# Securing 3D Deep Learning Models: Simple and Effective Defense Against Adversarial Attacks

Rosina F. Kharal    Saif Al-Din Ali    Usama Mohiuddin    Samir El Sayed

***Abstract*—In recent years, the growing vulnerability of deep neural networks (DNNs) to adversarial attacks has posed significant challenges in the field of machine learning, particularly in mission-critical applications such as computer vision. As a result, adversarial machine learning has emerged as a crucial research area focused on fortifying deep neural networks against these sophisticated threats. Simultaneously, the use of 3D datasets representing 3D objects has become increasingly important in applications like autonomous driving, robotics, and augmented reality. As the adoption of deep learning networks for processing and classifying both 3D and 2D data grows, concerns over their vulnerability to adversarial attacks have escalated.**

**Despite extensive research on adversarial defense in many areas, adversarial 3D deep learning models remains comparatively unexplored. This study examines the susceptibility of 3D deep learning to various forms of adversarial attacks, revealing significant weaknesses in current approaches. Our findings demonstrate these attacks, which involve subtle modifications to the data, can significantly degrade the performance of classifiers, leading to misclassification and potentially dangerous outcomes in real-world scenarios. We illustrate that adversarial inputs can compromise prediction performance, with accuracy dropping by over 20%.**

**In response, we propose an efficient defense strategy that does not overburden the learning model with a heavy adversarial training step. Our simplified defense strategy employs only the *best* classified 3D objects per class and not only restores network classification accuracy to baseline performance but can improve accuracy to above baseline performance in the event of adversarial attacks.**

***Index Terms*—Deep learning, 3D object classification, 3D model data, adversarial attacks, data security, robustness in 3D deep learning, neural networks, 3D CNNs, machine learning security**

## I. Introduction

This decade has witnessed phenomenal advancements in machine learning, particularly in image and video classification through deep learning models. These advancements have been pivotal not only in autonomous vehicles but also in various other areas such as medical imaging, facial recognition, security surveillance, and content moderation. In 2013, seminal work by Szegedy et al. [1] and Goodfellow et al. [2] illustrated the stark reality of the robustness of deep neural networks used for vision classification tasks; models were surprisingly susceptible to adversarial attacks leading to incorrect classification with high a degree of confidence [1]–[4]. An adversarial attack on images involves intentionally perturbing an image in a way that is often imperceptible to the human eye but causes a deep-learning model to misclassify the image. These attacks can be categorized into different families, including White-Box Attacks, where the attacker has full knowledge of the model's architecture and parameters; Black-Box Attacks, where the attacker has no direct access to the model's internals and instead relies on input-output queries; and Poisoning Attacks, where the attacker manipulates the training data to introduce vulnerabilities that can be exploited later. Adversarial attacks on deep learning networks and classifiers aim to deceive the network into making incorrect predictions, thus undermining the reliability of machine learning systems.

Adversarial attacks are a significant threat to the reliability of mission-critical classification tasks in 3-dimensional space, such as autonomous driving, where precise detection of objects is crucial for safety, robotics, where accurate environmental navigation is vital, and augmented reality, where correct object recognition enhances user experience. For example, perturbations to 3D point clouds in LiDAR systems can cause autonomous vehicles to misclassify objects, leading to potential accidents [5]. In robotics, similar attacks can interfere with environmental mapping, causing navigation systems to make incorrect decisions that affect movement and task execution [6]. In augmented reality applications, adversarial attacks can distort the recognition of virtual objects, undermining user interactions and experience [6]. Additionally, in medical imaging, where 3D scans like MRIs, CTs, or X-rays are used to diagnose conditions, adversarial perturbations can lead to incorrect diagnoses, such as confusing healthy tissue with tumors or failing to detect critical conditions, thereby compromising patient safety [7] [8].
However, in 3-dimensional space, the study of defenses against adversarial attacks has been primarily focused on 3D point cloud data [9]. 3D model data represents objects or scenes in three-dimensional space and is widely used in applications such as autonomous driving, robotics, and augmented reality. The study of 3D model datasets representing objects using 3D meshes has been limited [10]. 3D model data is also susceptible to adversarial attacks, where the goal is to subtly modify the data to mislead a classifier. These modifications can reduce the accuracy of the classification and can be particularly harmful in scenarios where precise object recognition is critical. Adversarial attacks on 3D datasets can take various forms, such as geometric perturbations [10], which involve slightly altering the shape or structure of the object, and feature manipulation [9], where the attributes of the object, such as texture or color, are changed to confuse the classifier.

