Point Cloud Comparisons from Monocular Images by MiDaS Convolutional Neural Networks and Dense Prediction Transformers

by

Seth Howells

Dr. Osama Abuomar, Yasmeen Labead

Masters project submitted in partial fulfillment of the

requirements for the Master of Science in Data Science degree

in the College of Arts and Sciences of

Lewis University

***Abstract*—This research paper introduces a comparative case study on inferring point clouds from calibrated single-view monocular images with two deep learning architectures, convolutional neural networks and dense prediction transformers. Three monocular depth estimation models, MiDaS Small-CNN, MiDaS Hybrid-DPT and MiDaS Large-DPT, were used to produce depth maps of input images. By applying a 4x4 disparity-to-depth mapping matrix, Q, from the calibrated intrinsic camera matrix, K, 2D input images were transformed to 3D point clouds. The research assesses computational efficiency of using each MiDaS model to generate depth map and corresponding polygon file at different image resolutions, monitor the difference in outlier points detected, visually compare 3D geometries of point clouds, and quantify y-axis measurements from a defined flat plane. The research also features original perspectives from objects' point-of-view in the opposite direction of the camera from the 2D input image. The production-rate analysis of the three models to produce depth map and polygon file results in MiDaS-Small CNN performing 9% faster than MiDaS Hybrid-DPT with 5% less memory, and 44% faster than MiDaS Large-DPT with 38% less memory. Despite the lack in computational efficiency, MiDaS Large-DPT ranked superior in accurate representations of 3D geometries that mimic closer to real-world coordinates while the CNN fell flat in Y-axis measurements and low z-depth range. Taking both computational efficiency and accuracy into consideration, MiDaS Hybrid-DPT stands out overall for use as a close-range monocular image 2D-to-3D model.**

Index Terms—convolutional neural network, deep learning, dense prediction transformer, depth estimation, point clouds

I. INTRODUCTION

In recent years, 3D acquisition technologies have helped further the understanding of surrounding environments for machines, which serves numerous applications across industries like robotics, autonomous driving, remote sensing, medical treatment, and social media platforms [1]. Rapid development in these types of technologies include 3D scanners, LiDARs, and RGB-D cameras (e.g. Kinect, RealSense, and Apple depth cameras) [2, 3]. While these new acquisitions can assist in producing 3D data, the cost of acquiring these technologies far exceeds the traditional monocular camera that has been around for decades. Furthermore, the majority of existing data related to vision, (i.e. images and videos) were captured with monocular cameras, which does not have the 3D-sensing qualities found in newer technology. Due to the large amount of existing single-view monocular data, it is important to consider approaches to utilize the data in ways that mimic newer technology.

A point cloud consists of a point set, which are a collection of points with 3D unstructured vectors [3]. The points in a point cloud (3D data) are called volumetric picture elements (voxels) which are analogous to a picture-element (pixel) in an image (2D data). Each point contains a vector with 3D spatial coordinates along with extra feature channels such as color, normals, or intensity values and are typically stored in LAS, PLY, XYZ or other 3D data text formats for import and export [4]. The research in this paper focuses on transforming single images into point clouds by leveraging depth maps from convolutional neural networks and vision transformers.

Moving from pixel to voxel, or otherwise 2D to 3D, from a single calibrated image has been a challenging task in computer vision until recent development in dense prediction transformers (DPT). Much previous work in this area has been centered around stereoscopic images, multi-angled images, or with newer technology [4, 5]. Sparse work has been done in generating 3D point clouds of an object from a single-view monocular image, which provides motivation to continue to improve on methods with fuller-scene point clouds [6].

The purpose of this research is to obtain single-view monocular images from datasets that provide the intrinsic camera matrix, K, to form a 4x4 disparity-to-depth mapping matrix, Q, in order to reproject the 2D image's pixels into 3D voxels by simultaneously transforming the input image and depth map into a polygon file. Monocular depth estimations will be inferred from a MiDaS depth-estimation model consisting of convolutional neural network (CNN), medium-size dense prediction transformer (DPT), and large dense prediction transformer (DPT). Using these deep learning approaches toward depth estimation, this research paper aims to assess two main areas in generating point clouds from monocular images: 1) quantify the computational efficiency, and 2) visually compare the accuracy of point clouds produced.

II. RELATED WORKS

There have been many attempts to transform 2D data into point clouds that range in the level of technological equipment. More traditional approaches use rectified stereo images which consist of two cameras, left and right, to mimic human eyesight in order to gauge depth from the point of disparity [7]. In the last decade, newer technological advancements in Light Detection and Ranging (LIDAR) "became a leading technology of detailed and reliable 3D environment perception" [8]. While sparse work has been done in single-view monocular images, the development of deep learning methods, namely with convolutional neural networks, has been successful for objects [6]. As opposed to existing methods, this research adopts vision transformers from MiDaS depth models [10] and compares the results in inferring point clouds from monocular images with point clouds inferred from convolutional neural networks.

