Towards Viewpoint-Invariant Visual Recognition via Adversarial Training

Shouwei Ruan[1], Yinpeng Dong[2,3], Hang Su[2,4,5*], Jianteng Peng[6], Ning Chen[2], Xingxing Wei[1*]
[1] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[2] Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center,
THBI Lab, BNRist Center, Tsinghua University, Beijing 100084, China
[3] RealAI [4] Peng Cheng Laboratory [5] Pazhou Laboratory (Huangpu), Guangzhou, China [6] OPPO
{shouweiruan,xxwei}@buaa.edu.cn, {dongyinpeng,suhangss,ningchen}@tsinghua.edu.cn, pengjianteng@oppo.com

Abstract

Visual recognition models are not invariant to viewpoint changes in the 3D world, as different viewing directions can dramatically affect the predictions given the same object. Compared to 2D transformations, the exploration of 3D viewpoint invariance deserves more attention for its greater practical significance. Motivated by the success of adversarial training in promoting model robustness, we propose Viewpoint-Invariant Adversarial Training (VIAT) to improve viewpoint robustness of common image classifiers. By regarding viewpoint transformation as an attack, VIAT is formulated as a minimax optimization problem, where the inner maximization characterizes diverse adversarial viewpoints by learning a Gaussian mixture distribution based on a new attack GMVFool, while the outer minimization trains a viewpoint-invariant classifier by minimizing the expected loss over the worst-case adversarial viewpoint distributions. To further improve the generalization performance, a distribution sharing strategy is introduced leveraging the transferability of adversarial viewpoints across objects. Experiments validate the effectiveness of VIAT in improving the viewpoint robustness of various image classifiers based on the diversity of adversarial viewpoints generated by GMVFool.

1. Introduction

The ability of learning invariant representations is highly desirable in numerous computer vision tasks [6, 17] and is conducive to model robustness under semantic-preserving image transformations. Previous works [15, 60, 9, 45] have striven to make visual recognition models invariant to image translation, rotation, reflection, and scaling. However, they mainly consider invariances to 2D image transformations, leaving the *viewpoint* transformation [58] in the 3D world less explored. It has been shown that visual recognition models are susceptible to viewpoint changes [2, 5, 13],

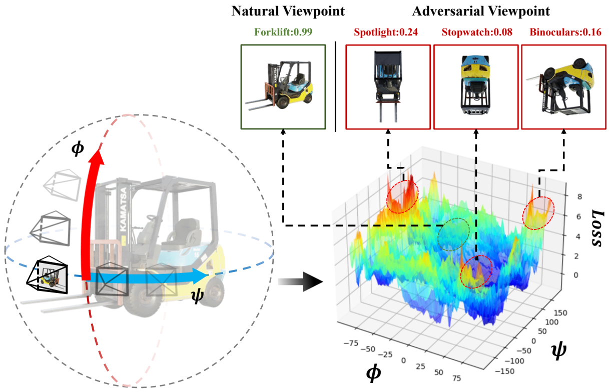

Figure 1. An illustration of viewpoint changes on model performance. We show the loss landscape w.r.t. yaw and pitch of the camera, which demonstrates multiple regions of adversarial viewpoints (We use ResNet-50 as the target model [21]).

exhibiting a significant gap from the human vision that can robustly recognize objects under different viewpoints [7]. Due to the naturalness and prevalence of viewpoint variations in safety-critical applications (*e.g.*, autonomous driving, robotics, surveillance, *etc*.), it is thus imperative to endow visual recognition models with viewpoint invariance.

Despite the importance, it is extremely challenging to build viewpoint-invariant visual recognition models since typical networks take 2D images as inputs without inferring the structure of 3D objects. As an effective data-driven approach, adversarial training augments training data with adversarially generated samples under a specific threat model and shows promise to improve model invariance/robustness to additive adversarial perturbations [63, 59, 54, 57], image translation and rotation [15], geometric transformations [28], *etc*. However, it is non-trivial to directly apply adversarial training to improving viewpoint robustness due to the difficulty of generating the worst-case adversarial viewpoints. A pioneering work [13] proposes *ViewFool*, which encodes real-world 3D objects as Neural Radiance Fields (NeRF) [36] given multi-view images and performs black-

[*] Corresponding author.

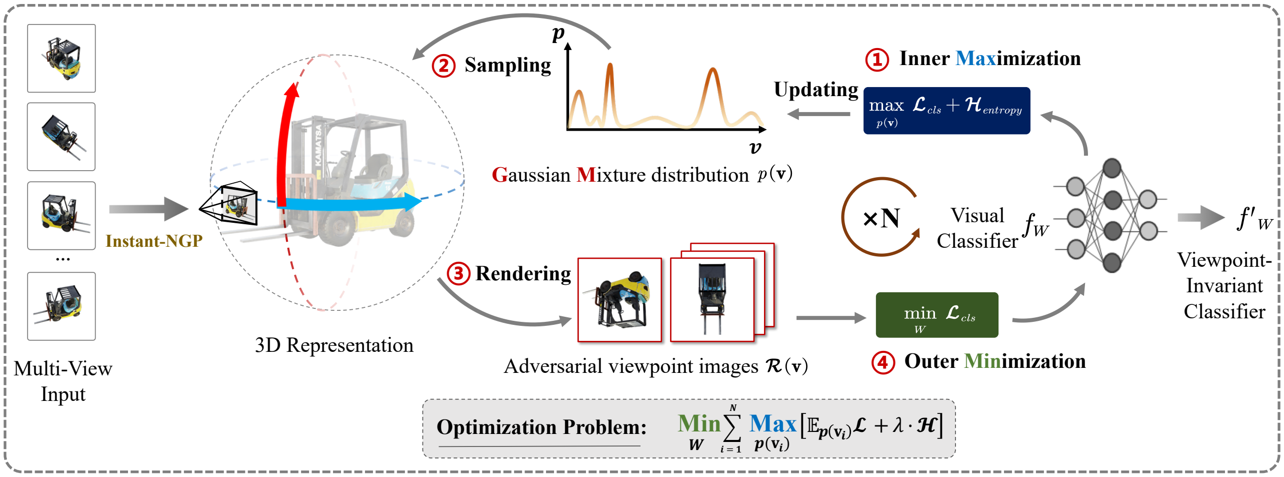
Multi-View Input
...
Instant-NGP
3D Representation
② Sampling
p
v
Gaussian Mixture distribution $p(\mathbf{v})$
③ Rendering
Adversarial viewpoint images $\mathcal{R}(\mathbf{v})$
① Inner Maximization
Updating
$\max_{p(\mathbf{v})} \mathcal{L}_{cls} + \mathcal{H}_{entropy}$
×N
Visual Classifier f_W
f'_W
Viewpoint-Invariant Classifier
$\min_{W} \mathcal{L}_{cls}$
④ Outer Minimization
Optimization Problem: $\underset{W}{\mathrm{Min}} \sum_{i=1}^{N} \underset{p(\mathbf{v}_i)}{\mathrm{Max}} [\mathbb{E}_{p(\mathbf{v}_i)} \mathcal{L} + \lambda \cdot \mathcal{H}]$

