# Enhanced AI On-the-Edge 3D Vision Accelerated Point Cloud Spatial Computing Solution

Gaurav Kumar Wankar[a] and Shubham Vohra[b]
*Neal Analytics Services Private Limited, Pune, Maharashtra, 411014, India*

Keywords: Industry - 4.0, Industry - 5.0, Digital Transformation, Business Transformation, Disruptive Technologies, 3D Vision, Product Market, Immersive Business Solutions, Artificial Intelligence, Deep Learning, Voxelization, PointNet, Pointpillars, Point Cloud, NVIDIA Jetson Tx2, Azure Kinect DK, GPU, Edge Applications, Edge AI, AI On-the-Edge, Transformative Experiences, Smart Everything Revolution.

Abstract: With the emergence of Industry - 5.0, the 3D Vision Product Market is growing rapidly. Leveraging Disruptive Technologies, we are exploring Artificial Intelligence driven Advanced 3D Vision immersive Business Solutions with transformative experiences leveraging Deep Learning accelerated with Voxelization, PointPillars and PointNet approaches for classification of Point Clouds enhancing the feature extraction to be more accurate bringing our work and data to life. NVIDIA Jetson Tx2 targeted at power constrained AI on-the-edge applications maintains awareness of its surroundings by visualizing in 3D space leveraging Azure Kinect DK depth sensing instead of 2D space thereby improving the performance in Edge AI computing device. Leveraging state of the art technologies converging AI and Mixed Reality we further encourage the readers to explore the possibilities of Next Generation services bringing accurate and immersive real-world information allowing decision-making based on Digital Reality driving Digital Transformation.

## 1 INTRODUCTION

With the beginning of 2nd quarter of 2021, the Business market is evolving rapidly exploring Digital Transformation. Now it is time for us to start thinking about reshaping the Organizational future course leveraging Industry - 4.0 solutions towards Industry - 5.0. While all the previous industrial revolutions are about utilizing technology to optimize and better the means of production. Industry 5.0 exclusively focuses on alliance between humans and smart systems.

As per the Gartner Hype cycle (Gartner, 2021) as shown in Figure 1.1, the current market is exploring towards the three themes. Interfaces and experiences, Business enablers and Productivity revolution.

Leveraging few of these Disruptive Technologies towards Edge AI, in this paper we are exploring new business solutions creating transformative experiences with cutting-edge spatial computing capabilities leading the Next Generation of Human Machine Interaction.

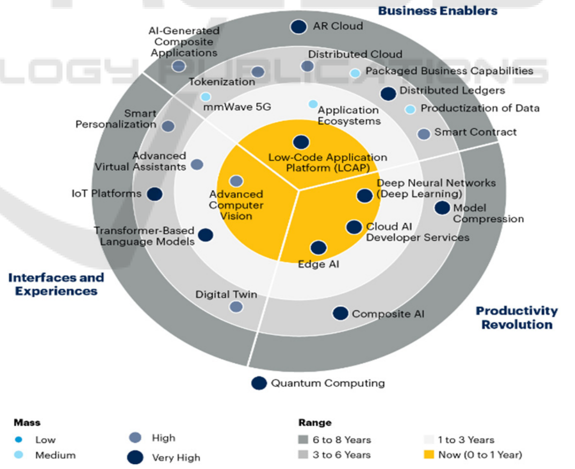


Figure 1.1: Impactful and Emerging Technologies from Gartner Trends for 2021 (Gartner, 2021).

With the advent of Industry - 4.0, the 2nd wave of computing is driven by GPUs for Edge Analytics and AI on-the-edge applications.

Designing hardware accelerators for AI on-the-edge applications involves performing acrobatics

[a] https://orcid.org/0000-0001-7298-0480
[b] https://orcid.org/0000-0002-9758-3390

Wankar, G. and Vohra, S.
Enhanced AI On-the-Edge 3D Vision Accelerated Point Cloud Spatial Computing Solution.
DOI: 10.5220/0010616203620369
In *Proceedings of the 10th International Conference on Data Science, Technology and Applications (DATA 2021)*, pages 362-369
ISBN: 978-989-758-521-0

amidst the constraints of low-power achieving high performance. NVIDIA's Jetson is a promising platform for Embedded AI achieving a balance between the above objectives.

GPUs are known for their efficient capabilities of parallel processing. This is very useful as the system needs to filter thousands of data frames quickly to react according to the situation.

Voxels till Industry - 4.0 has been considered a bottleneck for engineering, as the hardware was not capable for processing them. Now with the arrival of Industry - 5.0, leveraging the state of art technologies such as GPU, TPU, VPU, FPGA accelerated Deep Learning for High Performance Computing clusters, voxels are accelerating AI on-the-edge solutions providing more precise renderings for visualization.