III. METHODOLOGY

The method to produce point clouds from a single-view monocular image begins with applying a depth map of the image. The original image contains two-dimensional X, Y coordinates with either a three-channel color scale (e.g. RGB) or single-channel (e.g. grayscale). Three MiDaS monocular depth maps models [10] were used to produce the original image's corresponding depth estimation. The depth maps are representations of only the Z-depth of the objects in the given 2D scene. Although depth maps do not provide full three-dimensional information needed for point cloud creation, it is an important prerequisite to obtain the third dimension, Z-depth. Comparisons of depth maps from convolutional neural network and vision transformers are later described in greater detail in Section 3.2 and visualized in Section 4.1. A visual process of the sequence in methodology is shown in a diagram, Figure 1.

In addition to the acquisition of input image and corresponding depth map, a 4x4 disparity-to-depth mapping matrix Q is a necessary requirement to produce accurate point clouds because it is responsible for mapping 2D coordinates (pixels) to 3D locations (voxels) [11]. During the camera calibration phase, which occurs prior to capturing the input image, the camera's intrinsic parameter matrix is formed. The disparity-to-depth mapping matrix Q consists of information from the intrinsic parameters: principal point X, principal point Y, and focal length. Section 3.3 goes further into detail how the camera intrinsic parameters shape the point cloud. The final step before creating a point cloud in the form of a PLY file is to apply the color-channel onto the voxels.

After running this same operation for each of the three MiDaS models, several analyses are used to assess the difference between convolutional neural network implementations and transformers implementation. The analysis includes visual point cloud comparison, outlier analysis, point-based analysis, and production rate analysis. More information on results is found in Section 5.

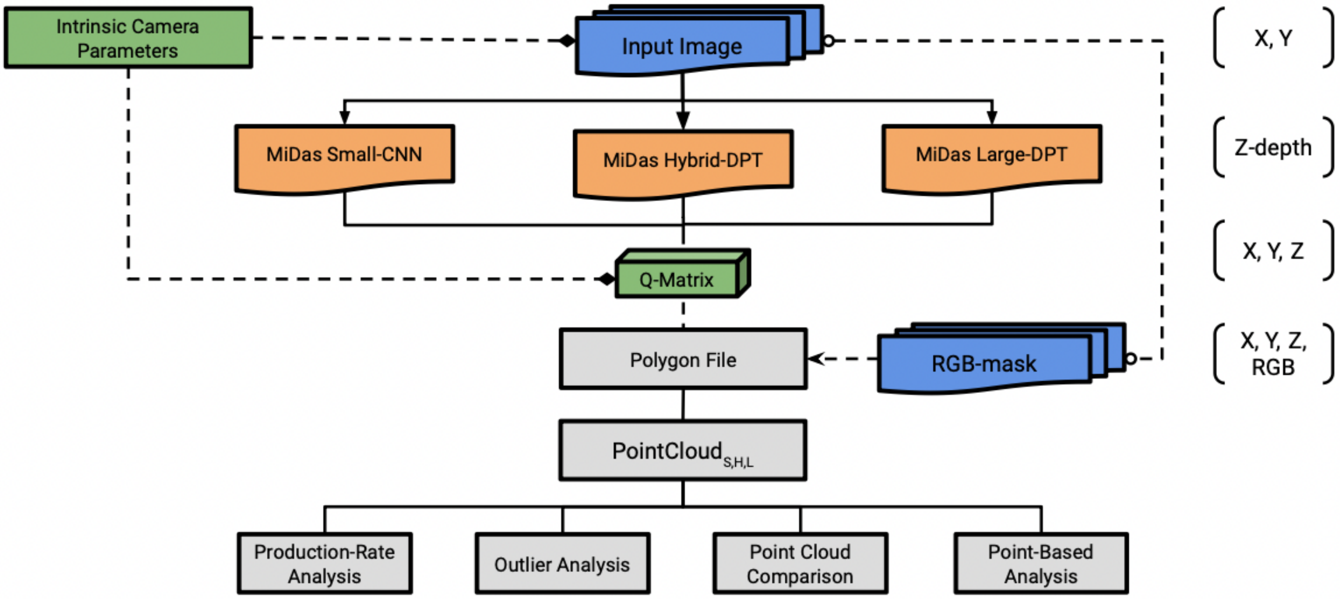


Figure 1: Diagram of methodology

3.1 Datasets

Two real-world datasets were used for the analysis of production rates, outlier detection levels, visual comparisons, and Y-axis differences. The datasets were chosen for implementation and analysis because each provided intrinsic camera parameters for the input image. Dataset 1, WILDTRACK Multi-Camera Person Dataset from The École polytechnique fédérale de Lausanne (EPFL) [12], focuses on human targets in an outdoor setting. Dataset 2, The Event-Camera Dataset and Simulator: Event-based Data for Pose Estimation, Visual Odometry, and SLAM from University of Zurich [13], analyzes outdoor architecture as well as indoor scenes.

3.2 Convolutional Neural Networks vs Dense Prediction Transformers

Previous work in monocular depth estimation has been almost entirely centered around pixelwise dense prediction, a task in which predicts a label for each pixel, which utilizes a convolutional neural network (CNN) [14]. While CNNs are known for their computational efficiency, Hinton notes that the architecture of CNNs produce “poor translational invariance and pose orientation” as well as “3D orientation relative to the camera or viewer”, all of which playing an integral role in point clouds [15]. In a different approach, dense prediction vision transformers (i.e. MiDaS monocular depth estimation models), used in this research, was proposed in 2021 and yields higher accuracy in understanding 3D geometries from 2D data, yet succumb to less computation efficiency.