Figure 2. An overview of our VIAT framework. We first train the NeRF representation of each object given multi-view images. The inner maximization learns a Gaussian mixture distribution of adversarial viewpoints by maximizing the expectation of classification loss and entropy regularization. The outer minimization samples adversarial viewpoints from the optimized distributions and renders 2D images from adversarial viewpoints, which are fed into the network along with clean samples to train viewpoint-invariant classifiers.

box optimization for generating a distribution of adversarial viewpoints. Though effective, ViewFool only adopts a unimodal Gaussian distribution, which is inadequate to characterize multiple local maxima of the loss landscape w.r.t. viewpoint changes, as shown in Fig. 1. We verify that this can lead to overfitting of adversarial training to the specific attack. Besides, ViewFool is time-consuming to optimize, making adversarial training intractable.

To address these problems, in this paper, we propose **Viewpoint-Invariant Adversarial Training (VIAT)**, the first framework to improve the viewpoint robustness of visual recognition models via adversarial training. As shown in Fig. 2, VIAT is formulated as a distribution-based minimax problem, in which the inner maximization aims to optimize the distribution of diverse adversarial viewpoints while the outer minimization aims to train a viewpoint-invariant classifier by minimizing the expected loss over the worst-case adversarial viewpoint distributions. To address the limitations of ViewFool, we propose **GMVFool** as a practical solution to the inner problem, which generates a Gaussian mixture distribution of adversarial viewpoints for each object, with increased diversity to mitigate overfitting of adversarial training. To accelerate training, we adopt a stochastic optimization strategy to reduce the time cost of training and adopt Instant-NGP [38], a fast variant of NeRF, to improve the efficiency of the optimizing process. In outer maximization, to further improve generalization, we propose a distribution-sharing strategy given the observation that adversarial viewpoint distributions are transferable across objects within the same class. We fine-tune classifiers on a mixture of natural and sampled adversarial viewpoint images to improve their viewpoint invariance.

To verify VIAT's ability of training a viewpoint-invariant model, a multi-view dataset is required. However, previous datasets [16, 29, 41, 10] usually have limited realism and viewpoint range, posing challenges when applying them to this topic. To address this, we devoted significant effort to creating a new multi-view dataset—**IM3D**, which contains 1k typical synthetic 3D objects from 100 classes, tailored specifically for ImageNet categories. IM3D has several notable advantages compared to previous datasets, as shown in Table 1: (1) It covers more categories. (2) It utilizes physics-based rendering (PBR) technology[1] to produce realistic images. (3) It has accurate camera pose annotations and is sampled from a spherical space, leading to better reconstruction quality and exploration of the entire 3D space. Thus, we mainly use it for training and further evaluating our method on other multi-view datasets. We will release our IM3D dataset, which includes multi-view images, 3D source files, and corresponding Instant-NGP weights.

We conduct extensive experiments to validate the effectiveness of both GMVFool and VIAT for generating adversarial viewpoints and improving the viewpoint robustness of image classifiers. Experimental results show that GMVFool characterizes more diverse adversarial viewpoints while maintaining high attack success rates. Based on it, VIAT significantly improves the viewpoint robustness of image classifiers ranging from ResNet [21] to Vision Transformer (ViT) [14] and shows superior performance compared with alternative baselines. Moreover, we construct a new out-of-distribution (OOD) benchmark—**ImageNet-V+**, containing nearly 100k images from the adversarial viewpoints found by GMVFool. It is 10× larger than the previous ImageNet-V benchmark [13]. We hope to serve it as a standard benchmark for evaluating viewpoint robustness in the future.

[1] A 3D modeling and rendering technique that enables physically realistic effects.

Dataset	#Objects	#Classes	PBR	Full 3D	Spherical Pose
ALOI [16]	1K	-	✗	✗	✓
MIRO [29]	120	12	✗	✗	✓
OOWL [25]	500	25	✗	✗	✗
CO3D [41]	18.6K	50	✗	✗	✗
ABO [10]	8K	63	✓	✓	✗
Dong *et al.* [13]	100	85	✓	✓	✓
IM3D (Ours)	1K	100	✓	✓	✓

Table 1. Comparison of our multi-view dataset with others.

2. Related Work

2.1. Robustness to Viewpoint Transformation

Since deep learning models have been applied in numerous safety-critical fields [56], it is necessary to study the robustness of visual recognition models to 3D viewpoint transformations. The ObjectNet [5], OOD-CV [61], and ImageNet-R [22] datasets introduce images including various uncommon camera viewpoints, object poses, and object shapes to evaluate out-of-distribution (OOD) generalization under viewpoint changes. But they are unable to evaluate the performance under the worst-case viewpoint transformation. Alcorn *et al.* [2] generate adversarial perspective samples for 3D objects using a differentiable renderer and find that the model is highly susceptible to viewpoint transformation. Hamdi *et al.* [19] demonstrate the effect of viewpoint perturbation on the model performance of 3D objects and use integral boundary optimization to find robust viewpoint regions for the model. Madan *et al.* [34] introduces diverse category-viewpoint combination images through digital objects and scenes to improve the model's generalization to OOD viewpoints. However, these methods all require 3D models. Dong *et al.* [13] further proposes ViewFool, which uses NeRF to build 3D representations of objects within multi-view images and optimizes the adversarial viewpoint distribution under an entropy regularizer. But it lacks the ability to discover diverse adversarial viewpoints. Our work differs from them mainly in that we focus on improving the viewpoint robustness of models rather than attacking them and then design a more efficient method to generate diverse adversarial viewpoints for this purpose.

2.2. Adversarial Training

The concept of adversarial training (AT) is introduced by Goodfellow *et al.* [18] and is widely recognized as the most effective way to enhance the robustness of deep learning models [3, 4]. Based on the classical AT frameworks such as PGD-AT [35], previous studies have proposed improvement strategies from different aspects [52, 44, 62, 59, 40, 27]. Adversarial training is being widely adopted for various deep learning tasks, such as visual recognition [18, 35, 59, 63], point cloud recognition [32, 64, 53], and text classification [37, 39]. For viewpoint robustness, Alcorn *et al.* [2] demonstrate that adversarial training can have an effect. They generate adversarial viewpoint images by the renderer. However, it only improves the robustness of known objects, while the generalization for unseen could be better. The difference from our work is that we don't rely on traditional renderers and 3D information of objects and can significantly improve the model's adversarial viewpoint generalization ability for unseen objects.

