi_i^p|$, which means that the magnitude of the perturbation on the spherical image is limited by the allowable size of PV perturbations. Therefore, the attack performance will be degraded when the adversarial PV image is projected to the spherical image. Towards mitigating this issue, we propose a distortion-aware attack operating on the PV domain to minimize the magnitude loss of the perturbation during the inverse perspective projection. We rewrite Eq.(3) as

$$\begin{aligned} &\max L(x_p^{adv}, y_o) \\ &\min L_p(e) \\ &s.t.\ ||e||_\infty \leq \epsilon, \end{aligned} \tag{5}$$

where e is the PV perturbation, $L(\cdot)$ is the loss function, $L_p(\cdot)$ is the pixel-level perturbation loss caused by the inverse perspective projection $P_I(\cdot)$, and ϵ limits the magnitude of perturbations. We seek for a transformation F_t to compensate the distortion introduced by $P_I(\cdot)$, then the second objective can be solved approximately by

$$\min ||P_I(F_t(e)) - e||_2. \tag{6}$$

In order to find an effective F_t, we deeply analyze the perspective projection, and model the position distortion and pixel intensity distortion between PV and spherical images according to spherical triangle formulas, then derive a pixel-wise transformation with geometry knowledge.

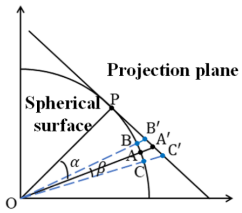


Figure 2. Distortion for the rectilinear projection.

Theorem 1. *Let D_I be the pixel intensity distortion introduced from the perspective projection, then*

$$D_I \propto \sqrt{\frac{\sin\beta}{\beta[\cos^2\frac{\beta}{2} - \sin^2\arccos(\cos\phi_A\cos\theta_A)]}}, \quad (7)$$

where α is the spherical angle corresponding to the arc between the perspective center and the projection point, β is the angle resolution of every sampling grid on the sphere, (θ_A, ϕ_A) is the spherical coordinate of the projection point.

Proof. As shown in Fig. 2, an arbitrary point A on the sphere centered at O can be projected to the tangent plane of the perspective center position P, and its projected point is denoted as A', and the angle between the rays OP and OA is denoted as α. As the grids on the plane are obtained by sampling, each of them represents a small portion of the sphere. For a small neighbourhood centered at A, its left and right endpoint are denoted by B and C respectively, and $\angle BOC$ is the angle resolution of the sampling grid, denoted as β.

An arc on the sphere is stretched to a line during the projection, and the position deviation distortion D_p can be formulated by the ratio between the length of the projected line $B'C'$ and its corresponding arc $\widehat{BAC}$

$$D_p = B'C'/\widehat{BAC} \quad (8)$$

If the spherical image is normalized to a unit sphere, then

$$D_p = [\tan(\alpha + \frac{\beta}{2}) - \tan(\alpha - \frac{\beta}{2})]/\beta. \quad (9)$$

According to the Spherical law of cosines

$$\alpha = \arccos(\sin\phi_P\sin\phi_A + \cos\phi_P\cos\phi_A\cos\Delta\theta), \quad (10)$$

where ϕ_P and ϕ_A are the longitudes of points P and A, and $\Delta\theta$ is the difference between the latitudes of these two points. Considering Eq.(9) and Eq.(10), we can obtain

$$D_p = \frac{\sin\beta}{\beta(\cos^2\frac{\beta}{2} - \sin^2\arccos(\cos\phi_A\cos\Delta\theta))}. \quad (11)$$

Based on the image energy formulation [31] and the Parseval's Theorem, the energy of a disk on the spherical image can be formulated by

$$E(I_S) = \Sigma_\phi\Sigma_\theta I^2_{\theta\phi}, \quad (12)$$

where I_S is the disk on the sphere, $I_{\theta\phi}$ is the pixel on the sphere, and $E(\cdot)$ is the energy function. Ideally, the projection process will follow the conservation of energy. In practice, the perspective projection introduces position distortion by stretching the ideal non-distorted plane, which is implemented by interpolation. Then the energy of the PV image I'_P can be represented by

$$E(I'_P) = \Phi[D_p \cdot E(I_S)], \quad (13)$$

where $\Phi(\cdot)$ denotes the projection function. If we denote the pixel intensity distortion D_I as the ratio between $E(I'_P)$ and $E(I_S)$, then in terms of Eq.(12) and Eq.(13), the relation between D_I and D_p can be approximately characterized as

$$D_I \propto \sqrt{D_p}. \quad (14)$$

Finally, with Eq.(11),

$$D_I \propto \sqrt{\frac{\sin\beta}{\beta[\cos^2\frac{\beta}{2} - \sin^2\arccos(\cos\phi_A\cos\theta_A)]}}. \quad (15)$$