University of Waterloo, Ontario, Canada. rkharal@uwaterloo.ca
Wilfrid Laurier University, Ontario, Canada.
Artifact: https://github.com/saifaldin14/AdversarialAttack3DCNN
Wilfrid Laurier University, Ontario, Canada.
University of Waterloo, Ontario, Canada.

1946-0759/24/$31.00 ©2024 IEEE
DOI 10.1109/ICMLA61862.2024.00234

In the context of 3D model datasets, adversarial training requires (i) generating adversarial 3D inputs from all or part of the full training dataset, and (ii) performing adversarial training with the newly generated inputs. The retraining step on adversarially attacked inputs imposes a significant burden on the network, making it an unrealistic and often impractical defense strategy [11]. Generating adversarial inputs using 3D datasets is extremely slow, which typically limits this step to only a small fraction of the training dataset. Our investigation confirms these findings, as shown in Figure 1. When using all of the objects in the dataset as the basis for training, the total training time is on average $8.8x$ longer per epoch than when we apply training using the strategy proposed in this work. Our proposed defense method uses only the *best-performing objects* from each class as a basis for generating adversarial inputs and subsequent adversarial training. Notably, we do not observe any performance degradation with this approach. These results are further discussed in Section IV.

This study investigates the vulnerability of 3D deep-learning models to adversarial attacks. We employed various techniques to assess how such attacks impact model accuracy and integrity. Our results reveal significant weaknesses, demonstrating that adversarial attacks on a 3D dataset can compromise a deep learning network's performance by 20-22%. These findings highlight the urgent need for enhanced defensive strategies to safeguard 3D deep learning networks from adversarial threats. We propose an efficient defense strategy that mitigates the burden of time-consuming adversarial retraining by employing only the *best-performing objects* per class. Using the ModelNet40 3D dataset, we examine our defense strategy across various forms of common adversarial attacks and find significant improvements in prediction accuracy following the adoption of our approach into a 3D deep learning network.

## Contributions

The key contributions of this work are as follows:

1) **Comprehensive Analysis of Adversarial Attacks on 3D Deep Learning Networks**: This study provides an in-depth examination of the impact of adversarial attacks on 3D deep learning models using a 3D model dataset. By uncovering critical vulnerabilities unique to 3D networks, it highlights the pressing need for robust defense mechanisms in this domain.
2) **Development of a Simplified and Effective Defense Strategy**: We propose a novel defense approach against adversarial attacks that significantly simplifies the adversarial training process. By leveraging the *best-performing objects* per class, our method reduces computational overhead while maintaining, or even enhancing, model performance compared to traditional adversarial training methods.
3) **Investigation of Adversarial Attack Variants and Combinations**: The work conducts a thorough investigation of various forms and combinations of adversarial attacks on a 3D model dataset. This includes assessing their impact on the proposed defense strategy, and demonstrating its robustness and effectiveness across varying adversarial scenarios.

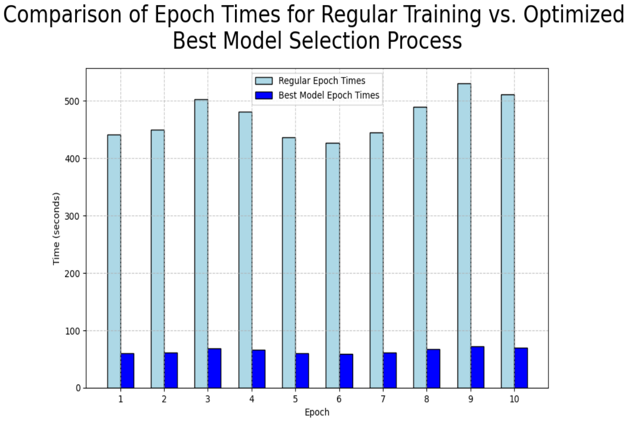

Fig. 1. Comparison of training times per epoch when using the full set of objects from the ModelNet40 dataset versus using *best-performing objects* only. The first set of bars shows the time taken per epoch when performing training on the full dataset. The second set of bars indicates the reduced training time when using the *best-performing objects*, demonstrating improved efficiency in the training process.