In this paper, we are leveraging Azure Kinect DK for obtaining the 3D vision point clouds for object detection along with the computing capabilities of NVIDIA's Jetson device for faster processing in real-time.

## 2 RELATED WORK

### 2.1 3D Vision Cameras

When we open our eyes on a well-known scene, we form an intuitive acquaintance of identifiable objects organized in the 3D spatial background. This is extended by our brain adapting to new surroundings captured though our eyes, sorting and combining them with prior knowledge to create the immersive experience.

Today with the emerging 3D vision cameras accelerating AI on-the-Edge solutions is pretty much feasible. Leveraging Azure Kinect DK with advanced computer vision models we are gaining deeper understanding of the physical environment and the objects present in the scene (David Coulter, 2019).

3D Vision based depth processing with object detection is quite fascinating and extensively studied for Embedded Vision applications.

Figure 2.1.1: Azure Kinect DK.

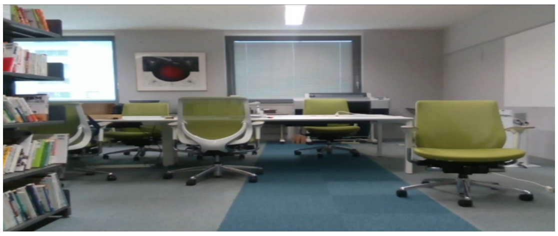


Figure 2.1.2: Scene captured by RGB Camera.

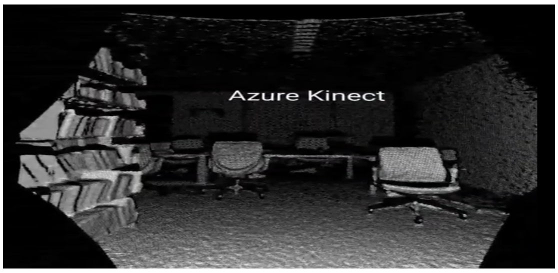


Figure 2.1.3: Scene visualized by Azure Kinect DK.

The figure 2.1.1 is representing Azure Kinect DK camera. The figures from 2.1.2 & 2.1.3 are representing the scene visualization from RGB Camera, Point Cloud data from Azure Kinect DK cameras respectively.

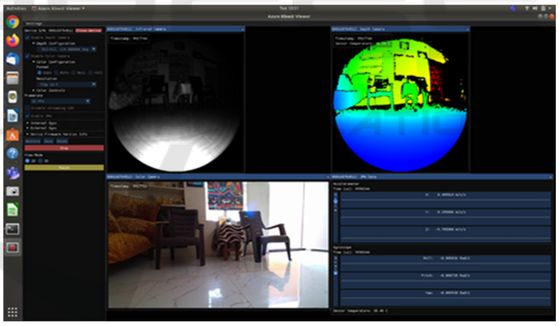


Figure 2.1.4: Azure Kinect 512x512 K4aviewer results visualizing the scenes from RGB, Infrared & Depth cameras corresponding to 720p resolution.

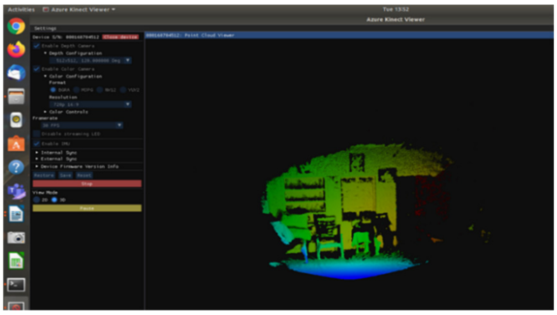


Figure 2.1.5: Azure Kinect 512x512 K4aviewer results visualizing the Point Cloud scene corresponding to 720p resolution.

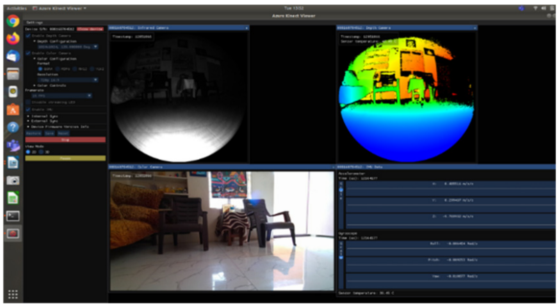


Figure 2.1.6: Azure Kinect 1024x1024 K4aviewer results visualizing the scenes from RGB, Infrared & Depth cameras corresponding to 720p resolution.