□

Theorem 1 indicates that the pixels in the PV image suffer from D_I distortion when projected onto the sphere. Therefore D_I can be seen as a mask operating on the image, and we can obtain

$$P_I(F_t(e)) = F_t(e)/D_I. \quad (16)$$

To solve Eq.(6), F_t is expected to satisfy $P_I(F_t(e)) \approx e$, then

$$F_t(e) = e \cdot D_I. \quad (17)$$

We propose to solve Eq.(5) by two steps: First, the perturbations are calculated with normal attack approaches. Then they are adjusted with the distortion mask D_I. Therefore, our derivated distortion compensation function can serve as an additional module to any existing attacks, such as FGSM, PGD, and MI-FGSM. In this paper, we choose PGD, and propose the Distortion-Aware Iterative Fast Gradient Sign Method (DAI-FGSM), in which the perturbation is manipulated with a distortion mask at every step. DAI-FGSM is summarized in Algorithm 1, where $\text{sign}(\cdot)$ is the sign function, and ∇ is the differential operator.

3.4. Inverse Perspective Projection and Saliency-aware Fusion

Given a set of adversarial PV images, in order to obtain the final adversarial spherical image, we first re-project the PV images to the sphere by using inversely Eq.(1) and Eq.(2). As each PV image is only projected to a portion of the sphere, we call it a spherical part. Due to the overlapped field of views among different PV images, different pixels in different PV images maybe be projected to the same position on the spherical surface, leading to overlaps across spherical parts. Therefore, how to merge multiple projection pixels in the same position on the sphere to one pixel is a crucial problem. A common method is to average them [3, 14], which is inefficient for generating adversarial

Algorithm 1 DAI-FGSM

Input: A PV image x_p with ground-truth label y_o, the angle resolution β, and a classifier C with loss function L

Input: Perturbation size ϵ, step size per iteration γ, and maximum iterations T

Output: An adversarial PV image x_p^{adv}

for all Positions (θ_A, ϕ_A) in x_p **do**
  Calculate the pixel intensity distortion mask $D_I(\theta_A, \phi_A)$
end for
$x_0^{adv} = x_p$
for $t = 0$ to $T-1$ **do**
  $e_{t+1} = \gamma \cdot \text{sign}(\nabla_{x_t^{adv}} L(x_t^{adv}, y_o))$
  Update $e_{t+1} = \text{Clip}^{\epsilon}\{D_I \cdot e_{t+1}\}$
  Update $x_{t+1}^{adv} = \text{Clip}_x^{(0,1)}\{x_p + e_{t+1}\}$
end for
return $x_p^{adv} = x_T^{adv}$

examples because this operation takes the multiple pixels equally important. With that in mind, we merge projection pixels by considering the difference between the original spherical pixel and its neighbors. For pixels similar with their neighbors, we consider farther spherical parts to collect more information of them. While for pixels significantly different from neighbors, we consider more on their close spherical parts. Considering saliency map implicitly reveals the variation among pixels, we propose a saliency-aware method to fuse spherical parts.

3.4.1 Saliency Detection Based on Spherical Spectral Residual

In this paper, we propose an efficient spherical spectral residual method for spherical saliency detection. Given a spherical image I, the pixel on the position (θ, ϕ) can be represented by the spherical harmonic function [27] as

$$I(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} f_l^m Y_l^m(\theta, \phi), \tag{18}$$

where f_l^m is the spherical harmonic coefficient, Y_l^m is the corresponding spherical harmonic function, l is the spherical harmonic degree, and m is the spherical harmonic order.