## II. Background and Related Work

### A. Adversarial Attacks

Adversarial attacks are methods of inducing misclassification of data points by a machine learning model, while remaining correctly classifiable by humans. Such attacks can be categorized into three models: white-box, black-box, and data-poisoning attacks. White-box and black-box adversarial attacks operate by introducing subtle perturbations into the set of data points, with the intent of causing the model to misclassify. Conversely, data poisoning attacks the training of a model, rather than the data points.

*1) White-box attacks:* In white-box adversarial attacks, the attacker has access to the parameters of the classifier. The most popular such attacks are the Fast Gradient Sign Method (FGSM), and Projective Gradient Descent (PGD), an iterative extension of FGSM. These attacks maximize a loss function (with respect to the true class) by applying gradient descent [12]. Gradient descent algorithms can be accelerated using momentum iterative methods (MIM), which work by incorporating the gradient at previous iterations to determine the optimal direction [13]. This method allows gradient descent to be resistant to poor local extrema.

*2) Black-box attacks:* In black-box adversarial attacks, the attacker is limited to labels and confidence scores from querying the model. A common black-box attack is the Gaussian Noise attack, which uses the particle swarm optimization genetic algorithm (GA-PSO) to find optimal adversarial examples, together with Gaussian noise [14].

*3) Data poisoning:* Another approach to attacking machine learning models, data poisoning, consists of injecting malicious training data into the model to influence the resulting classifier.

### B. Adversarial Defenses

Adversarial Defenses are methods to ensure machine learning models are resistant to adversarial attacks. These defenses can be categorized into active and passive defenses. Active defenses consist of ensuring a model is trained to be resistant to adversarial attacks. Common active defenses are adversarial training and network distillation. Passive defenses specifically target adversarial examples after a model has been developed, such as by adversarial detection or preprocessing inputs, without the cost of having to retrain a model.

*1) Network Distillation:* Network Distillation is a process by which the number of parameters of a model is reduced, while remaining accurate. The resulting neural network can be trained at a lesser computational cost. Papernot et al. found that distillation can significantly increase the robustness of a model, while remaining simple to implement [15].

*2) Adversarial Training:* Adversarial training, which involves training the classification model on both clean and adversarial examples, is one of the primary strategies used to mitigate the effects of an adversarial attack [16], [17]. This method requires iteratively solving optimization problems, which is computationally expensive and time-consuming [17]. Additionally, adversarial training decreases a model's efficacy on non-adversarial examples, introducing a balance between the accuracy of a model and its resistance to adversarial attacks. However, it adds the expensive overhead of (i) generating adversarial examples, (ii) retraining the network on training data modified to adversarial inputs (iii) increasing the model's generalization error. In the latter case, the model sacrifices accuracy on clean input data to protect against the possibility of inputs that have undergone an adversarial attack. Such factors can limit the practical applicability of adversarial training in real-time classification scenarios and time-sensitive applications. The burden of adversarial training is particularly intense on 3D model datasets where the size of the dataset and training a 3D deep learning network is already very expensive. Creating adversarial inputs and retraining on a proportion of the original 3D model dataset is simply unrealistic and often impossible to do with limited resources and hardware restrictions [11]. The limitations and resource-intensive nature of existing defense strategies such as adversarial training, have driven the search for more efficient and effective defense strategies in an increasingly data-driven world.

### C. 3D Machine Learning

The task of 3D shape classification has been explored across various representations of 3D objects, including point clouds, meshes, and voxels. 3D model data refers to the representation of objects in three-dimensional space, typically constructed using 3D meshes, which consist of vertices, edges, and faces that form the surface of an object, offering detailed geometric information essential for accurate classification tasks in fields such as autonomous driving, robotics, and augmented reality. The ModelNet40 dataset, is a dataset of models representing objects. It is available on Kaggle [18], is one of the most widely used benchmarks for 3D object classification and contains 40 categories of objects, including 9,843 models for training and 2,468 models for testing. Each object in this dataset is represented as a 3D mesh, which differs from point cloud data, where objects are captured as discrete, unstructured points in space. Research has shown that adversarial attacks on 3D mesh data, such as geometric modifications to 3D meshes, can significantly degrade classification performance, with success rates exceeding 99% when targeting classifiers like PointNet [19]. However, because these attacks are more perceptible in 3D data, defense strategies like frequency-domain filtering have been proposed to remove high-frequency perturbations and enhance the robustness of 3D mesh datasets against such attacks [20]. While adversarial attacks on 3D point clouds have received more attention, fewer studies have systematically examined attacks and defenses on mesh-based 3D models like those in ModelNet40. Nevertheless, recent research combining adversarial training with geometric transformations has shown promise in improving model resilience to adversarial threats. The goal of our research is to develop robust defense strategies against adversarial attacks on 3D model datasets that simplify the adversarial training process while providing a strong defense.