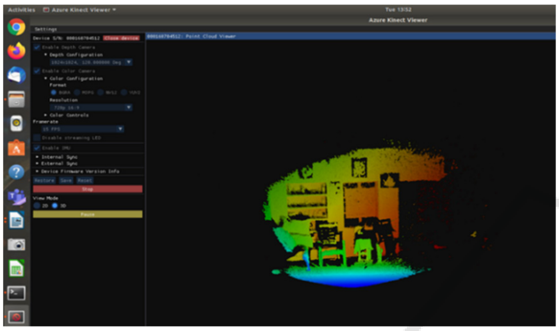


Figure 2.1.7: Azure Kinect 1024x1024 K4aviewer results visualizing the Point Cloud scene corresponding to 720p resolution.

The figures from 2.1.4 to 2.1.7 are representing the RGB, Infrared, Depth and Point Cloud results obtained from Azure Kinect DK correspondingly.

## 2.2 NVIDIA Jetson Tx2

NVIDIA Jetson Tx2 designed explicitly for High performance AI on-the-edge applications, housing both GPU and CPU on the same System-on-Chip (SoC) achieving faster processing supporting the emerging deep neural networks. (NVIDIA, 2021).

Jetson Tx2 leverages Pascal GPU architecture to increase performance upon Streaming Multiprocessor (SM). The Graphics Processing Cluster (GPC) includes multiple SM units and a Raster Engine for computing, rasterization, shading and texturing.

Jetson Tx2 runs more efficiently between 5 watts at max efficiency and 15 watts at max performance. (NVIDIA, 2021). It has better power efficiency making it ideal for AI on-the-edge applications such as autonomous vehicles, drones, virtual reality and mixed reality applications.

The figure 2.2.1 corresponds to NVIDIA Jetson Tx2 connected with the Azure Kinect DK leveraging the GPU parallel computing capabilities on the data obtained from the camera.

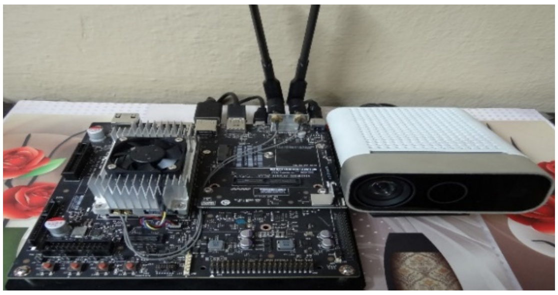


Figure 2.2.1: Edge AI Setup Leveraging Azure Kinect DK Accelerated with NVIDIA Jetson Tx2.

## 2.3 Emergence of AI-on-the-Edge

With the rise of Industry - 5.0 we are at the cusp of technology explosion driving innovation. Convergence of various technologies such as Big Data, Artificial Intelligence and Deep Learning incorporates AI on-the-edge applications.

With these rapidly evolving technologies such as Intel Movidius VPUs, NVIDIA GPUs, Intel Nervana Neural Network Processors (NNP), Google Tensor Processing Units (TPUs), Intel FPGA, Xilinx FPGA etc., the deep learning algorithms are emerging with advanced architectures leading to better performance, reducing latency and achieving higher throughput exploring unlimited possibilities leveraging Disruptive Technologies.

Leveraging the state of art technologies with deep learning architectures opens New Product Market providing immersive Business solutions. Exploring various approaches towards computer vision leads us to CNN, DNN, CuDNN, Mask R-CNN, Mesh R-CNN, LSTM, GoogLeNet, ResNet, SegNet and YOLO approaches which are trained on 2D data.

With the focus on Industry - 5.0 (Atwell, 2017) real-time applications, point-cloud based 3D deep learning is gaining pace rapidly as it involves an end-to-end deep learning network acquiring features directly from the point clouds.

3D object recognition, 3D object segmentation and point-wise semantic segmentation tasks are crucial component for applications which are tightly constrained by hardware resources and battery. Therefore, it is important to design efficient and fast 3D deep learning models for real-time applications on the edge such as virtual reality, mixed reality and autonomous driving.

# 3 ASSESSED ARCHITECTURES

In this paper, we are exploring few approaches for working on 3D data leveraging NVIDIA Jetson platform providing enhanced solutions.

We review both hardware and algorithmic approaches performed for running Deep Learning algorithms on NVIDIA Jetson and demonstrate the real-life applications. While the paper focuses on NVIDIA Jetson as an Edge device, these approaches also apply to the prevailing and future AI on-the-edge devices running AI algorithms on low-cost, low-power platforms. We are seeking to provide a glimpse of the recent progress towards the vision of "Smart Everything Revolution" (NVIDIA B. C., 2019).