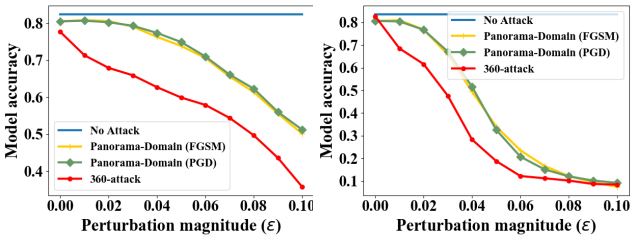
Generally, the spectral maps of spherical images are triangle matrices. To apply the residual approach on the spectrum maps, we first complement the matrices using the mean values of each column. After that, we compute the amplitude and phase of the spectrum, respectively, denoted as I_{am} and I_{ph}. Next, we calculate the log spectrum residual $\mathcal{R}(I)$ of the image:

$$\mathcal{R}(I) = \log(I_{am}) - HF_n * \log(I_{am}), \tag{19}$$

where HF_n is an $n \times n$ mean filter that is used to obtain the averaged log spectrum of the spherical image, and $*$ is the filtering operation.

As discussed in [18], considerable shape similarities can be observed from different spectrums of the input spherical image, and the statistical similarities imply redundancies in the image. Therefore, the information jumping out of the smooth curves deserves the attention, and the residual spectrum contains specific characteristics of the image. Finally, the saliency map S of the original spherical image can be achieved from the residual spectrum $\mathcal{R}(I)$ by

$$S(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} e^{\mathcal{R}_l^m + jI_{ph}} Y_l^m(\theta, \phi). \tag{20}$$



(a) (b)

Figure 3. The performance of the attacks on (a) M_s and (b) M_e.

3.4.2 Saliency-aware Fusion for Inversely Projected PV Images

Every pixel on the sphere is corresponding to a saliency score, implicitly indicating the degree of difference between it and its neighbors. The saliency score is used to guide the fusion of multiple inversely projected PV images.

Assuming a position (θ, ϕ) is covered by k spherical parts, in order to obtain its fused pixel value, we firstly compute the **haversine distances** between the centers of all spherical parts and it, and the distance (denoted as d_i) between the i-th center (θ_i, ϕ_i) and (θ, ϕ) is calculated by:

$$d_i = 2 \arcsin \sqrt{\sin^2(\frac{\theta_i - \theta}{2}) + \cos\theta_i \cos\theta \sin^2(\frac{\phi_i - \phi}{2})}. \tag{21}$$

Next the Gaussian function is used to calculate the weights of the spherical parts according to the saliency score $S(\theta, \phi)$:

$$g_i = e^{-\frac{(d_i - d_{min})^2}{2 \cdot \frac{1}{S(\theta,\phi)}}} = e^{-\frac{(d_i - d_{min})^2 \cdot S(\theta,\phi)}{2}}, \tag{22}$$

where $d_{min} = \min\{d_1, d_2, ..., d_k\}$. Finally, the value of the pixel at (θ, ϕ) is obtained by weighting the k spherical parts I_i with normalized gaussian weights w_i:

$$I_F(\theta, \phi) = \sum_{i=1}^{k} w_i I_i(\theta, \phi), w_i = \frac{g_i}{\sum_{i=1}^{k} g_i}, \tag{23}$$

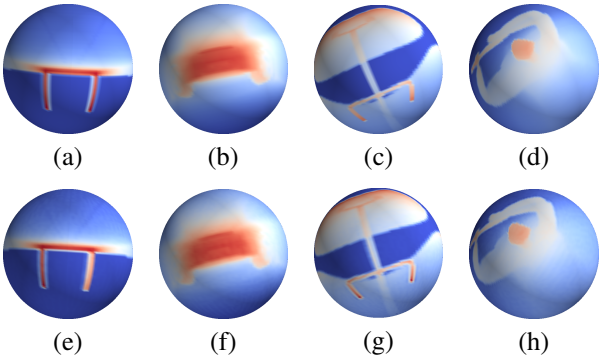


Figure 4. Some examples of 360-attack on the 3D object classification task. The top line shows the original benign images, and the bottom line shows the adversarial images.