### D. Patch-Based Defense Strategies

Adversarial patches have been widely utilized in malicious attack scenarios, with the primary aim of degrading the accuracy of image classification systems [3] [17] [4]. However, the pioneering work by Salman et al. [21] introduced the innovative concept of "*unadversarial patches*", designed to enhance rather than undermine classification accuracy. Building on this foundational idea, subsequent research [22] explored a data-augmentation technique that leverages patches to selectively improve the classification accuracy in specific subsets of image classification problems. This method employs a streamlined patch creation process, where the highest-performing images within each class serve as the basis for augmenting a validation dataset with patches. The highest performing image per class is also known as the *best image* per class from the training step.

In our approach we do not use patches as a defense strategy against attacks in 3D space, however, we do utilize the notion of the *best* classified image per class. We extend the notion of *best images* to the scope of 3D model datasets in order to find the *best-performing objects* per class. This is explained further in section III.

## III. Methodology

In this study, we developed a method to study the impact of 3D adversarial examples designed to deceive object recognition systems within a 3D environment. We develop a defense strategy that involves adversarial training which extends the concept of *best images* from previous work on 2D patches [22] into the 3D model dataset domain. This allows for a simplified adversarial model defense process to be applied dynamically to 3D deep learning networks.

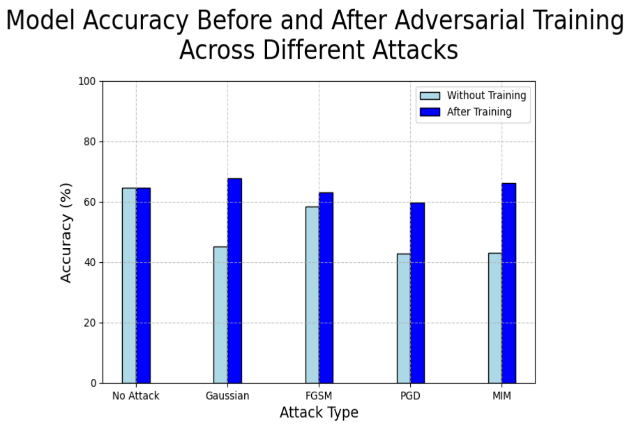

Fig. 2. Results showing the accuracy of the network *without* defense compared to the network with the proposed defense strategy before and after adversarial training. Adversarial attacks on an undefended network can compromise results by over 20%.

We explored the impact of these adversarial examples on the accuracy of object detection deep learning models, specifically focusing on their ability to manipulate model predictions under various attack scenarios. To achieve this, we utilized the ModelNet40 dataset, which contains 40 classes of 3D models. ModelNet40 is a dataset of 3D shapes commonly used in computer vision and machine learning. Each model in the ModelNet40 dataset is represented as a 3D mesh. A 3D mesh is a collection of vertices, edges, and faces that defines the shape of a 3D object.

The models are saved as '`.off`' (Object File Format) files. While comprehensive in their representation of mesh data, '`.off`' files can be extremely complex due to the variability between the number of vertices and faces across different objects. This irregular and variable-sized input data makes it difficult to apply standardized ML algorithms to the dataset. For this reason, the input data was first normalized and converted from 3D meshes to voxel grids. Voxel grids provide a uniform, consistent and fixed-sized data structure for the input data, thereby easily facilitating the application to standard 3D convolutional neural networks (CNNs). Furthermore, converting the data to voxel grids converts the complex surface representations of 3D meshes into binary occupancy grids, which are well-suited for tasks like object classification and recognition.

### A. 3D Deep Learning Model

We developed a custom 3D CNN for processing the voxel data generated by the ModelNet40 dataset. We refer to this network as the 3D-CNN. The network included four 3D convolutional layers followed by three batch normalization layers and ReLU activation functions. These layers served to extract features and spatial information from the input data. Our methodology included generating a *limited set* of geometric adversarial examples using PyTorch3D under various adversarial scenarios, followed by their integration into the training pipeline. The effectiveness of training with this sample set was tested under different attack types. We systematically evaluated the model's performance across varying attack strengths and combinations.