3. Methodology

The proposed Viewpoint-Invariant Adversarial Training (VIAT) is given here. We first introduce the background of NeRF in Sec. 3.1 and the problem formulation in Sec. 3.2, and then present the solutions of VIAT to the inner maximization in Sec. 3.3 and outer minimization in Sec. 3.4. An overview of VIAT is shown in Fig. 2.

3.1. Preliminary on Neural Radiance Fields (NeRF)

Given a set of multi-view images, NeRF [36] has the ability to implicitly represent the object/scene as a continuous volumetric radiance field $F : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \tau)$, where F maps the 3D location $\mathbf{x} \in \mathbb{R}^3$ and the viewing direction $\mathbf{d} \in \mathbb{S}^2$ to an emitted color $\mathbf{c} \in [0,1]^3$ and a volume density $\tau \in \mathbb{R}^+$. Then, using the volume rendering with stratified sampling, we can render a 2D image from a specific viewpoint. Given a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ emitted from the camera center $\mathbf{o}$ through a pixel on the image plane, the expected color $\hat{C}(\mathbf{r})$ of the pixel can be calculated by a discrete set of sampling points $\{t_m\}_{m=1}^M$ as

$$\hat{C}(\mathbf{r}) = \sum_{m=1}^{M} T(t_m) \cdot \alpha(\tau(t_m) \cdot \delta_m) \cdot \mathbf{c}(t_m), \tag{1}$$

where $T(t_m) = \exp(-\sum_{j=1}^{m-1} \tau(t_j) \cdot \delta_j)$, $\tau(t_m)$ and $\mathbf{c}(t_m)$ denote the volume density and color at point $\mathbf{r}(t_m)$, $\delta_m = t_{m+1} - t_m$ is the distance between adjacent points, and $\alpha(x) = 1 - \exp(-x)$. F is approximated by a multi-layer perceptron (MLP) network and optimized by minimizing the L_2 loss between the rendered and ground-truth pixels.

Although NeRF can render photorealistic novel views, both training and rendering are extremely time-consuming. Instant-NGP [38] proposes a fast implementation of NeRF by adaptive and efficient multi-resolution hash encoding. Therefore, in this paper, we adopt Instant-NGP to accelerate the training and volumetric rendering of NeRF.

3.2. Problem Formulation

In visual recognition, viewpoint invariance indicates that a model $f(\cdot)$ can make an identical prediction given two views of the same object as follows:

$$f(I(\mathbf{v}_1)) = f(I(\mathbf{v}_2)), \quad \forall(\mathbf{v}_1, \mathbf{v}_2) \tag{2}$$

where $I(\mathbf{v}_1)$ and $I(\mathbf{v}_2)$ are two images taken from arbitrary viewpoints $\mathbf{v}_1$ and $\mathbf{v}_2$ of the object. However, recent studies [5, 13, 2] have revealed that typical image classifiers are susceptible to viewpoint changes. As viewpoint variations in the 3D space cannot be simply simulated by 2D image transformations, it remains challenging to improve viewpoint invariance/robustness. Motivated by the success of adversarial training in improving model robustness, we propose **Viewpoint-Invariant Adversarial Training (VIAT)** by learning on worst-case adversarial viewpoints.

Formally, viewpoint changes can be described as rotation and translation of the camera in the 3D space [13]. We let $\mathbf{v} = [\mathbf{R}, \mathbf{T}] \in \mathbb{R}^6$ denote the viewpoint parameters bounded in $[\mathbf{v}_{\min}, \mathbf{v}_{\max}]$, where $\mathbf{R} = [\psi, \theta, \phi]$ is the camera rotation along the z-y-x axes using the Tait-Bryan angles, and $\mathbf{T} = [\Delta_x, \Delta_y, \Delta_z]$ is the camera translation along the three axes. Given a dataset $\{\text{obj}_i\}_{i=1}^N$ of N objects and the corresponding ground-truth labels $\{y_i\}_{i=1}^N$ with $y_i \in \{1, ..., C\}$, we suppose that a set of multi-view images is available for each object. With these images, we first train a NeRF model for each object using Instant-NGP to obtain a neural renderer that can synthesize new images from any viewpoint of the object. Rather than finding an adversarial viewpoint $\mathbf{v}_i$ for each object obj_i, VIAT characterizes diverse adversarial viewpoints by learning the underlying distribution $p(\mathbf{v}_i)$, which can be formulated as a distribution-based minimax optimization problem:

$$\min_{\mathbf{W}} \sum_{i=1}^{N} \max_{p(\mathbf{v}_i)} \big[\mathbb{E}_{p(\mathbf{v}_i)}[\mathcal{L}\left(f_{\mathbf{W}}\left(\mathcal{R}(\mathbf{v}_i)\right), y_i\right)] + \lambda \cdot \mathcal{H}(p(\mathbf{v}_i))\big], \tag{3}$$

where $\mathbf{W}$ denotes the parameters of the classifier $f_{\mathbf{W}}$, $\mathcal{R}(\mathbf{v}_i)$ is the rendered image of the i-th object given the viewpoint $\mathbf{v}_i$, $\mathcal{L}$ is a classification loss (*e.g.*, cross-entropy loss), and $\mathcal{H}(p(\mathbf{v}_i)) = -\mathbb{E}_{p(\mathbf{v}_i)}[\log p(\mathbf{v}_i)]$ is the entropy of the distribution $p(\mathbf{v}_i)$ to avoid the degeneration problem and help to capture more diverse adversarial viewpoints [13].

As can be seen in Eq. (3), the inner maximization aims to learn a distribution of adversarial viewpoints under an entropic regularizer, while the outer minimization aims to optimize model parameters by minimizing the expected loss over the worst-case adversarial viewpoint distributions. The motivation of using a distribution instead of a single adversarial viewpoint for adversarial training is two-fold. First, learning a distribution of adversarial viewpoints can effectively mitigate the reality gap between the real objects and their neural representations [13]. Second, the distribution is able to cover a variety of adversarial viewpoints to alleviate potential overfitting of adversarial training, leading to better generalization performance.

To solve the minimax problem, a general algorithm is to first solve the inner problem and then perform gradient descent for the outer problem at the inner solution in a sequential manner based on the Danskin's theorem [11]. Next, we introduce the detailed solutions to the inner maximization and outer minimization problems, respectively.