Perturbation magnitude ϵ		0.02	0.04	0.06	0.08
FGSM	M_e	0.649	0.290	0.132	0.108
	M_s	0.736	0.682	0.610	0.503
PGD	M_e	0.616	0.283	0.122	0.102
	M_s	0.679	0.627	0.579	0.497
MI-FGSM	M_e	0.608	0.276	0.115	0.107
	M_s	0.662	0.615	0.565	0.491

Table 1. Classification accuracy with different 2D attacks.

Perturbation magnitude ϵ		0.02	0.04	0.06	0.08
Panorama-domain	M_e	0.848	0.845	0.649	0.351
	M_s	0.847	0.842	0.838	0.774
PV-domain	M_e	0.793	0.758	0.375	0.163
	M_s	0.780	0.767	0.716	0.591

Table 2. Classification accuracy for fine-tuned models.

It can be seen from Eq.(23) that our fusion strategy gives the spherical parts closer to the fused position larger weight, especially for pixels on the salient areas. The reason is that the image content on those areas changes dramatically, and the pixels far from them contribute less to the fusion procedure. Therefore we pay more attention to close parts, which provide more accurate information for the fused pixel. On the contrary, for pixels on the non-salient areas, the change of the image content is relatively smooth, and the pixels within a large area may be similar. Weighting multiple spherical parts in non-salient areas helps consider more neighbor pixels. This saliency-aware fusion strategy avoids over-smoothing, reserving more adversarial perturbations.

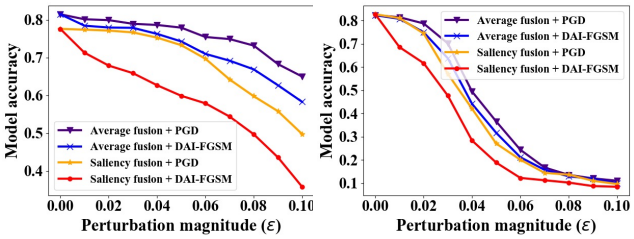
4. Experiments

4.1. 3D Object Classification

We first evaluate the performance of our attack on the shape classification task with the spherical ModelNet-40 dataset, which is a benchmark dataset of spherical models.

4.1.1 Evaluation Setup

We follow the operation in [8], generating a synthetic spherical dataset by projecting the ModelNet-40 [42] dataset to a spherical surface using a ray-mesh intersection method. Because there is no previous work investigating the issue of generating adversarial spherical images, the approach directly attacking the panorama is adopted as the baseline, in which the spherical images are first projected to panoramas, then the typical attack such as FGSM and PGD is carried on them to generate adversarial panoramas, and the adversarial panoramas are finally remapped to the spherical space. The resolution of the spherical images is set to 128×128, and thus the resolution of the panorama images is 64×128. The field of view for rendering PV images is set to $120°$, because it is approximate to that of human vision, and we enforce that the adjacent PV images overlap half with each other, resulting twelve PV images for one spherical image. We consider two target models, including Spherical CNN (M_s) and a standard cnn (M_e) taking panoramas as inputs. The planar model (M_t) used for generating adversarial panoramas and PV images is trained on a systhetic dataset consisting PV images and panoramas rescaled to 128×128. In our experiments, ϵ ranges from 0.01 to 0.10, and γ is set to $\epsilon/10$. The choice of T refers to [23], in which $T = \epsilon/\gamma + 4$.



(a) (b)

Figure 5. The ablation study performed on: (a) the spherical model M_s and (b) the planar model M_e.

4.1.2 Attack Performance

We evaluate the performance of the 360-attack against M_s and M_e, and the results are shown in Fig. 3. All the attacks successfully mislead the two models, although the models perform well with over 80% accuracy on benign images. It can be observed that M_s is more robust than M_e. When facing powerful attacks, M_s still keeps an accuracy above 30%, while the accuracy of M_e is lower than 10%. Among the attacks, the attack capability of the 360-attack is obviously superior to that of panorama-domain attacks. This is because panoramas are more distorted than PV images, making the adversarial panoramas consist of distorted information. The superiority of the 360-attack is more evident when attacking M_s, with more than 20% accuracy decline compared to baselines. However, the results in Fig. 3 (b)

illustrate that the effects of the three attacks on M_e tend to be the same when ϵ increases. It is caused by the fragility of M_e: When ϵ is small, the model has a weak immunocompetence to the adversarial examples, and thus more powerful attacks have more significant effect on the model. When the attack power increases, all attacks severely mislead the model, and finally achieve the similar attacking effects.