## 3.1 Voxelization Approach

With the emergence of Industry - 5.0, the craze for volumetric data is growing rapidly enhancing cutting edge abilities for visual communication by providing more precise renderings.

Voxel is referred as a Pixel that has a volume. It can be distinguished as a smallest cube forming a 3D perspective of an image. Pixel, on the other hand is referred as a smallest square from a 2D plane which contains a single-color value and positional data associated with it.

The process of adding depth to a 2D plane using volumetric data (x, y, z, v) where v stands for volume is called as Voxelization. This approach transforms the Point Cloud into a voxel grid by rasterizing using 3D CNN techniques.

### 3.1.1 VoxelNet Approach

VoxelNet is a 3D detection network which splits the point cloud into 3D voxels and alters them into unified representation of a single voxel feature encoding (VFE) layer. (Yin Zhou, 2017).

Thus, the point cloud is Voxelized and then connected to a Region Proposal Network to instantaneously acquire the discriminatory feature representation from the point clouds and predict accurate bounding boxes generating detections.

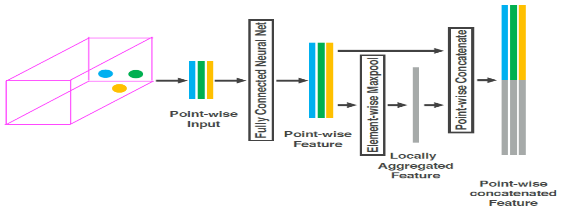

Point-wise Input
Fully Connected Neural Net
Point-wise Feature
Element-wise Maxpool
Locally Aggregated Feature
Point-wise Concatenate
Point-wise concatenated Feature

Figure 3.1.1: Voxel feature encoding (VFE) layer.

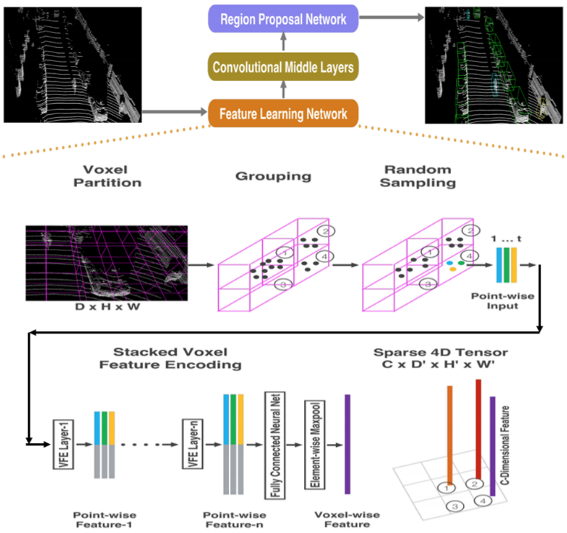


Figure 3.1.2: VoxelNet feature learning architecture.

The VFE layer converges point-wise and locally aggregated features. Stacking multiple VFE layers permits learning complex features for portraying local 3D shape data. (Yin Zhou, 2017). VoxelNet classifies the point cloud into 3D voxels, converts each voxel with stacked VFE layers and then DNN estimates local voxel features, transforming the point cloud into a high-level volumetric feature representation. Finally, RPN analyses this representation and yields the detection result.

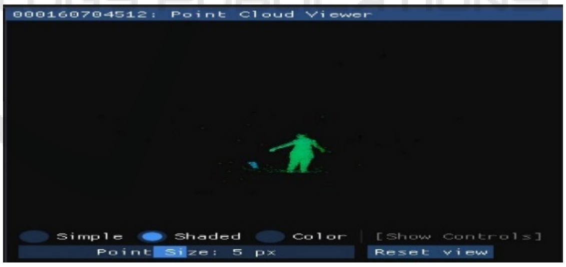


Figure 3.1.3: Azure Kinect Point Cloud Scene.

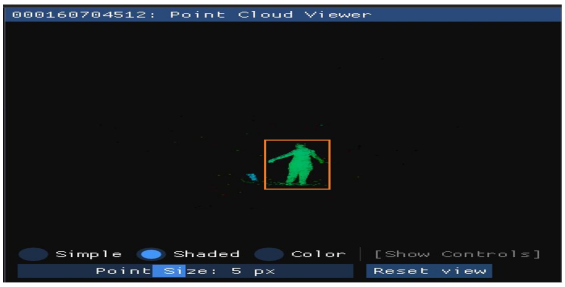
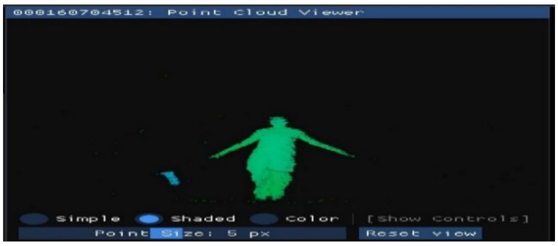
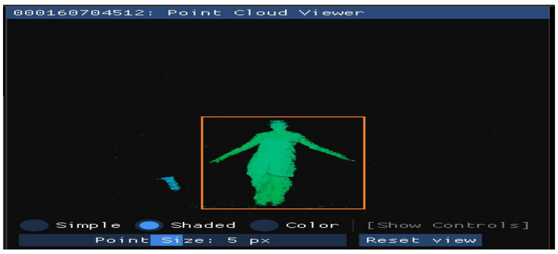