3.3. Inner Maximization: GMVFool

The key to the success of VIAT in Eq. (3) is the solution to the inner maximization problem. A natural way to solve the problem is to parameterize the distribution of adversarial viewpoints with trainable parameters. The previous method ViewFool [13] adopts a unimodal Gaussian distribution and performs black-box optimization based on natural evolution strategies (NES) [55]. However, due to the insufficient expressiveness of the Gaussian distribution, ViewFool is unable to characterize multiple local maxima of the loss landscape w.r.t. viewpoint changes, as shown in Fig. 1. Thus, performing adversarial training with ViewFool is prone to overfitting to the specific attack and leads to poor generalization performance, as validated in the experiment. To alleviate this problem, we propose **GMVFool**, which learns a Gaussian mixture distribution of adversarial viewpoints to cover multiple local maxima of the loss landscape for more generalizable adversarial training.

For the sake of simplicity, we omit the subscript i in this subsection since the attack algorithm is the same for all objects. Specifically, we parameterize the distribution $p(\mathbf{v})$ by a mixture of K Gaussian components and take the transformation of random variable approach to ensure that the support of $p(\mathbf{v})$ is contained in $[\mathbf{v}_{\min}, \mathbf{v}_{\max}]$ as:

$$\mathbf{v} = \mathbf{a} \cdot \tanh(\mathbf{u}) + \mathbf{b},\ p(\mathbf{u}|\Psi) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \mathbf{I}), \tag{4}$$

where $\mathbf{a} = (\mathbf{v}_{\max} - \mathbf{v}_{\min})/2$, $\mathbf{b} = (\mathbf{v}_{\max} + \mathbf{v}_{\min})/2$, $\Psi = \{\omega_k, \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k\}_{k=1}^K$ are the parameters of the mixture Gaussian distribution with weight $\omega_k \in [0, 1]$ ($\sum_{k=1}^K \omega_k = 1$), mean $\boldsymbol{\mu}_k \in \mathbb{R}^6$ and standard deviation $\boldsymbol{\sigma}_k \in \mathbb{R}^6$ of the k-th Gaussian component. Note that in Eq. (4), $\mathbf{u}$ actually follows a mixture Gaussian distribution while $\mathbf{v}$ is obtained by a transformation of $\mathbf{u}$ for proper normalization.

Now the probability density function $p(\mathbf{u}|\Psi)$ is in the summation form, which is hard to calculate the gradients. Thus, we introduce a latent one-hot vector $\mathbf{\Gamma} = [\gamma_1, ..., \gamma_K]$ determining which Gaussian component the sampled viewpoint belongs to, and obeying a multinomial distribution with probability ω_k, as $p(\mathbf{\Gamma}|\Psi) = \prod_{k=1}^K \omega_k^{\gamma_k}$. With the latent variables, we represent $p(\mathbf{u}|\Psi)$ as a multiplication form with $\mathbf{\Gamma}$ as $p(\mathbf{u}, \mathbf{\Gamma}|\Psi) = \prod_{k=1}^K \omega_k^{\gamma_k} \mathcal{N}(\mathbf{u}|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2 \mathbf{I})^{\gamma_k}$ and $p(\mathbf{u}|\Psi) = \sum_{\Gamma} p(\mathbf{u}, \mathbf{\Gamma}|\Psi)$, which is convenient for taking derivatives w.r.t. distribution parameters Ψ.

Given the parameterized distribution $p(\mathbf{v})$ defined in Eq. (4), the inner maximization problem of Eq. (3) becomes:

$$\begin{aligned} \max_{\Psi}\ & \mathbb{E}_{p(\mathbf{u}, \mathbf{\Gamma}|\Psi)} \big[\mathcal{L}(f_{\mathbf{W}}(\mathcal{R}(\mathbf{a} \cdot \tanh(\mathbf{u}) + \mathbf{b})), y) \\ & - \lambda \cdot \log p(\mathbf{a} \cdot \tanh(\mathbf{u}) + \mathbf{b})\big], \end{aligned} \tag{5}$$

where the second term is the negative log density, whose expectation is the distribution's entropy (proof in Appendix).

To solve this optimization problem, we adopt gradient-based methods to optimize the distribution parameters Ψ. To back-propagate the gradients from random samples to the distribution parameters, we can adopt the low-variance reparameterization trick [8, 31]. Specifically, we reparameterize $\mathbf{u}$ as $\mathbf{u} = \prod_{k=1}^{K}\boldsymbol{\mu}_k^{\gamma_k} + \prod_{k=1}^{K}\boldsymbol{\sigma}_k^{\gamma_k} \cdot \mathbf{r}$, where $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. With this reparameterization, the gradients of the loss function in Eq. (5) w.r.t. Ψ can be calculated. However, similar to ViewFool, although the rendering process of NeRF is differentiable, it requires significant memory overhead to render the full image. Thus, we also resort to NES to obtain the natural gradients of the classification loss with only query access to the model. For the entropic regularizer, we directly compute its true gradient. Therefore, the gradients of the objective function in Eq. (5) w.r.t. $\omega_k, \boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ can be derived as (proof in Appendix):

$$
\begin{aligned}
\nabla_{\omega_k} &= \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\left\{\gamma_k \cdot \left[\mathcal{L}_{\text{cls}} \cdot \frac{1}{\omega_k} - \lambda\right]\right\};\\
\nabla_{\boldsymbol{\mu}_k} &= \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\left\{\gamma_k \cdot \left[\mathcal{L}_{\text{cls}} \cdot \frac{\boldsymbol{\sigma}_k\mathbf{r}}{\omega_k} - \lambda{\cdot}2\tanh(\boldsymbol{\mu}_k + \boldsymbol{\sigma}_k\mathbf{r})\right]\right\};\\
\nabla_{\boldsymbol{\sigma}_k} &= \mathbb{E}_{\mathcal{N}(\mathbf{r}|\mathbf{0},\mathbf{I})}\left\{\gamma_k \cdot \left[\mathcal{L}_{\text{cls}} \cdot \frac{\boldsymbol{\sigma}_{\boldsymbol{k}}(\mathbf{r}^2 - 1)}{2\omega_k}\right.\right.\\
&\qquad\left.\left. +\lambda \cdot \frac{(1 - 2\mathbf{r} \cdot \tanh(\boldsymbol{\mu}_k + \boldsymbol{\sigma}_k\mathbf{r}) \cdot \boldsymbol{\sigma}_k}{\boldsymbol{\sigma}_k}\right]\right\};\\
\mathcal{L}_{\text{cls}} &= \mathcal{L}(f_{\mathbf{W}}(\mathcal{R}(\mathbf{a} \cdot \tanh(\prod_{k=1}^{K}\boldsymbol{\mu}_k^{\gamma_k} + \prod_{k=1}^{K}\boldsymbol{\sigma}_k^{\gamma_k} \cdot \mathbf{r}) + \mathbf{b})), y).
\end{aligned}
\tag{6}
$$

In practice, we use the Monte Carlo method to approximate the expectation in gradient calculation and use iterative gradient ascent to optimize the distribution parameters of each Gaussian component. In addition, we normalize ω_k after each iteration to satisfy $\sum_{k=1}^{K}\omega_k = 1$. Algorithm 1 outlines the overall algorithm of GMVFool.