The key of the success for the 360-attack lies in its capability for generating aggressive perturbations in the planar attack and reserving them in the fusion operation. In the planar attack step, the proposed DAI-FGSM method can compensate the perturbations to alleviate the impact of the following inverse perspective projection, which guarantees the high aggressiveness of the adversarial PV images. Moreover, in the fusion step, the salient pixels are fused by considering more on their close spherical parts, while the non-salient pixels are obtained by weighting more far spherical parts, which collects abundant information of the pixel and suppresses the impact of the projective distortion on attack.

Fig. 4 shows adversarial examples of 360-attack. All of the adversarial images successfully attack M_s and M_e with a minor modification to the original images, which demonstrates the effectiveness of 360-attack.

Note that our attack directly generates perturbations in the $\mathbb{Z}^2$ space, and then transfers the planar perturbations to the the spherical image. As the disturbed spherical image can successfully mislead the spherical model which operates in the SO(3) group, it demonstrates the transferability of adversarial perturbations from $\mathbb{Z}^2$ to SO(3) groups.

4.1.3 Combined with Different 2D Attacks

As claimed before, the distortion compensation function can be combined with any existing planar attacks. Therefore, we compare the attack results when integrating the distortion mask with FGSM, PGD, and MI-FGSM. The results in Tab. 1 indicates that PGD-based and MI-FGSM-based attacks have similar aggressiveness, and they are both superior to the FGSM-based attack. Intuitively, the FGSM-based attack compensates the distortion only once, resulting in less aggressive adversarial examples compared iterative methods. As for the similar performance between the iterative attacks, the reason may be that the critical factors influencing the attack are the distortion in PV images and strategy to fuse multiple adversarial PVs. The iterative attacks have similar compensation degrees for distortion, which leads to similar performance.

4.1.4 Evaluation against Adversarial Training

Adversarial training, which fine-tunes the model with correctly labeled adversarial examples, is one of the most effective defensive methods against adversarial attacks. We evaluate the performance of 360-attack in the adversarial training setting. Specifically, the victim models are fine-tuned with adversarial spherical images generated by panorama-domain attack and 360-attack with $\epsilon = 0.04$, while the model for generating planar adversarial images remains unchanged. The adversarial examples are then fed into the fine-tuned models, and the results are shown in Table 2.

The results indicate that adversarial training significantly improves the robustness of models. For example, for M_s, the accuracy on the adversarial examples of $\epsilon = 0.04$ generated from the panorama-domain attack improves from 0.7627 to 0.842, while for the 360-attack improves from 0.627 to 0.767. We observe that the panorama-domain attack has little effect on the fine-tuned M_s, while the 360-attack still severely misleads the models, further confirming that 360-attack is more aggressive than the panorama-domain attack. The behavior of M_e is a little different from that of M_s. When facing weak attacks, the improvement of performance is evident, and the model works normally. However, when the adversarial examples generated with a larger ϵ are fed into the model, its accuracy steeply degrades by 20-60%, and the defense is even useless against the 360-attack of $\epsilon = 0.08$. This is caused by the intrinsic instability of the planar model, just like the results in Fig. 3. In summary, the results in Table 2 demonstrate that the 360-attack is still highly effective in the adversarial training setting, remarkably outperforming the panorama-domain attack.

4.1.5 Ablation Study

To measure the effectiveness of the proposed distortion-aware attack and saliency-aware fusion, we perform an ablation study by replacing the DAI-FGSM with PGD in the 2D attack step and replacing the saliency-aware fusion with average fusion. The results of the ablation study are shown in Fig. 5. It can be seen that the lines with purple triangles, which show the results of the PGD attack with average fusion, are always above the blue x-mark lines showing the DAI-FGSM attack with average fusion. This indicates that DAI-FGSM method can reserve more adversarial perturbations than PGD, which benefits greatly from the ability of DAI-FGSM to alleviate the negative impact of distortion introduced by the inverse perspective projection. In addition, the blue lines are always above the red dotted lines that show the DAI-FGSM with saliency fusion. This means the saliency-aware fusion strengthens the aggressiveness of the attack, which results from its effect of reserving more accurate adversarial information. We can also observe that the blue lines are above the orange star lines that indicates the PGD attack with saliency fusion, which demonstrates the saliency-aware fusion contributes more to the attack performance than the DAI-FGSM. It may be because the smoothing effect of the average fusion severely degrades the impact of the perturbations. Overall, the DAI-FGSM method with saliency-aware fusion operations always performs better than any other attacks, further verifying their necessities.