Figure 3.1.4: Human detection by VoxelNet approach from the Azure Kinect Point Cloud.



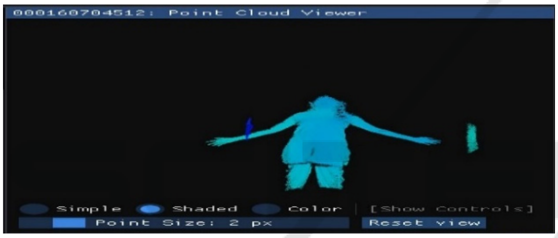
Point Size: 5 px

Figure 3.1.5: Azure Kinect Point Cloud Scene.



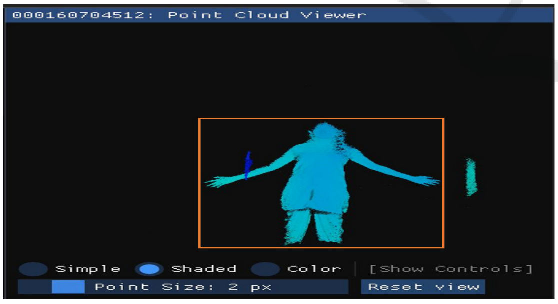
Point Size: 5 px

Figure 3.1.6: Human detection by VoxelNet approach from the Azure Kinect Point Cloud.



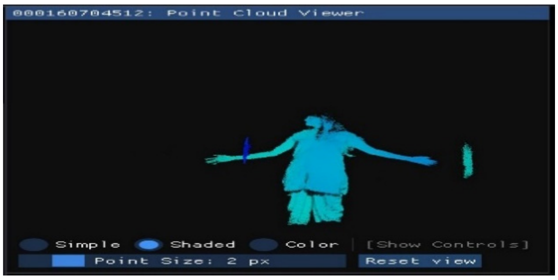
Point Size: 2 px

Figure 3.1.7: Azure Kinect Point Cloud Scene.



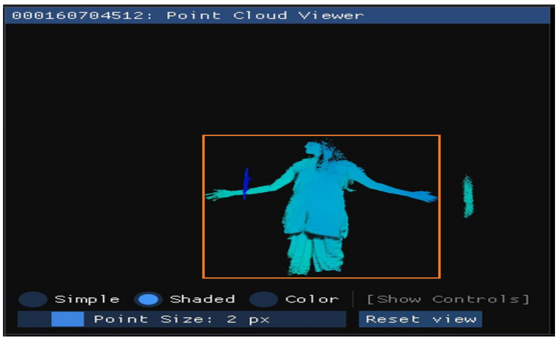
Point Size: 2 px

Figure 3.1.8: Human detection by VoxelNet approach from the Azure Kinect Point Cloud.



Point Size: 2 px

Figure 3.1.9: Azure Kinect Point Cloud Scene.



Point Size: 2 px

Figure 3.1.10: Human detection by VoxelNet approach from the Azure Kinect Point Cloud.

The Figures 3.1.3 to 3.1.10 are representing the VoxelNet approach operating on the raw point cloud data obtained from Azure Kinect Dk detecting humans in the point cloud scene.

The VoxelNet approach predicts and detects the human from the Point Cloud obtained from Azure Kinect DK with an accuracy of 90%.

Thus, benefiting from the sparse point cloud and parallel computing on the voxel grid.

## 3.2 PointPillars & PointNet Approach

PointPillars approach converts the raw point cloud data into a vertical pillar like structure having the same intensity values as the point cloud. PointNet performs semantic segmentation and classification over this data. It uses a novel feature encoder for performing feature extraction to precisely predict objects from a 3D plane. (Lang, 2019).

While the conventional approach of Voxelization requires volumetric data like Voxels causing deprecation in quality of details while sampling the data, the PointPillars approach overcomes this drawback by transforming the raw Point Cloud data into pillar like structures as an input to the simplified PointNet which can be then leveraged by the Deep Learning Network.

