Evaluating the Impact of Noisy Data on Time-Sensitive Point Clouds from Millimeter Wave Gesture Recognition Systems

Paul Jiang
Purdue University
West Lafayette, USA
jiang861@purdue.edu

Ellie Fassman
Cornwell University
Ithaca, USA
ofassman@iu.edu

Tao Li
IUPUI
Indianapolis, USA
tli6@iupui.edu

ABSTRACT

Point cloud data gathered through millimeter wave sensors has garnered increasing attention for its critical applications, including automotive radars, security systems, and notably, gesture recognition. It provides a non-intrusive and robust approach towards human-computer interactions; however, its reliance on real-time data makes resilience of paramount concern. Attacks on millimeter-wave sensors can have catastrophic effects. From real-time spoofing to data poisoning attacks or even just imperfect or poor data, systems based on 2D and 3D point cloud machine learning models can be extremely vulnerable. Despite this, there exist few studies prioritizing the robustness of time-sensitive point clouds. This study presents an in-depth examination on the effects of noisy data on frame based time-sensitive point clouds used in millimeter wave gesture recognition machine learning models. Noisy data can be introduced during the training stage where imperfect data is fed to the model, causing this model to misclassify test-time samples and lower the overall accuracy of the model. We stage and evaluate the impact of four different, simple data noising scenarios to observe vulnerabilities within this system and to emphasize the importance of robust machine learning models. Noisy databases are particularly relevant to deep learning systems because these models need large amounts of data to train, many of which commonly scraped from the internet with little to no manual inspection. Our findings highlight the importance to not only dedicate time and research towards innovations in mmWave gesture recognition, but also towards the robustness and resiliency of these systems in order to proactively prevent destructive effects.

KEYWORDS

gesture recognition, time-sensitive point clouds, machine learning, classification of point clouds, millimeter waves, noisy data, cybersecurity

ACM Reference Format:
Paul Jiang, Ellie Fassman, and Tao Li. 2023. Evaluating the Impact of Noisy Data on Time-Sensitive Point Clouds from Millimeter Wave Gesture Recognition Systems. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In recent years, the deployment of millimeter wave (mmWave) technologies in combination with exponential advancements in deep learning has brought a new wave of wireless communication and sensing systems. These systems leverage the unique characteristics of mmWave frequencies to achieve high data rates and enhanced spatial resolution, making them useful for a wide range of applications, including 5G networks, autonomous vehicles, and advanced radar systems [1].

One developing application of mmWave technology is in gesture recognition using time-sensitive point clouds [3]. Point clouds, comprising three-dimensional data points, serve as fundamental representations for object detection, localization, and mapping in various real-world scenarios. However, the accuracy and reliability of these applications heavily rely on the quality and integrity of the underlying point cloud data.

As mmWave-based point clouds become increasingly prevalent in critical domains such as autonomous vehicles and advanced radar systems, ensuring robustness of these systems is of the utmost importance. This research focuses on the vulnerability of specifically, time-sensitive and frame based gesture recognition systems under three distinct neural networks (long short-term memory, convolutional, and transformer). Despite the numerous benefits offered by mmWave-based point clouds, their susceptibility to adversarial attacks and noisy databases is a growing concern. In contrast to sophisticated and highly targeted data poisoning attacks, noisy data is simpler yet still potentially impactful. These noisy data set scenarios involve the introduction of noise or perturbations into the raw point cloud data, be it deliberate or accidental, which disrupts the accuracy and reliability of the underlying processing algorithms. Imperfect data labeling is a large issue since typical data labeling for large data sets is outsourced and can be subject to errors. Due to the scale of prospective data sets and their dynamic nature (in the case of our research), the annotation process is inherently complex and subsequent labeling is often conjoined with noise. In addition, errors in data collection can lead to noisy data. For example, in gesture recognition, different postures of the person articulating each gesture and different environments can contribute to noisy data sets. Furthermore, faulty or even misaligned equipment can similarly result in flawed point clouds. These errors in data collection and labeling can lead to models misclassifying gestures during test-time, lowering and potentially compromising the overall accuracy of gesture recognition systems. While there exist many studies examining the utilities of these technologies, numerous data poisoning attacks [2][4][6], and even explorations into adapting frameworks to be robust against perturbations [5], we are unaware of any investigating the robustness of dynamic and time-sensitive

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

systems, particularly gesture recognition, under various models and noise.