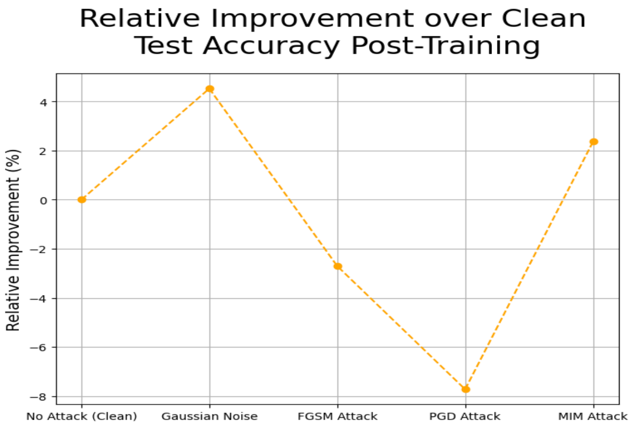

Fig. 3. The % improvement in network accuracy after adversarial training occurs using only *best-attacked objects* relative to baseline performance where no attack occurs on the network. Positive relative % improvement indicates the retraining process on *best-attacked objects* increases classification accuracy overall.

### B. Adversarial Attacks

The adversarial attacks employed in our study were Gaussian noise [14], Fast Gradient Sign Method (FGSM) [12], Projected Gradient Descent (PGD) [23], and Momentum Iterative Method (MIM) [13]. These methods have been extensively documented in the literature for their effectiveness in generating adversarial examples that can deceive neural networks [14], [23]–[25]. By incorporating these attacks, we ensure that our study is grounded in proven methodologies, allowing us to compare our findings with existing research and building upon the collective knowledge in this domain.

The FGSM, PGD, and MIM use a parameter, epsilon ($\epsilon$), to adjust the attack strength. The value of $\epsilon$ controls the magnitude of the perturbation applied to the input data during adversarial attacks. Epsilon defines the balance between the human-visible impact of the adversarial perturbation and the effectiveness of the attack. A larger $\epsilon$ value increases the perturbation, leading to stronger adversarial attacks that are more likely to mislead the neural network, while a smaller $\epsilon$ results in weaker perturbations, which may be harder to detect but are typically less effective at compromising the network's predictions.

The proposed method for generating and applying 3D adversarial examples involves a few key steps outlined below:

### C. 3D Adversarial Training Process

First, all training models from the ModelNet40 dataset are trained on our 3D-CNN classifier. As training occurs, we extract the set of objects that achieve the highest classification accuracy per class. We label this set as the *best-performing* *objects* and use these for subsequent steps. We apply varying adversarial attacks from Section III-B to the set of *best-performing objects* and now refer to this set as the *best-attacked objects*. There are 40 classes in the ModelNet40 dataset therefore, we have 40 *best-attacked objects* which are used to retrain the 3D-CNN. This step is referred to as adversarial training, however, we drastically reduce the amount of retraining required by employing only *best-attacked objects*. The training pipeline was modified to incorporate these adversarial examples. This ensures that during training, the network is exposed to both clean and adversarially perturbed 3D data.

### *D. Experiments*

In our experimental evaluation, we use three varied approaches in generating the *best-attacked objects*: (1) 1 adversarial attack is performed per model using 1 common attack parameter ($\epsilon$ value of 0.01), (2) Two attacks, one after another are performed on the *best-performing objects*, again using common attack parameters (3) We vary the attack strengths on the *best-performing objects* by testing across different $\epsilon$ values. Once the *best-attacked objects* are generated, and the 3D-CNN is retrained on these objects, we perform the same adversarial attack on the validation dataset that was used to generate the *best-attacked objects*. The validation dataset is now an adversarially attacked validation dataset (AAVD). We test the 3D-CNN on the AAVD and report results in the next section.

### *E. Evaluation and Analysis*

The performance of the 3D-CNN was evaluated using accuracy metrics on both clean and adversarially perturbed datasets. This included testing the network's robustness against various attacks, such as Gaussian noise, FGSM, PGD, and MIM, and comparing the results to the baseline accuracy.