The PointPillars network (Lang, 2019) has 3 stages. In the 1$^{st}$ stage the novel encoder transforms the point cloud to a sparse pseudo image by transforming point cloud into PointPillars and then using a sorted edition of PointNet network learns the representation of point clouds as PointPillars.

PointNet identifies each object from vivid classes making a remarkable progress in scene semantic segmentation classifying every voxel belonging to the particular class of objects. Thus, maintaining the variance after performing various transformation like translating or even rotation of objects from the 3D plane.

In the 2nd stage the 2D Convolutional backbone processes this pseudo image in a high-level representation. Finally in the 3rd stage the detection head detects objects and creates bounding boxes around them.

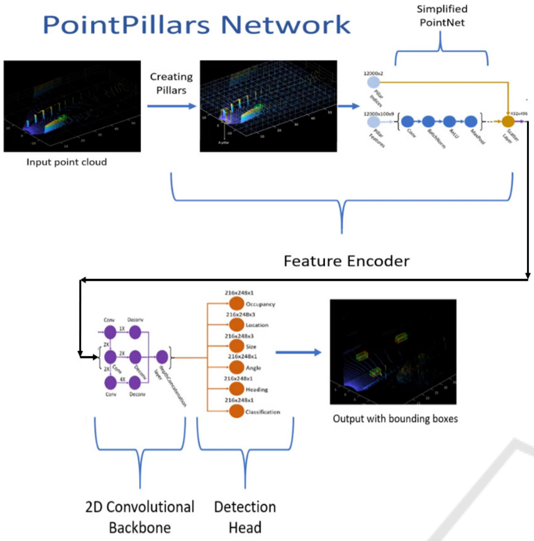


Figure 3.2.1: Network architecture for PointPillars approach.

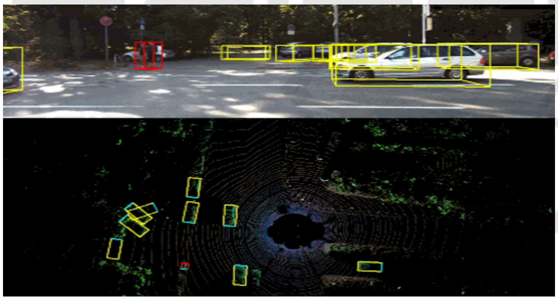


Figure 3.2.2: Visual perception using PointPillars.

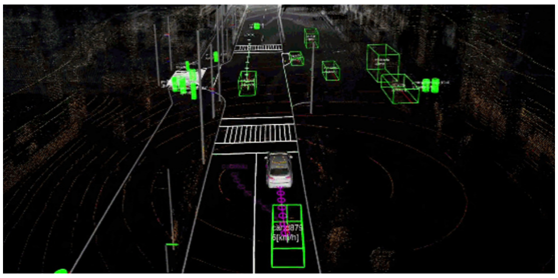


Figure 3.2.3: Numerous objects detected by PointPillars.

The Figures 3.2.2 and 3.2.3 are representing the PointPillars approach discovering numerous objects in the Point Cloud visualization.

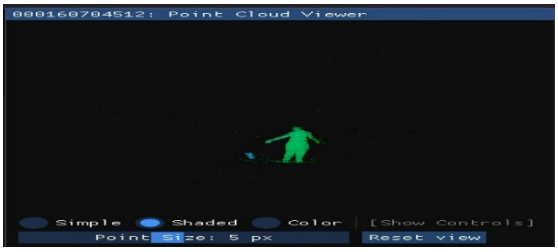


Figure 3.2.4: Azure Kinect Point Cloud Scene.

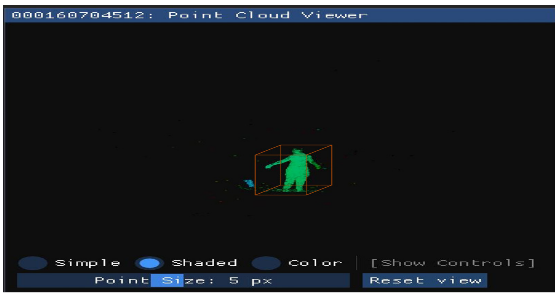


Figure 3.2.5: Human detection by PointPillars and PointNet approach from the Point Cloud.

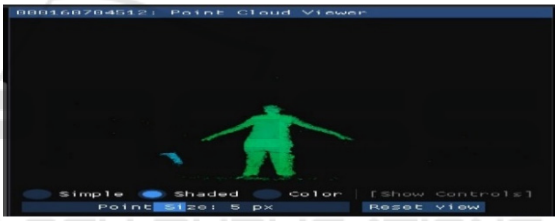


Figure 3.2.6: Azure Kinect Point Cloud Scene.