3.4. Outer Minimization

In outer minimization of Eq. (3), our goal is to minimize the loss expectation over the learned adversarial viewpoint distributions. However, there are two problems of adversarial training: inefficiency and overfitting. We next detail how we address these two problems with the **stochastic update strategy** and **distribution sharing strategy**, respectively.

Although we introduce the efficient Instant-NGP, the inner maximization still needs many gradient steps to converge for rendering images from new viewpoints. This process is typically unacceptable for adversarial training, as each optimization step of outer minimization needs to solve the inner maximization problem for a batch of objects. To accelerate adversarial training, we propose a stochastic update strategy for the inner problem. First, we perform full inner optimization to generate adversarial viewpoint distributions for all objects given a pre-trained image classifier. At each fine-tuning epoch, we only update the distribution parameters for one randomly selected object in each category while keeping those for other objects unchanged. Note that all objects can be sufficiently optimized within multiple epochs. The rationale is that GMVFool is able to learn a sufficiently wide range of adversarial viewpoints, making the distribution effective for adversarial training over an extended period. This strategy can significantly improve efficiency and make adversarial training feasible.

Algorithm 1 GMVFool

```
Input: Image classifier $f_{\mathbf{W}}$, rendering function $\mathcal{R}$, true label $y$, number of iterations $T$, number of Monte Carlo samples $q$, learning rate $\eta$, number of Gaussian components $K$, and balance hyperparameter $\lambda$.
Initialize the Gaussian mixture distribution parameters of the object $\Psi^0 = \{\omega_k^0, \boldsymbol{\mu}_k^0, \boldsymbol{\sigma}_k^0\}_{k=1}^K$;
for $t = 1$ to $T$ do
    Sample $\{\mathbf{r}_j\}_{j=1}^q$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$;
    Sample $\{\boldsymbol{\Gamma}_j\}_{j=1}^q$ from the multinomial distribution with probability $\omega_k^t$;
    Calculate $\{\mathbf{u}_j\}_{j=1}^q$;
    Calculate $\nabla_{\Psi^t} = \{\nabla_{\omega_k^t}, \nabla_{\boldsymbol{\mu}_k^t}, \nabla_{\boldsymbol{\sigma}_k^t}\}$ by Eq. (6);
    Update the parameters:
        $\Psi^{t+1} \leftarrow \Psi^t + \eta \cdot \nabla_{\Psi^t}$;
    Normalize $\omega_k^{t+1} \leftarrow \omega_k^{t+1} / \sum_{k=1}^K \omega_k^{t+1}$;
end for
Output: Parameters of adversarial viewpoint distribution: $\Psi^T = \{\omega_k^T, \boldsymbol{\mu}_k^T, \boldsymbol{\sigma}_k^T\}_{k=1}^K$.
```

Besides, we find that as the training epochs increase, the learned adversarial viewpoint distributions would degenerate, *i.e.*, in the late stage of training, the diversity of adversarial viewpoints decreases, leading to overfitting of adversarial training and inferior results. To alleviate this problem, we propose a distribution sharing strategy, in which we share the distribution parameters of different objects within the same category. It is based on our finding that the adversarial viewpoint distributions of objects within the same class are highly similar, as shown in Fig. 3. For each object in training, we choose its own distribution parameters or randomly select other distribution parameters of another object for sampling based on a probability π.

The training process of VIAT can be summarized as follows: at each fine-tuning epoch, the parameters of the adversarial viewpoint distribution are updated using GMVFool. In particular, all objects' parameters are initially optimized at the first epoch, while in each subsequent epoch, the object's parameters within each class are randomly updated. Next, the adversarial viewpoints are sampled from a distribution based on the sharing probability. Then the corresponding adversarial examples are generated by Instant-NGP. These examples are fed to the network with the clean

	Method	ResNet-50			ViT-B/16		
		ImageNet	ViewFool	GMVFool	ImageNet	ViewFool	GMVFool
Standard-trained	-	85.60	8.28	8.98	92.88	25.70	29.10
Augmentation	Natural	85.76	16.52	19.30	92.78	43.32	46.48
	Random	85.82	34.80	33.52	92.78	62.03	67.34
VIAT	ViewFool	85.66	55.12	58.75	92.70	79.53	82.03
	GMVFool	85.70	**59.84**	**59.61**	92.56	**82.81**	**83.13**

Table 2. The **classification accuracy** (%) from evaluation protocols with ResNet-50 and ViT-B/16, which are trained via ImageNet subset only (standard-trained), data augmentation by natural and random viewpoint images, and VIAT framework with ViewFool and GMVFool.

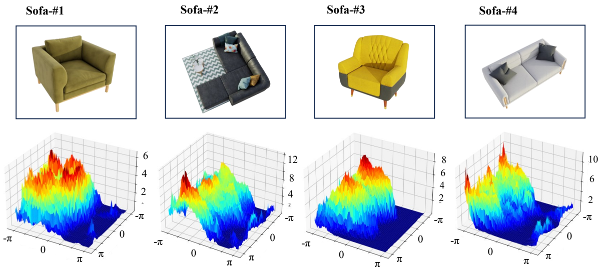

Figure 3. The adversarial viewpoint regions of objects within the same class are similar. We show the **loss landscape** w.r.t. ψ and ϕ of four different sofas based on ResNet-50, in which we keep $[\theta, \Delta_x, \Delta_y, \Delta_z]$ as $[0, 0, 0, 0]$.

samples from ImageNet to calculate the cross-entropy loss. Finally, the network parameters are optimized to obtain a viewpoint-invariant model.

Moreover, VIAT boasts acceptable time costs, which can be attributed to three factors: (1) It optimizes the mixture distribution of adversarial viewpoints instead of individual ones, allowing for the generation of diverse adversarial viewpoints through distribution sampling instead of multiple optimizations. (2) Using the efficient Instant-NGP accelerates the optimization of adversarial viewpoint distribution and rendering of viewpoint samples. (3) The stochastic optimization strategy based on distribution transferability further reduces time consumption in adversarial training.