The four types of noisy data scenarios we induce in this study are: mislabeling, rotated point clouds, missing frames, and misordered frames. To represent mislabeling, we apply simple label flipping to invert the labels of a percentage of frames to induce misclassification. In order to simulate misalignments and variations in the creation of training data, we introduce rotated point clouds: a rotation of coordinates within point cloud frames by a parameterized angle. To replicate missing data scenarios and faulty equipment, we script the removal of frames from critical gestures within the training data. Finally, to explore the effects of disruptions in temporal flow and structure of data samples, we use seeded randomization to shuffle the order of frames and affect gesture recognition.

While data noising situations and imperfect data might not be as subtle or stealthy as their more complex counterparts, they can still lead to detrimental consequences in time-sensitive applications. The injected noise can distort the geometry of the point clouds, mislead object detection algorithms, and compromise localization accuracy, ultimately putting the safety and performance of the entire system at risk.

This research paper focuses on the impact of noisy data on time-sensitive mmWave point clouds in a mid-air gesture classification system. We induce data noising and subsequently review classification accuracy of various model as understanding these realistic scenarios is paramount to developing robust defenses against them.

2 BACKGROUND

2.1 mmWave

Millimeter waves (mmWaves) are a portion of the electromagnetic spectrum that falls within the microwave frequency range. Their wavelengths typically range from 1 to 10 millimeters (frequencies between 30 and 300 gigahertz). Applications of mmWaves include wireless communication, radar systems, imaging, and sensing. For example, mmWave sensing can be used for occupancy sensing, through-wall sensing, and gesture recognition.

2.2 Point Clouds

2D and 3D point clouds are a representation of two to three dimensional data composed of individual points in a coordinate system. Each point in the point cloud is primarily defined by its X, Y, and sometimes Z coordinates, representing its position in space. Point clouds capture the geometric information of objects and scenes, making them valuable for various applications in computer vision, augmented reality, and autonomous vehicles. One can easily leverage spatial information through point clouds which can aid in the classification of gestures such as biannual and circular gestures.

2.3 Models

The three most common model options for gesture based recognition systems are convolutional neural networks (Conv), long short-term memory networks (LSTM), and transformer neural networks (Trans). Convolutional neural networks are a class of deep learning models specifically designed for processing and analyzing visual data. The key components of convolutional neural networks are convolutional layers, pooling layers, activation functions, fully connected layers, training, and backpropagation. Long short-term memory networks are a type of recurrent neural network designed to handle sequential data (which is particularly applicable to time-sensitive point cloud data). Long short-term memory networks consist of specialized memory cells and gates that control the flow of information. Transformer neural networks are a type of deep learning architecture. The transformer model aids in natural language processing and various other sequence-to-sequence tasks by using a self-attention mechanism without using recurrent or convolutional layers. Transformer networks consist of an encoder and a decoder, which both use layers of self-attention and feed-forward neural networks. The encoder processes the input sequence while the decoder generates the output sequence in sequence to sequence tasks. In this research we study the accuracy of gesture classification associated with these three different models and investigate the robustness of each model to noisy data sets.

3 RELATED WORKS

3.1 Pantomime

Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds lays the framework for mid-air gesture recognition systems. Pantomime uses a hybrid model architecture for optimized spatio-temporal feature extraction which is designed to recognize sparse motion gestures [2]. In the classification system, local features are first extracted. This process is iterative until features of the whole point cloud are computed. Multiple set abstraction levels are used to mimic the multiple convolution levels in CNNs. Pantomime uses 21 types of mid-air gestures including bimanual, linear, and circular gestures. Pantomime provides real-time recognition and achieves 95% accuracy of classification for the 21 gestures.

3.2 Learning With Noisy Labels

Different methods have been proposed in mitigating the impact of noisy data on model accuracy. A recent work introduces the Point Noise-Adaptive Learning (PNAL) framework, tailoring its strategies to the nuances of point cloud data, such as spatially variant noise rates. PNAL incorporates novel methodologies, including point-wise confidence selection based on historical predictions and cluster-wise label correction to enhance the accuracy of model training with noisy labels, leading to improved performance, even in scenarios where a significant portion of the labels is inaccurately annotated. [5].