To gain deeper insights into the vulnerability of deep learning networks to adversarial attacks, the network's performance was also evaluated using the AAVD *without* training on *best-attacked objects*. This approach helps to better understand how an undefended 3D deep learning network can be compromised by malicious attacks.

## IV. Experiments and Results

### *A. Baseline vs Adversarial Attacks*

The first experiment tested in the context of 3D adversarial attack scenarios is establishing a ***baseline***. In this baseline, the 3D-CNN is trained and evaluated on clean data from ModelNet40 without any form of adversarial attack taking place and no adversarial training step occurs. Additionally, we test scenarios where the network is trained on clean data but evaluated on adversarially attacked data, as illustrated in Figure 2. These results indicate the significant impact that adversarial attacks have on a 3D deep learning network used to predict 3D model data. The network, as is, is not designed to withstand such attacks.

As expected, all experiments where the 3D-CNN is subjected to adversarial attacks perform worse than the baseline experiment. However, the model under the FGSM attack showed the closest performance to the baseline compared to other attack types. The $\epsilon$ value used in these experiments is 0.001 which is the weakest attack strength. After establishing baseline performance across different attack scenarios, we employ adversarial training using *best-attacked objects* as outlined in Section III-C. Figure 2 illustrates that the network retrained with *best-attacked objects* consistently outperforms their non-adversarially trained counterpart. Figure 3 shows the network trained and evaluated against Gaussian noise and MIM even outperforms the baseline case, while the other attacks perform below the baseline.

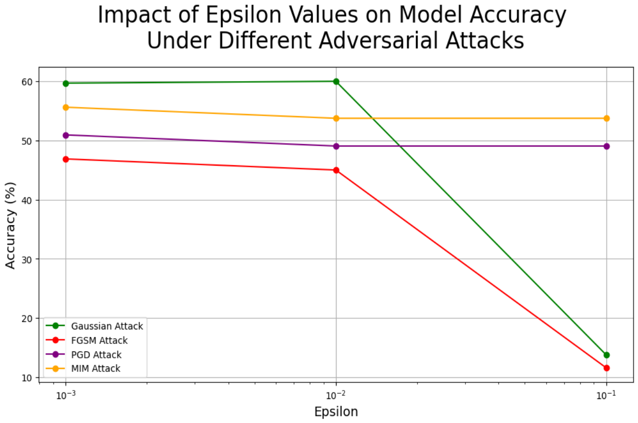

Fig. 4. Accuracy results from measuring network (3D-CNN) performance after adversarial attacks have been applied, testing adversarial attacks with varied attack strength (epsilon values).

What is clear from Figures 2 and 3 is the significant impact of adversarial attacks on the performance of the 3D-CNN, with prediction accuracy dropping by as much as 22%. The proposed defense strategy of retraining with *best-attacked objects* successfully restores performance to baseline levels (i.e., the unattacked network). The exception is the case of the PGD attack, where results fall slightly short of baseline performance. This discrepancy is related to a phenomenon known as generalization error, which can occur with adversarial training and was discussed in Section II.

### *B. Efficiency Test*

To empirically demonstrate the benefits of training using only the *best-performing objects*, we compared it to the traditional approach of training on the full dataset. Figure 1 illustrates the time required to traing using all objects in the dataset versus using only the *best-performing objects* from each class. We observe a significant reduction in training time when utilizing the smaller training dataset. The full dataset requires $8.8x - 9.2x$ more time than using the *best-performing objects* only. The *best-performing objects* reduce training time by approximately 90%.

It is important to emphasize that our approach—using the *best-performing objects* per class—is fundamentally different

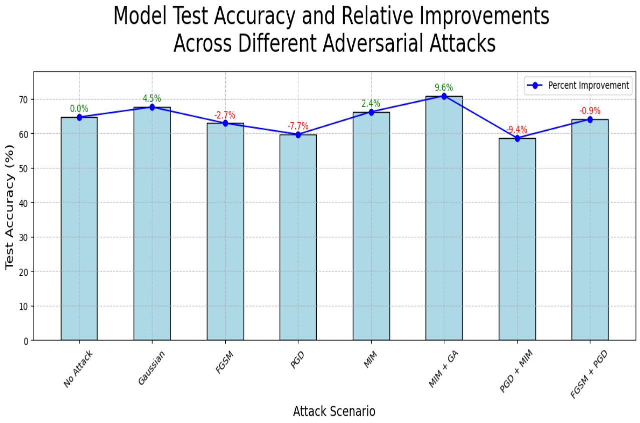

Fig. 5. Accuracy results from varied attack scenarios (single attack, composite attacks) and the percent improvement in accuracy relative to baseline.

from using all of the objects from the full dataset. By focusing on the most representative objects, we significantly reduce the time and computational resources required for adversarial training while achieving results that match the baseline performance. This selective strategy offers a highly efficient defense without sacrificing accuracy or robustness.