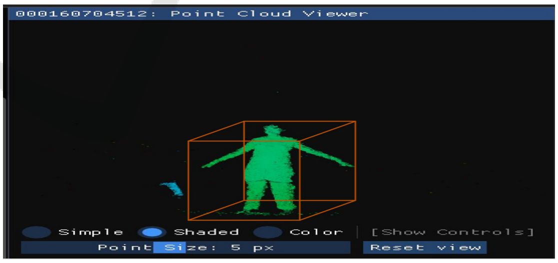


Figure 3.2.7: Human detection by PointPillars and PointNet approach from the Point Cloud.

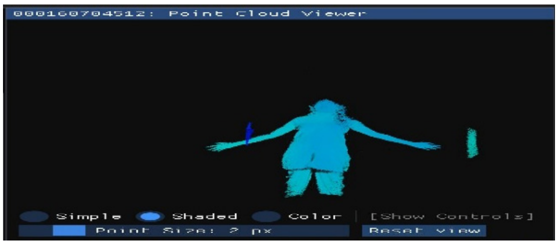


Figure 3.2.8: Azure Kinect Point Cloud Scene.

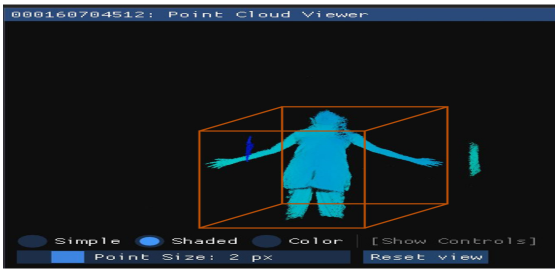

000160704512: Point Cloud Viewer
○ Simple ● Shaded ○ Color [Show Controls]
Point Size: 2 px
Reset view

Figure 3.2.9: Human detection by PointPillars and PointNet approach from the Point Cloud.

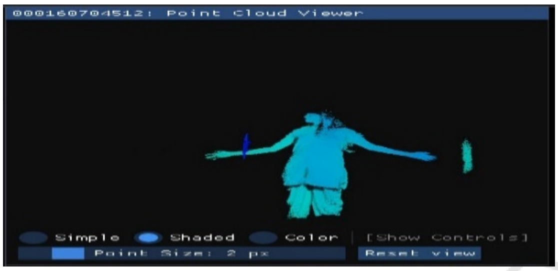

000160704512: Point Cloud Viewer
○ Simple ● Shaded ○ Color [Show Controls]
Point Size: 2 px
Reset view

Figure 3.2.10: Azure Kinect Point Cloud Scene.

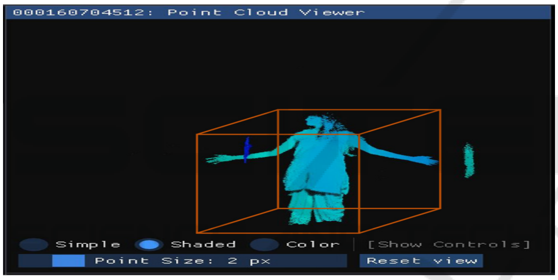

000160704512: Point Cloud Viewer
○ Simple ● Shaded ○ Color [Show Controls]
Point Size: 2 px
Reset view

Figure 3.2.11: Human detection by PointPillars and PointNet approach from the Point Cloud.

The Figures 3.2.4 to 3.2.11 are representing the PointPillars approach operating on the raw point cloud data obtained from Azure Kinect Dk detecting humans in the point cloud scene.

The PointPillars approach predicts and detects the human from the Point Cloud obtained from Azure Kinect DK with an accuracy of 95%.

Thus, benefiting from the sparse point cloud and parallel computing on the PointPillars leveraging NVIDIA Jetson Tx2.

## 4 FUTURE PERSPECTIVE

With the emergence of Industry - 5.0, the 3D Vision Product Market is growing tremendously. Azure Kinect DK working alongside with Microsoft's Azure leveraging Cognitive Services enhances the feature extraction to be more accurate.

Leveraging vision and speech analytics we can effectively enhance the immersive detection and decision-making experience.

Thus, providing innovative Business-to-Business (B2B) and Business-to-consumer (B2C) solutions across various industry verticals such as Robotics, Digital Twins, Self-Driving Vehicles, Drones, Photogrammetry, Healthcare, Retail and Manufacturing industries.

Converging AI and Mixed Reality opens a New Generation of services bringing accurate immersive real-world information allowing decision-making based on Digital Reality driving Digital Transformation.