The reason that CNNs perform poorly in areas important to 3D representations is largely due to the loss of information in the encoder and decoder architecture, Figure 2 (left). In image classification research, convolutional backbones progressively downsample the whole image as part of learning low-level features to high-level features. However, in this process of downsampling and upsampling, granularity and resolution are lost deep in the model and struggle to recover in the decoder.

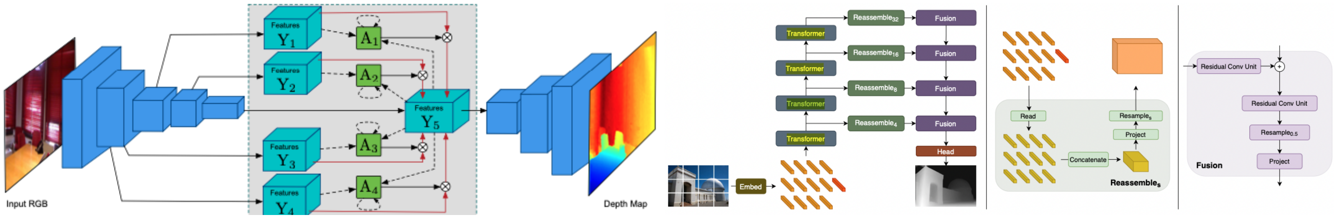


Figure 2: Convolutional neural network architecture for depth estimation [9] (left). Vision transformer architecture for depth estimation [10] (right).

Dense prediction transformers (DPT) work around the issue with CNNs by adopting attention-based models, which have been successful in natural language processing [10]. Here, vision transformers work as a bag-of-words representation where the input image is divided into image patches (or, "tokens") and reassembled for final prediction, Figure 2 (right). The overall DPT structure still maintains an encoder-decoder architecture, but decreases the loss of information during downsampling and upsampling because it is working on image patches as opposed to the image as a whole. From natural language processing to computer vision, this operation is analogous to words in a sentence/bag (NLP) and image patches to the overall image (CV).

Leveraging dense prediction transformers in place of CNNs helps solve the issues of 3D pose orientation and in return produce more accurate point clouds. The role depth maps play in point cloud creation is to segment the image into closest and furthest objects in relation to one another. Once depth maps are produced, the original image's RGB color is applied onto the depth map as a mask. In return, the newly formed image contains the original properties from the input image while obtaining the Z-dimension from the depth map.

3.3 Disparity-to-Depth Mapping Matrix *Q*

Disparity-to-depth mapping matrix typically involves stereovision where the distance between left and right viewpoints are calculated to triangulate the point of disparity. Because disparity is "proportional to the inverse of the corresponding scene point," depth can be gained. However, in the implementation of single-view monocular images, the depth map from MiDaS depth estimation models take the role of disparity maps and utilize the calibrated camera's intrinsic matrix. The diagram in Figure 3 shows the interplay between the camera's coordinate system *C,X,Y* to the image plane *x,y*, i.e. input image, and *Z*-depth for each *u,v* point on the image plane.

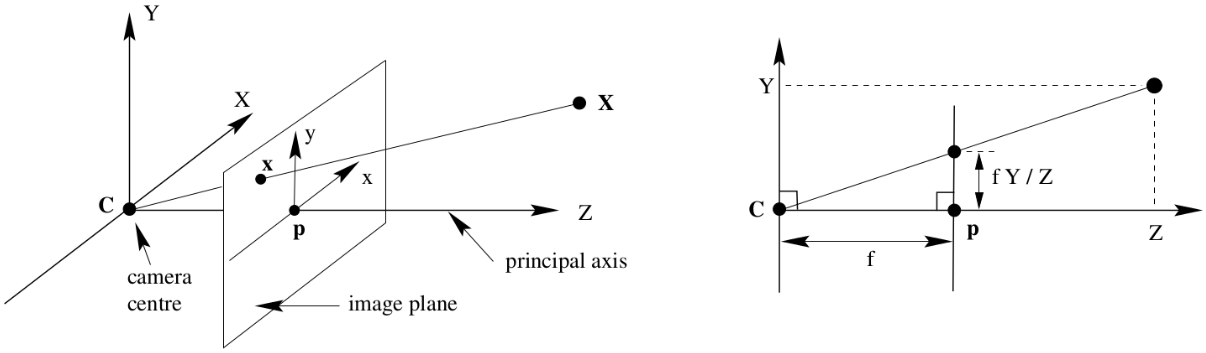


Figure 3. Diagram of camera coordinate system relationship image plane points [16]

The Pinhole Model in Equation (1) translates 2D points on the image plane (*u,v*) to the camera's coordinate system (*X,Y,C*) through the intrinsic matrix in Equation (2). The intrinsic matrix refers to a 3x3 matrix converting camera-to-image transformation with image-to-pixel transformation. The 3x3 intrinsic matrix, K, contains pixel focal length (f_x and f_y) and principal point (c_x and c_y), where K is expressed in pixel units.