### C. *Attack Strengths*

We conducted further tests by evaluating the network under attacks of varying strengths, using higher epsilon values. As shown in Figure 4, all experiments experienced their largest drop in accuracy at epsilon values of 0.1. FGSM and Gaussian noise performed particularly poorly at these higher epsilon values. Reducing epsilon to 0.001 led to slight improvements compared to 0.01, though these gains were marginal. Next, we evaluated the network's performance when combining multiple attacks in sequence, relative to the baseline, as depicted in Figure 5. In these experiments, two adversarial attacks were applied to the *best-performing objects* to generate the *best-attacked objects*. Valid attack combinations tested included MIM and Gaussian noise, PGD and MIM, and FGSM and PGD. Other combinations resulted in errors due to incompatible perturbations between different attacks, causing the network to crash. Overall, the test accuracy across all experiments, whether subjected to single or combined adversarial attacks, remained similar to the baseline model without any attacks. This indicates that applying two attacks to the *best-performing objects* is unnecessary. Notably, in one case (Gaussian-MIM), we observed a performance improvement, surpassing baseline accuracy by 9.6%.

These experiments demonstrate that employing a simplified defense strategy using 3D adversarial training with *best-attacked objects* significantly improves the performance and prediction accuracy of a 3D deep learning network compared to using no adversarial training at all. Specifically, combining different attacks, such as MIM and Gaussian noise, enhanced the network's resiliency and accuracy relative to the baseline.

## V. Conclusions and Future Work

In this work, we investigate the impact of adversarial attacks in the 3D data domain. We illustrate that 3D deep learning networks are highly susceptible to performance degradation by the use of small perturbations in the dataset that throw off the network. Our experiments showed that even the simplest adversarial attacks, such as FGSM, PGD, and MIM, significantly degrade the performance of 3D convolutional neural networks (3D-CNNs) used for object recognition in 3D environments. Attacks using minimal perturbation parameters resulted in accuracy drops of up to 22%, demonstrating the vulnerability of these models when subjected to adversarial perturbations.

We presented a novel and efficient defense approach to defending 3D deep learning networks against adversarial attacks. We extend the concept of *best images* from 2D to 3D adversarial training using *best-attacked objects*. Our defense strategy is far more efficient in 3D space where datasets and 3D deep learning paradigms are particularly resource-intensive. This approach was shown to significantly improve model robustness while reducing the retraining time by up to 90% compared to traditional adversarial training on the full dataset. Additionally, we demonstrated that combining different attack types, particularly MIM and Gaussian noise, yielded superior results in terms of accuracy surpassing baseline performance.

The results from this study validate the effectiveness of our simplified and efficient adversarial training process using *best-attacked objects*, especially in mitigating the impact of adversarial attacks on 3D object detection frameworks. Importantly, the proposed defense strategy successfully restored network performance to baseline levels in most cases, highlighting its potential as a practical solution for defending real-world 3D deep learning applications.

### A. *Future Work*

While this study provides valuable insights into the effectiveness of adversarial training in 3D environments, with a focus on 3D model datasets, there are several avenues for future work. First, further exploration of adversarial attack types and their interactions in 3D space is necessary. This includes investigating more sophisticated and adaptive attack strategies, which could further challenge the robustness of 3D deep learning networks. Additionally, the defense strategy presented in this work will be extended to handle other types of 3D input data formats beyond voxel grids, such as point clouds and surfaces. The aim is to provide a broader understanding of how well adversarial training using *best-attacked objects* can be adapted across different data representations.

Overall, the work presented in this study lays the groundwork for future advancements in defending 3D deep learning models against adversarial attacks, and we believe that our proposed methodology will serve as a strong foundation for further exploration in this domain.