Attack		Panorama-domain attack	360-attack (PGD)	360-attack (Average)	360-attack (DAI-FGSM)
UNet	IoU	0.359/0.347/0.255	0.359/0.340/0.234	0.359/0.357/0.298	0.359/**0.331/0.217**
	Acc.	0.558/0.543/0.496	0.558/0.536/0.473	0.558/0.549/0.509	0.558/**0.531/0.447**
UG-SCNN	IoU	0.413/0.398/0.320	0.413/0.385/0.298	0.413/0.406/0.350	0.413/**0.387/0.273**
	Acc.	0.569/0.553/0.490	0.569/0.547/0.477	0.569/0.553/0.511	0.569/**0.543/0.463**

Table 3. Attack performance on semantic segmentation task. ($\epsilon = 0/0.03/0.08$)

Attack	CFL		LayoutNet	
	IoU	Accuracy	IoU	Accuracy
Panorama-attack	0.595/0.535/0.329	0.932/0.917/0.855	0.564/0.450/0.250	0.911/0.906/0.792
360-attack (PGD)	0.595/0.530/0.318	0.932/0.916/0.839	0.564/0.444/0.239	0.911/0.911/0.773
360-attack (Average)	0.595/0.552/0.357	0.932/0.932/0.865	0.564/0.479/0.31	0.911/0.915/0.830
360-attack (DAI-FGSM)	0.595/**0.522/0.282**	0.932/**0.908/0.830**	0.564/**0.420/0.212**	0.911/**0.870/0.750**

Table 4. Attack performance on the 3D layout reconstruction task. ($\epsilon = 0/0.03/0.08$)

4.2. Tasks on Real-world 360° Datasets

Aforementioned experiments show that 360-attack is effective on the synthetic dataset. We may wonder whether it is still effective on the real-world datasets. Thus, we perform experiments on the real-world datasets for tasks including semantic segmentation and layout prediction. In these experiments, we compare our attack (360-attack (DAI-FGSM)) with panorama-domain attack, 360-attack with PGD, and 360-attack with average fusion, and the experimental setting is the same as that of the 3D object classification. Note that the last two attacks are modified from the proposed attack, and the comparison experiments between them and our attack can be considered as ablation studies.

4.2.1 360° Semantic Segmentation

In this task, UNet [32] and UG-SCNN [22] models are choosen as the target models, and the experimental dataset is the Standford 2D/3D dataset. The experimental results are shown in Table 3. The results indicate 360-attack performs the best in reducing IoU and accuracy of the model prediction among the attacks. It is worth noting that the impact of the attacks on semantic segmentation models is less than those on classification models, and the reason may be that the measurements are calculated from the predictions on all of the pixels, and the slight perturbations added by the attacks only change the predictions of part of the pixels.

4.2.2 3D Layout Reconstruction

We also test our attack against the models trained for 3D layout reconstruction task. In this experiment, we consider CFL [13] and LayoutNet [49] models as our target models, and the test dataset is the SUN360 dataset [44]. Table 4 shows the results of this experiment. Similar to the experiments on the semantic segmentation task, the effect of our attack on the prediction of the target models is greater than that of other compared attacks. Compared to the classification models, the models used in this experiment are more robust to adversarial attacks, and it is due to the simple target of this task: Only eight corners and their corresponding contour lines are expected to be predicted.

4.3. Broader Impact and Limitations

This work can potentially contribute to deeper understanding of DNNs, especially for those processing 360° images. Although we reveal the spherical models are also vulnerable to adversarial examples, our work focuses on assisting researchers to perform more thorough evaluations on DNNs, rather than on attacking real-world systems. We firmly believe that our work can help researchers design new robust models and efficient defenses. In the future, we will focus on addressing the limitation of relying on assigned positions to render PV images, and work for selecting adaptively PV images to implement the attack.