## 5 CONCLUSIONS

We would like to conclude that, with the rise of Industry - 5.0 we are moving towards the convergence of various fields concentrating on alliance between humans and smart systems. We are exploring transformative experiences accelerated with NVIDIA Jetson Tx2 edge computing leveraging the state of art Deep Learning algorithms such as Voxelization, PointNet and PointPillars Network approaches for enhancing the feature extraction and object detection form the Point Cloud data with cutting-edge spatial computing capabilities.

Thus, leading towards the Next Generation of Human Machine Interaction breathing New Life into the future of Immersive Business Transformation.

## ACKNOWLEDGEMENT

It gives us immense pleasure to extend our gratitude towards Neal Analytics Services Private Limited for encouraging, motivating and providing us the necessary support along with the computational resources required throughout the entire journey of exploring this project.

## REFERENCES

Alex H. Lang, S. V. (2019). PointPillars: Fast Encoders for Object Detection from Point Clouds. https://arxiv.org/abs/1812.05784.

Atwell, C. (2017). Yes, Industry 5.0 is Already on the Horizon. https://www.machinedesign.com/automation-iiot/article/21835933/yes-industry-50-is-ready-on-the-horizon. Last accessed on Sept 12, 2017.

B. Li, T. Z. (2016). Vehicle detection from 3d lidar using fully convolutional network. Robotics: Science and Systems.

Cabanes, Q. S. (2017). Object detection and recognition in smart vehicle applications: Point cloud-based approach. Ninth International Conference on Ubiquitous and Future Networks, (pp. 287-289).

Chen, X. M. (2017). Multi-View 3D Object Detection Network for Autonomous Driving. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6526-6534). doi: 10.1109/CVPR.2017.691.

David Coulter, P. M. (2019). Azure Kinect DK documentation. https://docs.microsoft.com/en-us/azure/kinect-dk/. Last accessed on Jun 26, 2019.

Gartner. (2021). 4 Impactful Technologies from the Gartner Emerging Technologies and Trends Impact Radar for 2021. https://www.gartner.com/smarterwithgartner/4-impactful-technologies-from-the-gartner-emerging-technologies-and-trends-impact-radar-for-2021/. Last accessed on Jan 18, 2021.

Geiger, A. L. (2013). Vision meets Robotics: The KITTI Dataset. International Journal of Robotics Research (IJRR).

Intel. (2019). Beginner's guide to depth. https://www.intelrealsense.com/beginners-guide-to-depth/. Last accessed on Jul 15, 2019.

Lang, A. H. (2019). PointPillars: Fast Encoders for Object Detection from Point Clouds. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 12689-12697).

Microsoft. (2021). Azure Kinect DK: Developer kit with advanced AI sensors for building computer vision and speech models. https://azure.microsoft.com/en-in/services/kinect-dk/.

NVIDIA. (2021). DATA SHEET NVIDIA Jetson TX2 Series System-on-Module. https://developer.download.nvidia.com/assets/embedded/secure/jetson/TX2/docs/Jetson-TX2-Series-Module-Datasheet-v1.8.pdf?nstEXJXxf7pNVJaKPAZFKcUQJis2UqnnKPC9kw04moG3zJ3rKBd9ECfWOzQ9i3Ispxa4ZLf4eTKJ95-Gryf3wYjgTgpDpREROe_ejvwPx7fDboumngyRgRvA6boB37hIe_BCWy. Last accessed on Apr, 2021.

NVIDIA, B. C. (2019). 5G Meets AI: NVIDIA CEO Details 'Smart Everything Revolution,' EGX for Edge AI, Partnerships with Leading Companies. https://blogs.nvidia.com/blog/2019/10/21/5g-meets-ai-nvidia-egx-edge-ai/. Last accessed on Oct 21, 2019.

Qi, C. R. (2018). Frustum PointNets for 3D Object Detection from RGB-D Data. IEEE/CVF Conference on Computer Vision and Pattern Recognition, (pp. 918-927).

Shaoshuai Shi, C. G. (2019). PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. https://arxiv.org/abs/1912.13192.

Stanisz, J., Lis, K., Kryjak, T., & Gorgon, M. (2020). Optimization of the PointPillars network for 3D object detection in point clouds. https://doi.org/10.36227/techrxiv.12593555.v1. TechRxiv.

Tang, H. C. (2017). Multi-cue pedestrian detection from 3D point cloud data. IEEE International Conference on Multimedia and Expo (ICME), (pp. 1279-1284.).

X. Chen, K. K. (2016). Monocular 3d object detection for autonomous driving. IEEE CVPR.

Yin Zhou, O. T. (2017). VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. https://arxiv.org/abs/1711.06396.