3.3 Additional Studies

Several studies have explored the vulnerabilities and potential impact of targeted data poisoning attacks on mmWave-based point clouds, leading to valuable insights and defense mechanisms. Some related research in this area includes the following papers. Defending against 3D Adversarial Point Clouds via Adaptive Diffusion in which the authors proposed a defense strategy against simple data noising attacks on mmWave point clouds [6]. Leveraging adversarial training, they trained point cloud processing models with augmented datasets containing adversarially noised samples. The results showed resilience to data noising attacks. In Shape-invariant

3D Adversarial Point Clouds, the researchers introduced shape-invariant perturbations, which imposed minimal changes to point cloud geometry while causing significant misclassification [3].

This research highlights the growing concern over simple data-based attacks on time-sensitive mmWave point clouds. Consequently, researchers have been actively exploring defense strategies, detection methods, and robust algorithms to ensure the resilience of mmWave-based applications in the face of such attacks.

4 PROBLEM SETTING

In this study, we investigate the robustness of a time-sensitive point cloud gesture recognition system on three common models (LSTM, Conv, Trans) in the presence of different types of noisy data. Gesture recognition plays a critical role in human-computer interaction, enabling natural and intuitive control of various applications. However, real-world scenarios often introduce various forms of noise during training time that can degrade the performance of gesture recognition systems. We focus on the following types of noisy data:

(1) Rotation Noise: Variations in device orientation or gesture execution may lead to slight rotations in point clouds, affecting the system's ability to accurately recognize gestures.
(2) Mislabeled Data: Noise introduced by incorrect gesture labels in the training data set can result in confusion during recognition, impacting the system's reliability. This can occur at various stages during the training process.
(3) Frame Loss: Missing or incomplete frames in the input point cloud sequence could disrupt the temporal context and challenge the system's ability to maintain accurate recognition over time; it can often be introduced through faulty equipment.
(4) Unordered Frames: Disordered frames in the input sequence may disrupt the temporal sequence, requiring the system to handle out-of-order data.

5 METHODOLOGY

To comprehensively assess the robustness of time-sensitive gesture recognition systems, we conducted extensive experimentation using a vast dataset comprising 7402 point cloud sequences encompassing nine distinct gestures: up, down, left, right, clockwise, counterclockwise, s, x, z – the last three gestures formed through tracing the respective letter in the air. Each sequence consisted of between 10 to 20 point cloud "frames" in order to induce temporal structure. The training set was formed from 70% of this data while the test set, the remaining 30%. To mimic real-world noisy scenarios, we employed data augmentation during training, introducing four distinct types of noise (rotation, mislabeling, frame loss, and unordered frames) to the clean dataset. In each case, we introduced controlled variations into the data, and employed equal testing on LSTM (Long Short-Term Memory), Convolutional, and Transformer models for our evaluations on robustness. We trained each model over 10 epochs, subsequently testing them on the remaining clean data to obtain our accuracy for each trial. 10 epochs was chosen due to resource and time constraints; training and testing on an entirely clean data set with these specifications resulted in between 93 - 97 percent accuracy on all models.

Table 1: Accuracy on clean training set

Model	Baseline Validation Accuracy (%)
LSTM	97.21
Conv	95.5
Trans	93.2

In the case of mislabeled data, we conducted isolated experiments, systematically incrementing the amount of noise in the data set by 10%, eventually reaching a scenario with 100% mislabeled data. To deliberately induce mislabeling, we employed an algorithmic approach to interchange each gesture with its opposing label. For the letter gestures, we implemented a circular swapping strategy to further augment the mislabeling process.

For rotational noise assessment, we systematically applied rotations ranging from 15 to 90 degrees in intervals of 15 degrees, replicating conceivable perturbations in practical applications. These modifications were evaluated under two scenarios: one with an entirely noisy data set and another with 50% of the data seeded randomly to be afflicted by noise.

To assess the impact of frame loss, we simulated the random removal of frames at intervals of 25%, 50%, and 75% across noise levels of 25%, 50%, 75%, and 100% in order to ensure a wide array of measurements. The increments for this scenario were larger due to the added variable of frame loss percentages leading to a larger number of trials in combination with time contraints.