5. Conclusion

We investigate the vulnerability of DNNs trained for spherical images against adversarial attack by transferring adversarial perturbations from the PV domain to the spherical domain. Two key procedures are proposed to preserve more embedded perturbations against the conversion of attack domain space. In the planar attack step, a distortion-aware attack is proposed to suppress the impact of distortion introduced by the projection between spherical and PV images. In the fusion step, we proposed a saliency-aware fusion approach to merge multiple inversely projected spherical parts to the final adversarial spherical image. A systematic study on the spherical and panorama-based models with various synthetic and real-world datasets demonstrates the effectiveness of the proposed attack. Finally, our work also demonstrates the transferability of the adversarial examples between the 2D and 3D spaces.

6. Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 61771469 and the Cooperation project between Chongqing Municipal undergraduate universities and institutes affiliated to CAS (HZ2021015).

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the International Conference on Machine Learning (ICML*, 2018. 2

[2] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Machine Learning (ICML)*, 2018. 2

[3] Peter J Burt and Edward H Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)*, 2(4):217–236, 1983. 4

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 1

[5] 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*, 2017. 2

[6] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representation (ICLR)*, 2019. 2

[7] Oliver J Cobb, Christopher GR Wallis, Augustine N Mavor-Parker, Augustin Marignier, Matthew A Price, Mayeul d'Avezac, and Jason D McEwen. Efficient generalized spherical cnns. *International Conference on Learning Representations (ICLR)*, 2021. 1

[8] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 6

[9] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning (ICML)*, 2020. 2

[10] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[11] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[12] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[13] Clara Fernandez-Labrador, José M Fácil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and José J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *arXiv:1903.08094*, 2019. 1, 2, 8

[14] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 4

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representation (ICLR)*, 2015. 1, 2

[17] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *Experimental Robotics*, pages 477–491. Springer, 2014. 1

[18] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on computer vision and pattern recognition (CVPR)*, 2007. 5

[19] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. Deep 360 pilot: Learning a deep agent for piloting through 360 sports videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[20] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[21] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning (ICML)*, 2018. 2

[22] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical cnns on unstructured grids. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 8

[23] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representation (ICLR)*, 2017. 2, 6

[24] Yeonkun Lee, Jaeseok Jeong, Jongseob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[25] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representation (ICLR)*, 2018. 2

[27] Claus Müller. *Spherical harmonics*, volume 17. Springer, 2006. 5

[28] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[29] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[30] 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 ACM Asia Conference on Computer and Communications Security (Asia CCS)*, 2017. 2

[31] William K Pratt. *Introduction to digital image processing*. CRC press, 2013. 4

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 8

[33] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. A deeper look at 3d shape classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[34] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. In *Advances in Neural Information Processing Systems (NIPS)*, pages 529–539, 2017. 1, 2

[35] Yu-Chuan Su and Kristen Grauman. Making 360 video watchable in 2d: Learning videography for click free viewing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[36] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representation (ICLR)*, 2014. 1, 2

[37] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[38] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[39] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su. Sparse adversarial perturbations for videos. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1

[40] Alex Wong, Mukund Mundhra, and Stefano Soatto. Stereopagnosia: Fooling stereo networks with adversarial perturbations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 1

[41] Weibin. Wu, Yuxin. Su, Xixian. Chen, Shenglin. Zhao, Irwin. King, Michael. R. Lyu, and Yu-Wing. Tai. Boosting the transferability of adversarial samples via attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[42] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[43] 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 (CVPR)*, 2019. 1

[44] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8

[45] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[46] Wenyan Yang, Yanlin Qian, Joni-Kristian Kämäräinen, Francesco Cricri, and Lixin Fan. Object detection in equirectangular panorama. In *International Conference on Pattern Recognition (ICPR)*. IEEE, 2018. 1, 2

[47] Yihuan Zhang, Jun Wang, Xiaonian Wang, and John M Dolan. Road-segmentation-based curb detection method for self-driving via a 3d-lidar sensor. *IEEE Transactions on Intelligent Transportation Systems (TITS)*, 19(12):3981–3991, 2018. 1

[48] Denis Zorin and Alan H. Barr. Correction of geometric perceptual distortions in pictures. In *Proceedings of the Annual Conference on Computer Graphics and Interative Techniques (SIGGRAPH)*, 1995. 2

[49] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8