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{1}$$

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The 4x4 depth mapping Q matrix, Equation (3), transforms the K matrix into a projection matrix to map out pixel-to-voxel spatial coordinates.

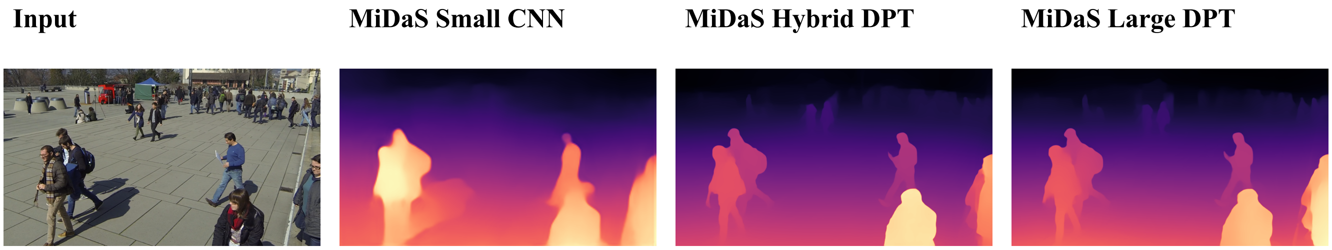
$$Q = \begin{bmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -\frac{1}{T_x} & \frac{c_x - c_x'}{T_x} \end{bmatrix} \tag{3}$$

IV. IMPLEMENTATION

The implementation of applying CNN or vision transformer depth maps on a single image from a monocular camera and transforming the pixels to voxels with information from the camera intrinsic parameters shows that deep learning techniques can build 3D models from 2D images. Due to loss of information in the decoding stage, convolutional neural networks struggle to regain the 3D geometric relationships from pixel-to-voxel transformation. Vision transformers, on the other hand, overcome such challenges by reassembling and fusing the image, which allows less information loss during the decoding stage.

4.1 Comparing Depth Maps

Comparisons of MiDaS models in an outdoor setting with moving pedestrians as objects are displayed in Figure 4. Compared to the fully-convolutional network, Dense Prediction Transformers (Hybrid and Large) show greater definition of objects (e.g. outlines of pedestrians), gauges depth from further distances (e.g. pedestrians in background, fourth row), and ignores shadows provided from the pedestrians (e.g. first column). Large-DPT outperforms foreground definitions greater than Hybrid (e.g. foot of nearest person, third row).