4. Experiments

4.1. Performance of VIAT

Experimental settings. (A) Datasets: To address the lack of multi-view images in ImageNet, we construct IM3D: a dataset composed of 1K typical synthetic 3D objects from 100 ImageNet categories, each category containing 10 objects. We acquire the multi-view images on the upper hemisphere with the corresponding camera poses for each object, and then we learn the NeRF representation using Instant-NGP. Objects of each type are divided into a training set and a validation set with a ratio of 9:1, which are utilized for adversarial training and validation of viewpoint invariance. **(B) Model:** Two classifiers are considered to perform experiments, including the CNN-based ResNet-50 [21] and the Transformer-based ViT-B/16 [14]. We train the classifiers on the subset of ImageNet, which corresponds to our 3D synthetic object's category, achieving 85.60%, 92.88% Top-1 accuracy on the ImageNet subset. **(C) VIAT Setting:** Following [13], we initialize the camera at $[0, 4, 0]$, the range of rotation parameters are set as $\psi \in [-180°, 180°]$, $\theta \in [-30°, 30°]$, $\phi \in [20°, 160°]$, the range of translation parameters are set as $\Delta_x \in [-0.5, 0.5]$, $\Delta_y \in [-1, 1]$, $\Delta_z \in [-0.5, 0.5]$, and the balance hyperparameter $\lambda = 0.01$. Based on the results of the ablation studies, we set the components number K=15, and distribution sharing probability π=0.5. For the inner maximization step, we approximate the gradients in Eq. (6) with q=100 MC samples and use the Adam [30] optimizer to update Ψ for 50 iterations in the first epoch, then iterate 10 times under the previous Ψ for subsequent epochs. After obtaining the model trained on the ImageNet subset, we continue to train the model for 60 epochs with the adversarial viewpoints and ImageNet clean samples, with a ratio of 1:32.

Evaluation metrics. To fully explore the viewpoint invariance of models, we use Top-1 accuracy as the evaluation metric and set up four evaluation protocols: (a) ImageNet: the accuracy is calculated under the validation set of ImageNet. (b) ViewFool: the accuracy is calculated under renderings of adversarial viewpoints generated by ViewFool. (c) GMVFool: the accuracy is calculated under renderings of adversarial viewpoints generated by our GMVFool.

Experimental results. The experimental results are shown in Table 2. We compare VIAT with three baselines: (a) Data augmentation with the most common viewpoint renderings from training objects' natural states. For this, we define a range of views frequently appearing in ImageNet for each class (*e.g.* hotdogs are usually in the top view). (b) Data augmentation with random viewpoint rendering of objects in the training set. (c) VIAT uses ViewFool as the inner maximization method. To be fair, we also use Instant-NGP for accelerating ViewFool in the adversarial training. From the table, we can draw the following conclusions:

(1) VIAT significantly improves the viewpoint invariance of the model. Under the adversarial viewpoint generated by GMVFool and ViewFool, the accuracy of ResNet-50 is improved by 50.63% and 51.56%, while that of ViT-B/16 is improved by 54.03% and 57.11%, respectively,

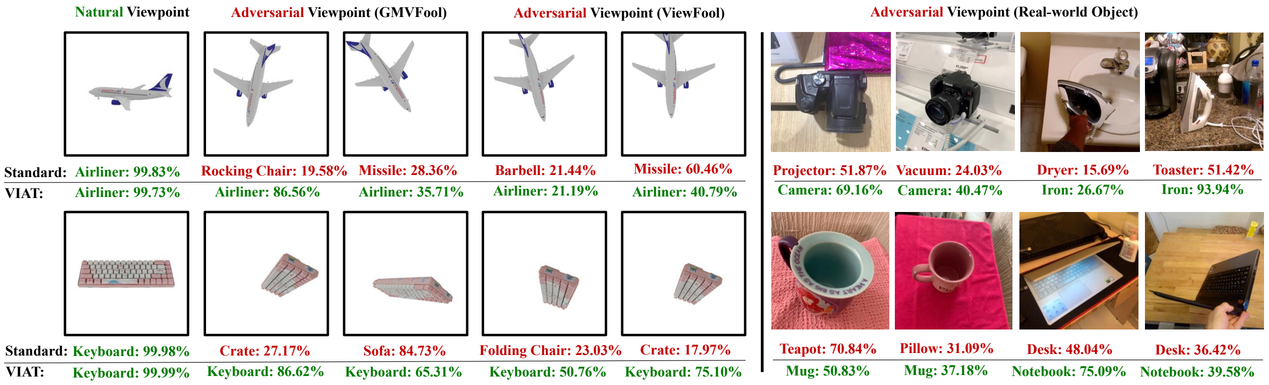

Figure 4. The **prediction** examples of Standard-trained and VIAT-trained ResNet-50 under natural and adversarial viewpoint images. Green and red text represent correct and incorrect predictions, respectively, and the corresponding number is the confidence value.

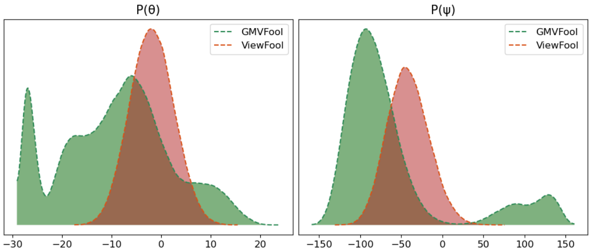

Figure 5. The **probability density curves** of the adversarial distribution under viewpoint parameters θ and ψ, which are optimized by GMVFool ($K = 5$) and ViewFool, respectively.

compared to the standard-trained model.

(2) Adopting GMVFool as VIAT's inner maximization method results in higher accuracy than adopting ViewFool under the adversarial viewpoint images. The smaller accuracy gap between the two attack methods means that VIAT+GMVFool can better generalize to different viewpoint attacks. We think it benefits from the Gaussian mixture modelling of GMVFool which can generate diverse adversarial viewpoints to be learned by the network.

(3) Data augmentation methods using natural and random viewpoint images have great limitations in improving the model's performance under adversarial viewpoints.

(4) ViT-B/16 is better than ResNet-50 in resisting adversarial viewpoint attacks. This phenomenon may benefit from its transformer structure, which is also confirmed by the benchmark results of ImageNet-V+ in Sec. 4.4.

Visualization. Fig. 4 shows the visualization results of the natural and adversarial viewpoints rendering of objects, as well as the output and confidence of the standard-trained and VIAT-trained ResNet-50. The results demonstrate that VIAT-trained model can still predict the correct labels when facing the adversarial viewpoint. Additional examples will be presented in the Appendix.

4.2. Additional Results and Ablation Studies

The effects of K and π. We further conduct ablation experiments to investigate the impact of the number of Gaussian components (K) and distribution sharing probability (π). Fig. 6 presents the classification accuracy of the model against GMVFool attacks after VIAT training with different settings. We observe a positive correlation between the model's ability to resist viewpoint attacks and the number of components used by VIAT. Additionally, a suitable sharing probability benefits the model in achieving better viewpoint invariance. However, a high component number and sharing probability will lead to the opposite situation.