Lastly, in the context of unordered frames, we introduced randomly seeded shuffling of frames to disrupt the temporal sequence of frames within data set subsets. We similarly conducted experiments at 10 intervals, progressively increasing the noise levels from 10% to 100%.

The primary evaluation metric employed was accuracy (dividing the number of correct predictions by the total number of predictions) quantifying the system's correct recognition of gestures amidst noisy conditions. To ensure results, each experiment was repeated three times, and average accuracy was computed. We decided against a greater number of trials due to time constraints.

In summary, our experimental setup entailed training the LSTM, Convolutional, and Transformer models on the augmented datasets with varying noise levels, followed by rigorous testing and performance assessment. This comprehensive evaluation framework aimed to provide insights into the resilience of gesture recognition systems in real-world noise-ridden scenarios, offering valuable perspectives for improving their practical utility and performance.

6 EVALUATION

To start, we compared different deep learning model options. We wanted to investigate how noisy data impacted the validation accuracy of the system using these three different models: convolutional neural networks (CNN), long short-term memory networks, and transformer neural networks. CNN's are mainly used for image processing and object detection. It is clear why using a CNN would be beneficial in gesture classification. Long Short Term Memory Networks (LSTM's) can learn long-term dependencies because they

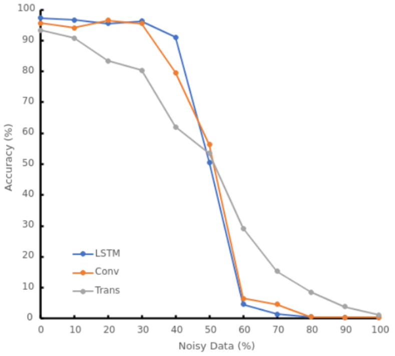

Figure 1: Accuracy with mislabeled gestures

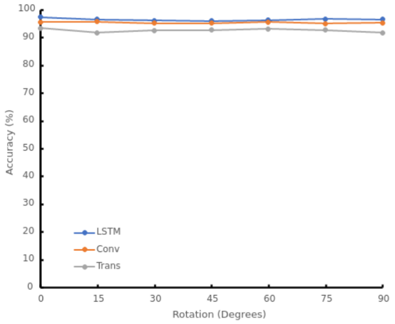

Figure 2: Accuracy with rotation on 50% of data

retain information over time. LSTM's are advantageous in time-series prediction so they pose an advantage over CNN's because we are working with time-sensitive point cloud data. Transformer neural networks (transformers) attain high accuracy in classification due to their ability to learn contextual relationships between input data, allowing for more accurate predictions.Transformers are designed to handle sequential data, but do not require that the sequential data be processed in order. Thus transformers are able to find the relationship between sequential elements when these elements are out of order.

Looking at the clean data (without data noise injected), we can see from Table 1 that the baseline validation accuracy is highest for LSTM and lowest for Trans, though all are over 90% accurate. Validation accuracy is the accuracy of the gesture recognition on the clean test data set.

Beginning with label flipping, we can see in Figure 1 that all three of the models drop from high accuracy to an accuracy consistent with guessing when 50% of the data has been tampered with. After 60% of data and beyond, the accuracy plummets to less than 20% of gestures correctly classified for all three models. Interestingly, the LSTM had the highest accuracy in low data noising situations, but the lowest accuracy in high data noising situations. Conversely, the transformer model was more stable in its performance throughout the experiments.