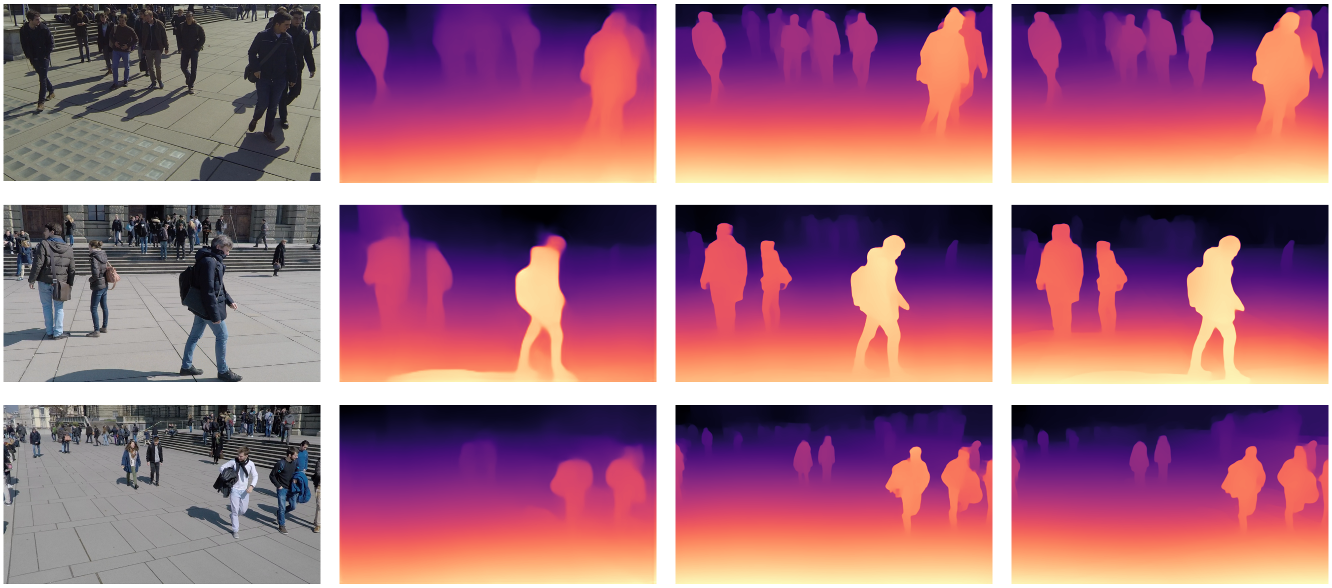
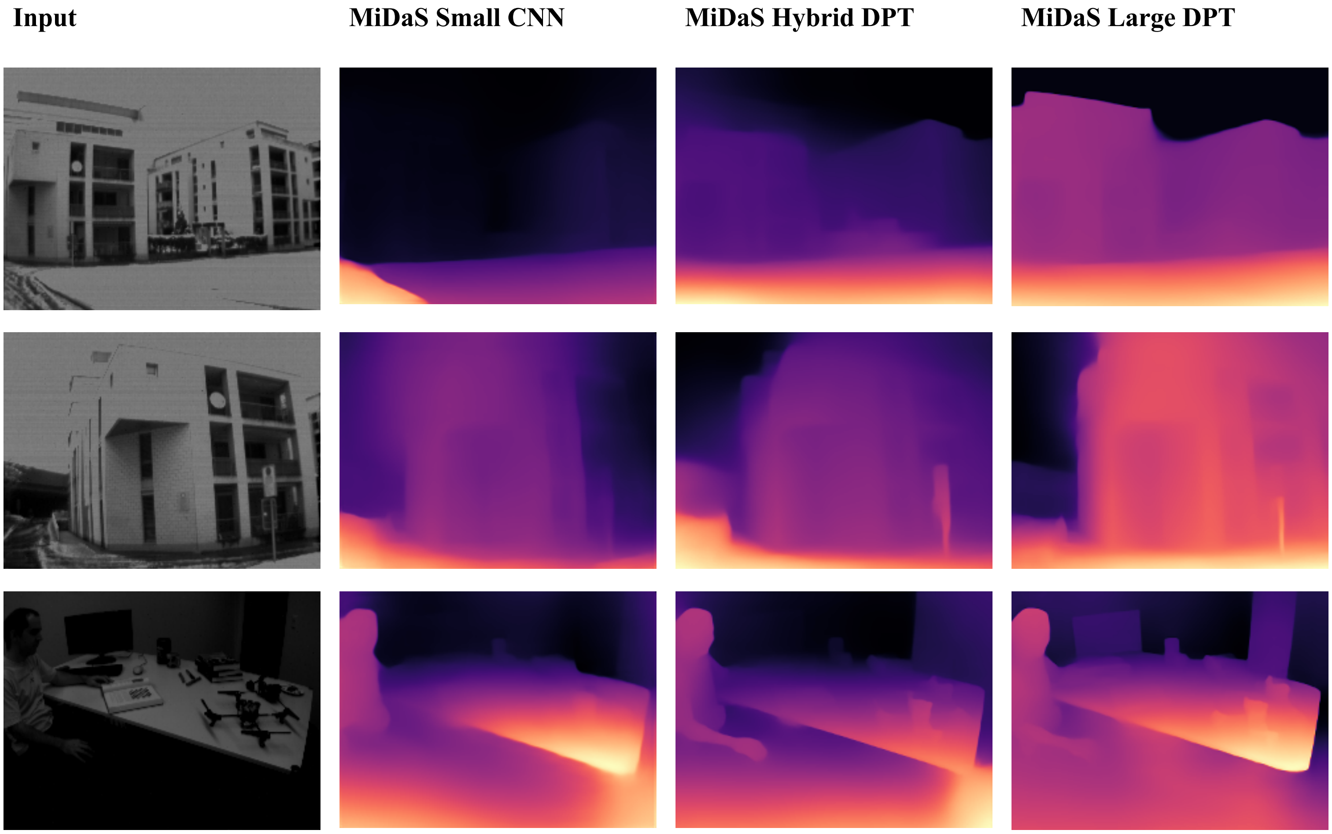


Figure 4: Sample results of monocular depth estimation from (3) MiDaS models of WILDTRACK Seven-Camera HD Dataset

Figure 5 compares the models with outdoor building architecture structures as well as indoor office scenes. Stricter outlines of objects are seen in Dense Prediction Transformers than with convolutional neural networks in all 5 sampled images. Deeper shading of concave structures within a building are progressively better with the Large-DPT model (e.g. balcony, second row). Finer-grained details are also progressively better with transformers as seen with the furthest objects on the desk (e.g. cup and monitor, third row).



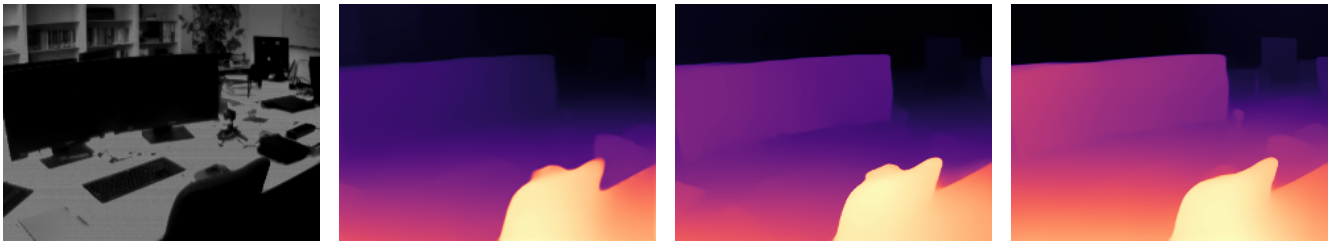


Figure 5: Sample results of monocular depth estimation from (3) MiDaS models of UC Davis 240C dataset.