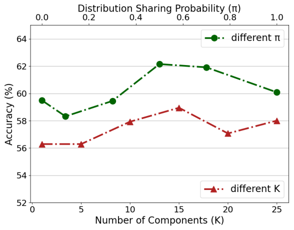

Figure 6. The **accuracy** (%) of VIAT-trained ResNet-50 against adversarial viewpoints, using various sets of K and π.

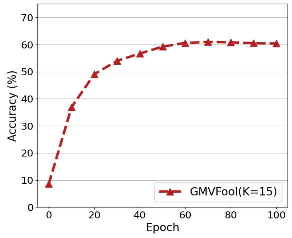

Figure 7. The **accuracy** (%) of VIAT-trained ResNet-50 against GMVFool attack with different training iterations.

Convergence discussion. As a distribution-based adversarial training framework, the convergence of VIAT is guaranteed in theory [12]. Furthermore, We study the convergence of VIAT with learning rate 0.001, $K = 15$, $\pi = 0.5$. The accuracy under adversarial viewpoints generated by GMVFool is presented in Fig. 7, indicating that VIAT can converge well under experimental setting.

The superiority of GMVFool. Contributed by the mixture distribution design, we can control K to balance viewpoint diversity and attack performance in different tasks. Specifically, for adversarial training, GMVFool with larger K is used to improve the adversarial viewpoint diversity, which is crucial for achieving better robustness via adversarial training. As depicted in Fig. 5, GMVFool ($K = 5$) captures more comprehensive and diverse adversarial viewpoints than ViewFool. Table 3 quantitatively confirms this

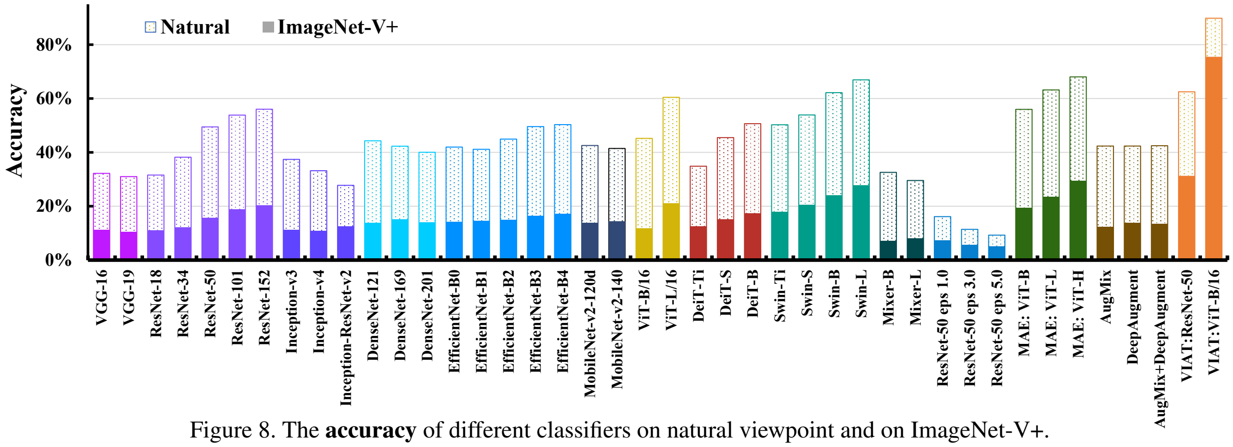

Figure 8. The **accuracy** of different classifiers on natural viewpoint and on ImageNet-V+.

Method	ResNet-50 [21]		EN-B0 [49]		DeiT-B [51]		Swin-B [33]	
	$\mathcal{R}(p^*(\mathbf{v}))\uparrow$	$\mathcal{H}(p^*(\mathbf{v}))\uparrow$	$\mathcal{R}(p^*(\mathbf{v}))\uparrow$	$\mathcal{H}(p^*(\mathbf{v}))\uparrow$	$\mathcal{R}(p^*(\mathbf{v}))\uparrow$	$\mathcal{H}(p^*(\mathbf{v}))\uparrow$	$\mathcal{R}(p^*(\mathbf{v}))\uparrow$	$\mathcal{H}(p^*(\mathbf{v}))\uparrow$
Random Search	64.12	-	79.18	-	32.25	-	19.88	-
ViewFool	91.50	-10.14	95.64	-10.38	80.37	-10.41	73.85	-10.71
GMVFool (K=1)	**92.13**	-10.17	**95.78**	-10.38	**80.61**	-10.44	**74.02**	-10.62
GMVFool (K=3)	89.20	-3.75	**95.78**	-3.94	79.65	-3.92	72.48	-4.05
GMVFool (K=5)	91.56	**-0.69**	95.62	**-0.85**	78.91	**-1.00**	70.16	**-1.07**

Table 3. The **attack success rate** (%) and the **entropy** of methods against various classifiers. $\mathcal{R}(\cdot)$ denotes the rendering process.

observation through the entropy $\mathcal{H}(p^*(\mathbf{v}))$. For viewpoint attacks, GMVFool can be more effective with a smaller K. Table 3 compares the attack performance of GMVFool ($K=1$, 3, and 5, respectively) with previous methods. It shows that GMVFool (K=1) performs best on the *optimal distribution of adversarial viewpoints* $p^*(\mathbf{v})$ and GMVFool (K=3 and 5) maintains good performance under $p^*(\mathbf{v})$.

4.3. Evaluation on Other Datasets

Performance on real-world adversarial viewpoints. We conduct evaluation experiments on objectron [1], which contains object-centric videos in the wild.. The accuracy of standard-trained and VIAT-trained ResNet-50 is presented in Table 4, demonstrating that VIAT enhances the model's performance in unnatural viewpoints. Some cases are shown in Fig. 4. Furthermore, we conduct experiments on other OOD datasets that contain viewpoint perturbations, the results in Table 5 demonstrate the VIAT-trained model's ability to resist different natural perturbations.

	shoe	*camera*	*mug*	*chair*	*computer*
Standard	76.28	81.42	97.67	38.82	60.19
VIAT (Ours)	**76.73**	**86.93**	**99.07**	**39.75**	**63.40**

Table 4. The **accuracy** (%) of standard-trained and VIAT-trained ResNet-50 in the real-world images from the objectron dataset.

Performance on multi-view datasets. We conduct evaluation experiments on other multi-view datasets, for which we obtain the categories that overlap with our training categories. Table 5 presents the accuracy of the model under various viewpoints, indicating the improved robustness of the VIAT-trained model against viewpoint transformations.