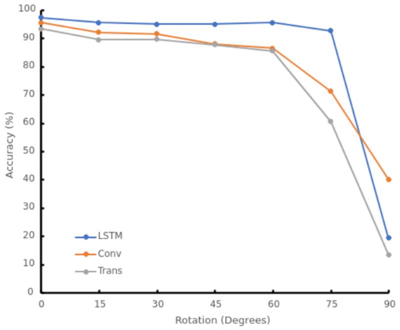

Figure 3: Accuracy with rotation on entire data

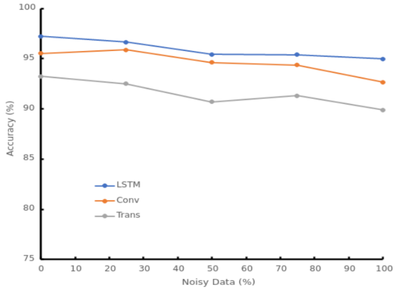

Figure 4: Accuracy with 25% frame loss

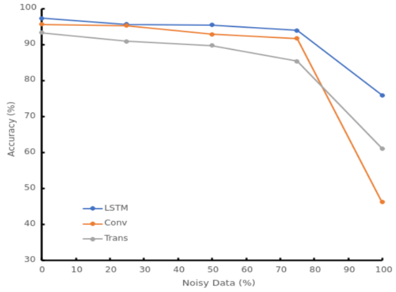

Figure 5: Accuracy with 50% frame loss

Figures 2 and 3 show the effect that coordinate rotation has on accuracy. With 50% of the data rotated, all three models perform remarkably well with no model having an accuracy below 90% for all the rotation angles. With 100% of the data rotated we can see that all three models perform well until a rotation of about 70 degrees, when the accuracies of all three models drastically drop. In comparing the three models against each other, all three perform comparably with each other, though with the convolutional neural network attaining the most stable classification accuracy when 100% of the data is rotated.

We can see from Figures 4-6 the impact that frame deletion has on validation accuracy. With 25% loss all three models have stable performance even when 100% of the data is impacted, with all the models achieving accuracy above 90%. When there is 50% of frames deleted, there is a clear accuracy drop when 75% or more data is

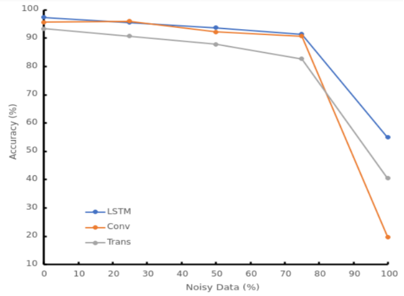

Figure 6: Accuracy with 75% frame loss

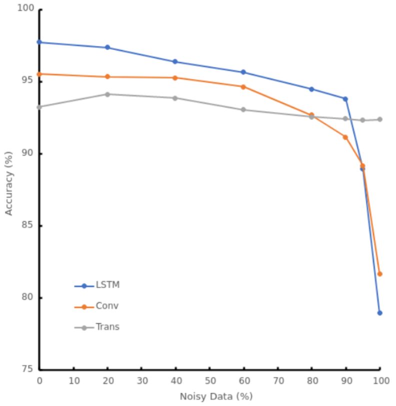

Figure 7: Accuracy with unordered frames

impacted. Interestingly, when there is 75% frame loss, all three models have a very similar trend, dropping in accuracy when 75% of the data is impacted. However, we can see that when 100% of the data has frame loss, the accuracies attained by all three models are lower than in the 50% loss case. In both the 50% and 75% loss cases, the LSTM model performs the best when a high percentage of the data is impacted, and the CNN model performs the worst.

When looking at the impact that frame scrambling has on validation accuracy in Figure 7, we can see that all three models perform well (over 90% accuracy) up until 90% of the data is impacted. Then, both LSTM and CNN are impacted, while the transformer model stays stable above 90% accuracy. Counterintuitively, all three models achieve a high accuracy even when a vast majority of the data has scrambled frames. For the transformer model, this makes more sense because as mentioned above, transformers do not require that the sequential data be processed in order. However, both the LSTM and CNN models perform well even when 100% of data is scrambled (still attaining over 75% accuracy). This may suggest that the time-sensitivity of this data is not extremely important. All three models classify gestures based on physical and temporal features of the frame sequences. If the accuracy is only mildly affected by frame scrambling, this implies that this gesture classification relies heavily on the physical features of the frame sequences and not the temporal features. This claim is justified by the fact that all three models achieve remarkably high accuracy when 100% of the data is impacted by frame scrambling. This suggests that gesture recognition systems relying on time-sensitive point clouds only depend heavily on the physical features of these point clouds and do not actually depend very much on the temporal features of these sequences. This could mean that gestures can be compressed into only one or two frames, rather than kept as a sequence of frames, since these gestures can still be accurately classified. Training and testing the gesture recognition system on fewer frames would help save on computation and time requirements of this system.