Samples from both datasets, Figure 4 and Figure 5, clearly show the progressive quality in distinguishing objects' depth estimation from the camera. Qualities of finer-grained detail, global coherence, shading of concave structures, ignorance of non-object shapes (i.e. shadows), and foreground-to-background distinction play an integral role in point clouds. Each X,Y pixel on the depth map will later be reprojected into a 3D space where these qualities determine how the projection of X,Y will be viewed in X,Y,Z.

V. RESULTS

This section will contain graphs, tables, and figures of the results from each MiDaS model inferring point clouds. Section 6 will contain discussion of the results.

5.1 Production-Rate Analysis

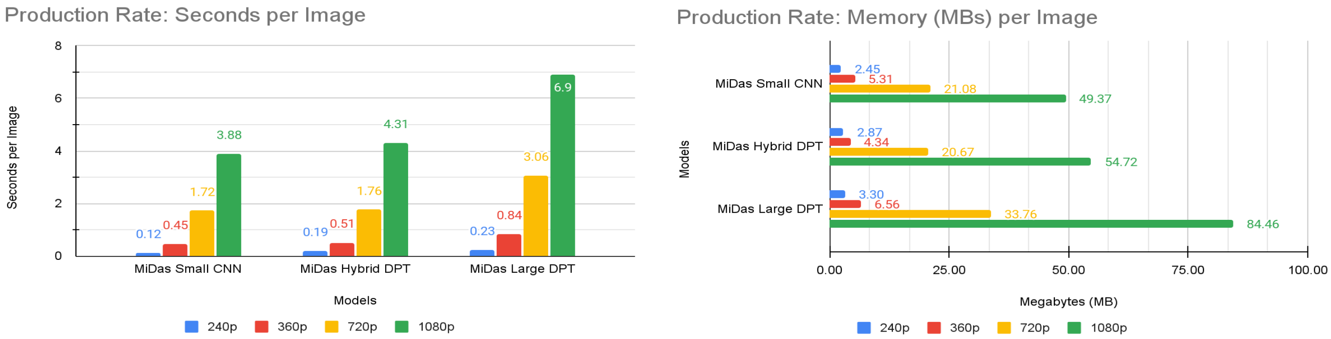


Figure 6: Sampled at (4) different resolution: 240p, 360p, 720p, 1080p. Time to produce depth map (PNG) and point cloud (PLY) file per image (left). Amount of memory required for depth map and point cloud files (right).

5.2 Outlier-Based Analysis

Outliers were determined by a statistical measurement of nearby neighbors' average distance. Points that exceeded the maximum distance, defined by average distance (average of 50 neighbors) plus standard deviation (threshold) of 2.0, were computed as an outlier.

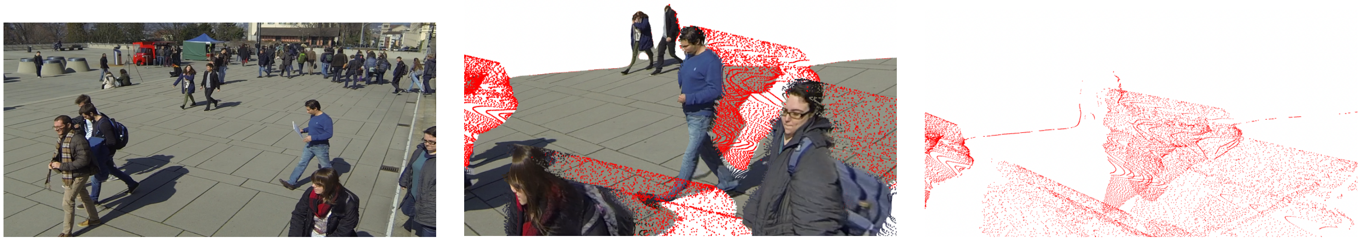


Figure 7: Ground truth (left). Outlier detection on image (middle). Outliers only (right)

	Camera 1	Camera 2	Camera 5	Camera 6
MiDaS Small CNN	0.03486	0.03114	0.04243	0.05645
MiDaS Hybrid DPT	0.01806	0.02478	0.01384	0.02859
MiDaS Large DPT	0.01744	0.02542	0.01560	0.02810

Table 1: Heatmap chart of outlier detection in percentages- Dataset 1

	urban_1	urban_2	office_1	office_2
MiDaS Small CNN	0.03428	0.04340	0.05182	0.03085
MiDaS Hybrid DPT	0.04642	0.04672	0.04953	0.02469
MiDaS Large DPT	0.01047	0.03371	0.03517	0.01796

Table 2: Heatmap chart of outlier detection in percentages - Dataset 2

5.3 Point Cloud Comparison

Split into two parts, the first portion, Figure 8-9, of point cloud comparison concerns the comparison of 3D spatial accuracy with specificity geared toward Y-axis shape and Z-axis range. The second portion, Figure 10, focuses on experimental visualization by assessing a new perspective not directly provided by the ground truth image by relocating the virtual point cloud camera to two pedestrians, A and B, with (4) pedestrians as the target.