	Dataset	Standard	Aug	VIAT (Ours)
OOD Datasets	ObjectNet [5]	35.92	36.02	**37.39**
	ImageNet-A [24]	19.03	18.65	**20.32**
	ImageNet-R [22]	45.06	45.09	**46.51**
	ImageNet-V [13]	28.15	32.83	**38.96**
Multi-view Datasets	MIRO [29]	57.41	58.78	**65.86**
	OOWL [25]	51.41	51.24	**52.13**
	CO3D [41]	64.68	64.90	**66.04**

Table 5. The **accuracy** (%) of standard-trained, Augmentation with random renderings (Aug) and VIAT-trained ResNet-50 in various OOD and multi-view datasets.

4.4. ImageNet-V+ Benchmark

We utilize GMVFool to construct a larger benchmark dataset, ImageNet-V+, to evaluate the viewpoint robustness of visual recognition models. It comprises 100K adversarial viewpoint images of 1K synthetic objects belonging to the 100 ImageNet classes. The details and visualizations will be included in the Appendix. We adopt ImageNet-V+ to evaluate 40 different models pre-trained on ImageNet, including models with different structures (the CNN-based VGG [46], ResNet [21], Inception [48, 47], DenseNet [26], EfficientNet [49], MobileNet-v2 [43], the transformer-based: ViT [14], DeiT [51], Swin Transformer [33], and the MLP Mixer [50]), different training paradigms (adversarial training [42] and mask-autoencoder [20]), different augmentation methods (AugMix [23], DeepAugment [22]). For comparison, we also evaluate the model trained with VIAT.

Fig. 8 illustrates the accuracy of various models on natural viewpoint images and ImageNet-V+. When exposed to adversarial viewpoints, the accuracy of all models decreases significantly. We observe that the model's performance with the same architectures is positively related to its size, with transformer-based models outperforming CNN-based models. Among them, MAE with ViT-H performs best in ImageNet-V+, achieving 29.37% accuracy. Models using data augmentation and adversarial training, which is robust to adversarial examples and image corruption in previous work, perform poorly from the adversarial viewpoint. Finally, ViT-B/16 trained with VIAT outperform all models using standard training, achieving an accuracy of 75.49%.

5. Conclusion

This paper proposed the VIAT framework to obtain viewpoint invariance for visual recognition via adversarial training and contributed GMVFool, an efficient method for generating diverse adversarial viewpoints. We also provided a new multi-view dataset—IM3D and conducted extensive experiments to verify the effectiveness of VIAT in enhancing viewpoint invariance. Moreover, we introduced ImageNet-V+, a large viewpoint OOD benchmark including 100K adversarial viewpoint images of 1K synthetic objects, and provided the accuracy on various models.

6. Acknowledgement

This work was supported by the Project of the National Natural Science Foundation of China (No.62076018), and the Fundamental Research Funds for the Central Universities.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 8

[2] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019. 1, 3

[3] Anish Athalye and Nicholas Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018. 3

[4] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021. 3

[5] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9453–9463, 2019. 1, 3, 8

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 1

[7] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 1

[8] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 5

[9] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 1

[10] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022. 2, 3

[11] John M Danskin. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 2012. 4

[12] Yinpeng Dong, Zhijie Deng, Tianyu Pang, Jun Zhu, and Hang Su. Adversarial distributional training for robust deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8270–8283, 2020. 7

[13] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. In *Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 4, 6, 8

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 6, 8

[15] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR, 2019. 1

[16] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61:103–112, 2005. 2, 3

[17] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, volume 22, 2009. 1

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[19] Abdullah Hamdi and Bernard Ghanem. Towards analyzing semantic robustness of deep neural networks. In *European Conference on Computer Vision*, pages 22–38. Springer, 2020. 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 8

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 2, 6, 8

[22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 3, 8, 9

[23] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 9

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 8

[25] Chih-Hui Ho, Brandon Leung, Erik Sandstrom, Yen Chang, and Nuno Vasconcelos. Catastrophic child's play: Easy to perform, hard to defend adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9229–9237, 2019. 3, 8

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 8

[27] Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao Sr. Improving fast adversarial training with prior-guided knowledge. *arXiv preprint arXiv:2304.00202*, 2023. 3

[28] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4441–4449, 2018. 1

[29] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5010–5019, 2018. 2, 3, 8

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 5

[32] 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. 3

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 8

[34] Spandan Madan, Timothy Henry, Jamell Dozier, Helen Ho, Nishchal Bhandari, Tomotake Sasaki, Frédo Durand, Hanspeter Pfister, and Xavier Boix. When and how do cnns generalize to out-of-distribution category-viewpoint combinations? *arXiv preprint arXiv:2007.08032*, 2020. 3

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 1, 3

[37] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016. 3

[38] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 2, 3

[39] Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138, 2022. 3

[40] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020. 3

[41] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2, 3, 8

[42] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020. 8

[43] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[44] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis,

Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019. 3

[45] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1233–1240, 2013. 1

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8

[47] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 8

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 8

[49] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 8

[50] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 34:24261–24272, 2021. 8

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 8

[52] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 3

[53] Ruibin Wang, Yibo Yang, and Dacheng Tao. Art-point: Improving rotation robustness of point cloud classifiers via adversarial rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14371–14380, 2022. 3

[54] Xingxing Wei, Bangzheng Pu, Jiefan Lu, and Baoyuan Wu. Physically adversarial attacks and defenses in computer vision: A survey. *arXiv preprint arXiv:2211.01671*, 2022. 1

[55] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(1):949–980, 2014. 4

[56] Yang Wu, Ding-Heng Wang, Xiao-Tong Lu, Fan Yang, Man Yao, Wei-Sheng Dong, Jian-Bo Shi, and Guo-Qi Li. Efficient visual recognition: A survey on recent advances and brain-inspired methodologies. *Machine Intelligence Research*, 19(5):366–411, 2022. 3

[57] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17:151–178, 2020. 1

[58] Richard Zemel and Geoffrey E Hinton. Discovering viewpoint-invariant relationships that characterize objects. In *Advances in neural information processing systems*, pages 299–305, 1990. 1

[59] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482, 2019. 1, 3

[60] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 1

[61] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to individual nuisances in real-world out-of-distribution shifts. In *ICML 2022 Shift Happens Workshop*, 2022. 3

[62] Shiji Zhao, Xizhe Wang, and Xingxing Wei. Mitigating the accuracy-robustness trade-off via multi-teacher adversarial distillation. *arXiv preprint arXiv:2306.16170*, 2023. 3

[63] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, pages 585–602. Springer, 2022. 1, 3

[64] Yue Zhao, Yuwei Wu, Caihua Chen, and Andrew Lim. On isometry robustness of deep 3d point cloud models under adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1201–1210, 2020. 3