## Acknowledgment

The authors of this work would like to thank Trevor Brown for his guidance and support. This work was supported by: The Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Program grant: RGPIN-2019-04227, the Canada Foundation for Innovation John R. Evans Leaders Fund with equal support from the Ontario Research Fund CFI Leaders Opportunity Fund: 38512, NSERC Discovery Launch Supplement: DGECR-2019-00048, and the University of Waterloo.

## References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv*, 2013. [Online]. Available: https://arxiv.org/abs/1312.6199

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

[3] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2018. [Online]. Available: https://arxiv.org/abs/1712.09665

[4] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Transactions on Privacy and Security (TOPS)*, vol. 22, no. 3, pp. 1–30, 2019.

[5] Z. Li, Z. Tu, B. Cheng, W. Zhang, and J. Lu, "Dynamic adversarial attacks on autonomous driving systems," *arXiv preprint arXiv:2006.13555*, 2020. [Online]. Available: https://arxiv.org/abs/2006.13555

[6] S. Zheng, W. Liu, S. Shen, Y. Zang, C. Wen, M. Cheng, and C. Wang, "Adaptive local adversarial attacks on 3d point clouds," *Pattern Recognition*, vol. 144, p. 109825, Dec 2023.

[7] K. He, H. Wang, and J. Chen, "Universal adversarial attacks on deep neural networks for medical image classification," *BMC Medical Imaging*, vol. 20, pp. 1–13, 2020.

[8] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019. [Online]. Available: https://www.science.org/doi/10.1126/science.aaw4399

[9] H. Naderi and I. V. Bajić, "Adversarial attacks and defenses on 3d point cloud classification: A survey," *IEEE Access*, 2023.

[10] J. Zhang, L. Chen, B. Liu, B. Ouyang, Q. Xie, J. Zhu, W. Li, and Y. Meng, "3d adversarial attacks beyond point cloud," *Information Sciences*, vol. 633, pp. 491–503, 2023.

[11] X. Zeng, C. Liu, Y.-S. Wang, W. Qiu, L. Xie, Y.-W. Tai, C.-K. Tang, and A. L. Yuille, "Adversarial attacks beyond the image space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4302–4311.

[12] J. Ma, J. Zhang, G. Shen, A. Marshall, and C.-H. Chang, "White-box adversarial attacks on deep learning-based radio frequency fingerprint identification," May 2023.

[13] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," Jun 2018.

[14] Y. Wang, Y. a. Tan, W. Zhang, Y. Zhao, and X. Kuang, "An adversarial attack on dnn-based black-box object detectors," *Journal of Network and Computer Applications*, vol. 161, p. 102634, 2020.

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016.

[16] W.-A. Lin, C. P. Lau, A. Levine, R. Chellappa, and S. Feizi, "Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks," 2020.

[17] C. Liu, M. Salzmann, T. Lin, R. Tomioka, and S. Süsstrunk, "On the loss landscape of adversarial training: Identifying challenges and how to overcome them," in *Advances in Neural Information Processing Systems*, 2020.

[18] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[19] C. Xiang, C. R. Qi, and B. Li, "Generating 3d adversarial point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 9136–9144. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Xiang_Generating_3D_Adversarial_Point_Clouds_CVPR_2019_paper.html

[20] Anonymous, "Lpf-defense: 3d adversarial defense based on frequency analysis," *PLOS ONE*, vol. 17, no. 8, p. e0271388, 2022. [Online]. Available: https://doi.org/10.1371/journal.pone.0271388

[21] H. Salman, A. Ilyas, L. Engstrom, S. Vemprala, A. Madry, and A. Kapoor, "Unadversarial examples: Designing objects for robust vision," 2021.

[22] R. F. Kharal, "Towards augmentation based defense strategies against adversarial attacks," in *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2023, pp. 1430–1437.

[23] L. M. P and M. Gunasekaran, "Adversarial attacks against neural networks using projected gradient descent with line search algorithm," May 2023.

[24] S. M. Naqvi, M. Shabaz, M. A. Khan, and S. I. Hassan, "Adversarial attacks on visual objects using the fast gradient sign method," *Journal of Grid Computing*, vol. 21, no. 4, Sep 2023.

[25] J. Zhang, W. Qian, R. Nie, J. Cao, and D. Xu, "Generate adversarial examples by adaptive moment iterative fast gradient sign method," *Applied Intelligence*, vol. 53, no. 1, p. 1101–1114, Apr 2022.