5.3.1 Views From Different Angles

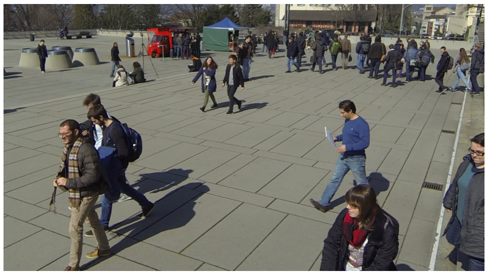
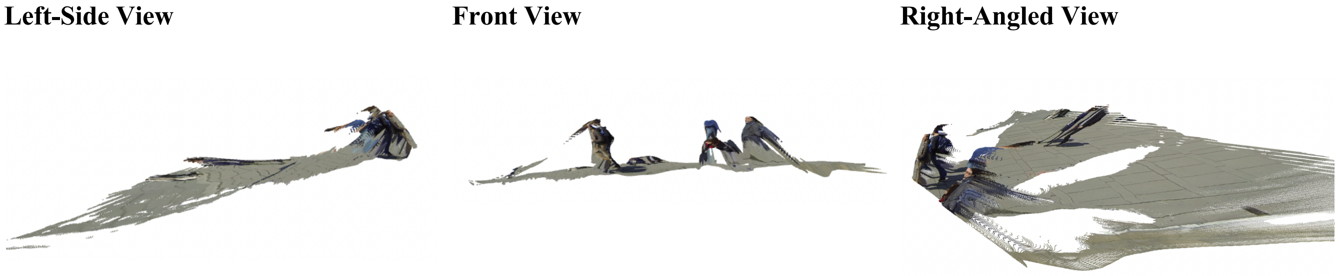


Figure 8: Ground truth image



Left-Side View

Front View

Right-Angled View

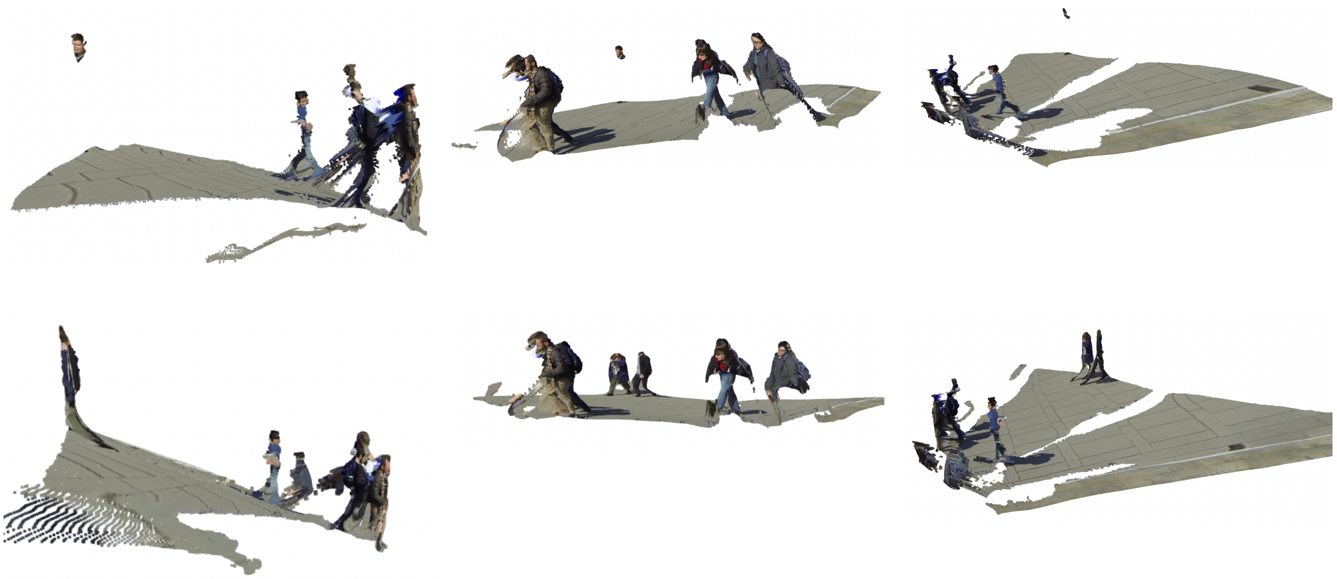
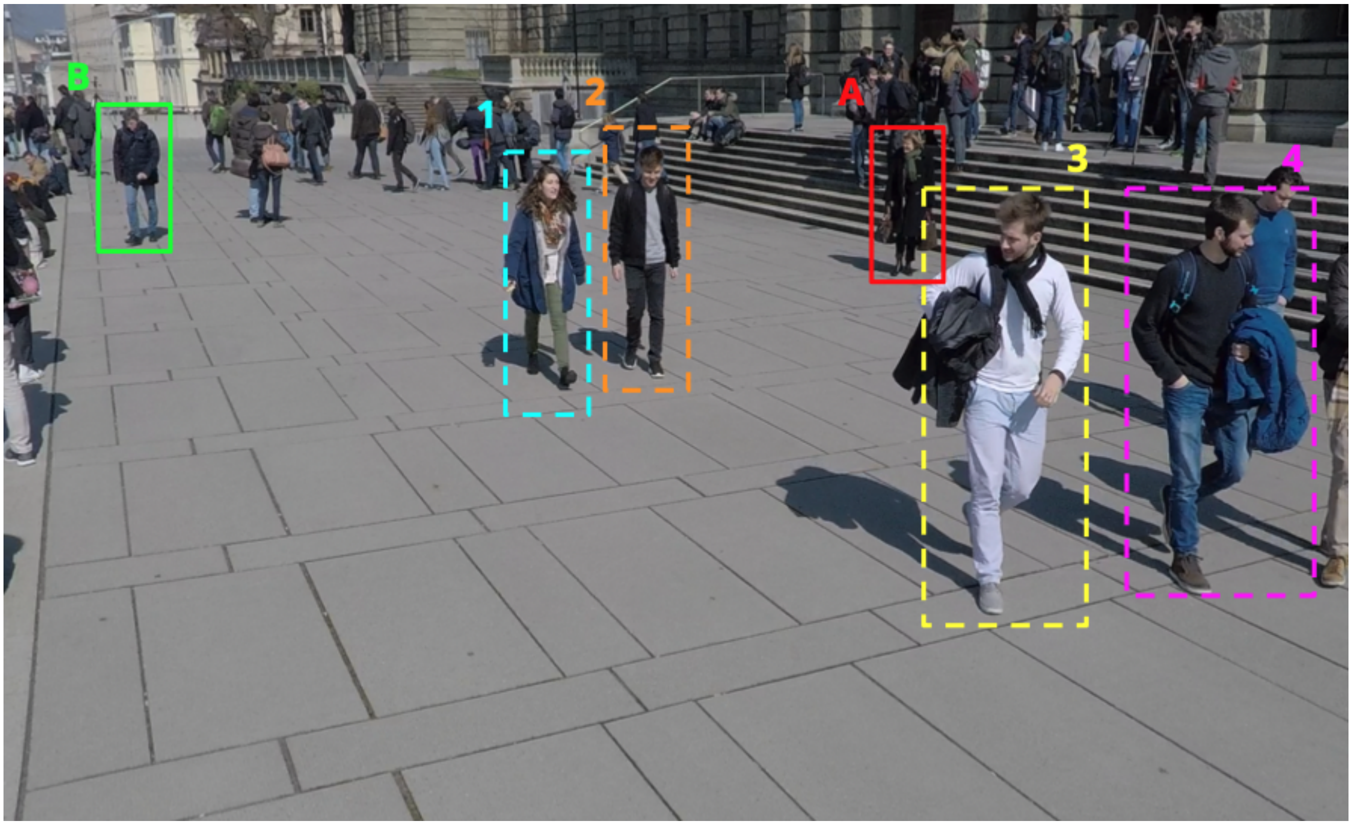