7 LIMITATIONS AND FUTURE WORKS

While this study provides valuable insights into the robustness of time-sensitive gesture recognition systems utilizing mmWave-based point clouds, it is important to recognize certain limitations that shape the scope of our findings. Firstly, our examination of noisy data scenarios was intentionally simplified to facilitate controlled experimentation. In real-world scenarios, noise can exhibit intricate and unpredictable patterns, potentially even merging various types of noise and yielding distinct effects on the system's behavior. Furthermore, the uniform distribution of noise across our dataset might not accurately mirror the variability of noise patterns encountered in these real-world environments. As a result, the translation of our controlled scenarios to actual noisy data occurrences in real world settings should be approached with caution.

Secondly, the focus of our investigation centered on a specific set of nine distinct and relatively simple gestures. While these gestures serve as foundational examples, the applicability of our findings to a broader array of gestures and intricate interactions warrants further exploration. The influence of dataset specifics on our results cannot be overlooked, and examining robustness across a more diverse range of datasets could provide a richer context for understanding the generalizability of our conclusions.

Our study also employed a specific selection of model architectures—CNNs, LSTMs, and Transformers—to assess robustness. Other architectures, which were not explored in this study, could potentially offer different perspectives on the impact of noisy data; future investigations could encompass a broader spectrum of architectures to attain a more holistic understanding.

Finally, due to time constraints, we were unable to further research methods in which to improve model robustness. In the future there should be research on techniques applicable in making gesture recognition systems more robust to noisy data. Potential paths may be accomplished through data cleaning, integration, transformation, or reduction. Data cleaning is the process of filling missing values, smoothing and removing noisy data and outliers. Data integration means integrating data from multiple sources. Data transformation is normalization and aggregation of data. Data reduction reduces the number of attributes and dimensions of the data. In the case of this time-sensitive point cloud data, we have found that the time sequences of frames is not very important in classification accuracy over a Transformer neural network. This means that the dimensionality of the gesture point clouds can be reduced. Developing classification systems using the above techniques will help these systems be more robust and resilient to data noise and data poisoning attacks.

In conclusion, while our study contributes valuable insights into the challenges of noise-induced robustness in mmWave gesture recognition, the outlined limitations underscore the need for careful interpretation of our results. Acknowledging these limitations creates the path for future research endeavors to delve deeper into the complexities of robustness and to cultivate a more comprehensive comprehension of the practical implications of noisy data on time-sensitive point cloud systems.

8 CONCLUSION

In our research, we underscore the critical importance of robustness in time-sensitive gesture recognition systems using millimeter wave point cloud data. Through controlled experiments, we explore the effects of distinct noise types – label flipping, coordinate rotation, frame loss, and unordered frames – on gesture recognition accuracy. Our findings reveal that even simple noise scenarios can significantly impact accuracy, highlighting the vulnerability of these systems. However, recognizing our study's limitations, such as controlled data noise and specific model architectures, we advocate for a comprehensive investigation into complex noise patterns with more diversity in time-sensitive datasets.

As gesture recognition continues shaping various domains, addressing noisy data implications remains paramount. By fortifying systems against noise, we pave the way for seamless human-computer interactions and heightened safety across critical applications.

ACKNOWLEDGMENTS

To Dr. Feng Li and Sheila Walter for the bagels, guidance and support throughout this research process. Additional thanks and gratitude to the National Science Foundation for providing and funding this research and opportunity.

REFERENCES

[1] Saifullahi Aminu Bello, Shangshu Yu, and Cheng Wang. 2020. Review: deep learning on 3D point clouds. arXiv:2001.06280 [cs.CV]

[2] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Hang Zhou, Weiming Zhang, and Nenghai Yu. 2022. Shape-invariant 3D Adversarial Point Clouds. arXiv:2203.04041 [cs.CV]

[3] Sameera Palipana, Dariush Salami, Luis A. Leiva, and Stephan Sigg. 2021. Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 27 (mar 2021), 27 pages. https://doi.org/10.1145/3448110

[4] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. 2021. Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks. arXiv:2006.12557 [cs.LG]

[5] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 2021. Learning with Noisy Labels for Robust Point Cloud Segmentation. arXiv:2107.14230 [cs.CV]

[6] Kui Zhang, Hang Zhou, Jie Zhang, Qidong Huang, Weiming Zhang, and Nenghai Yu. 2022. Ada3Diff: Defending against 3D Adversarial Point Clouds via Adaptive Diffusion. arXiv:2211.16247 [cs.CV]