Figure 9: Comparing point clouds inferred from (3) different MiDaS monocular depth estimation models: MiDaS Small-CNN (top), MiDaS Hybrid-DPT (middle), MiDaS Large-DPT (bottom). First column displays the point cloud from a left-side view in order to determine the Y-axis perpendicularity shape of pedestrians on the flat plane (ground). Second column displays a front-view of the point cloud to display another Y-axis perpendicularity assessment and shows X-axis skewness. Third column is an angled right-side viewpoint to display Z-axis, depth, as well as Y-axis. From top-row down to bottom-row presents progressive accuracy in depicting 3D reality from the 2D image.

5.3.2 Experimental Results: Original Perspectives From 2D Image



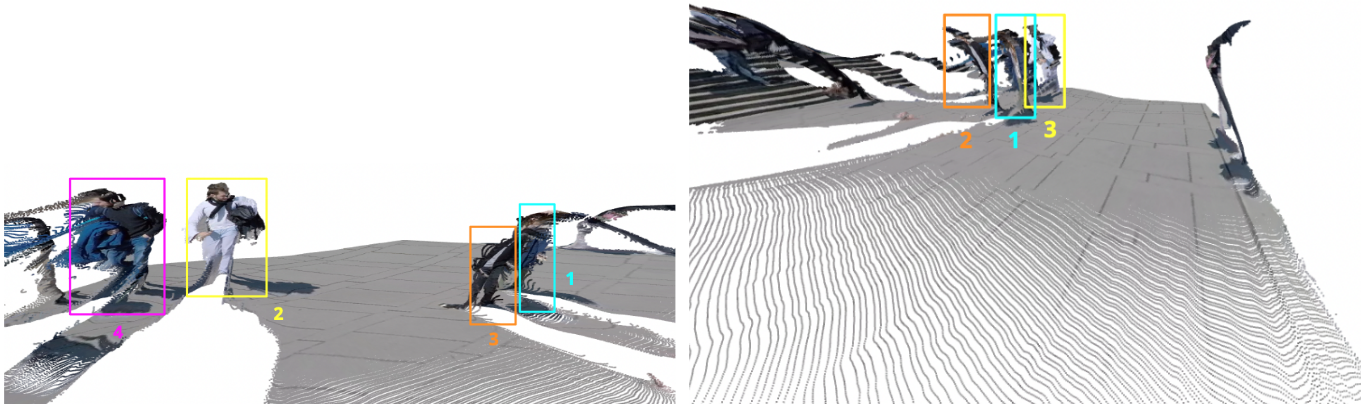
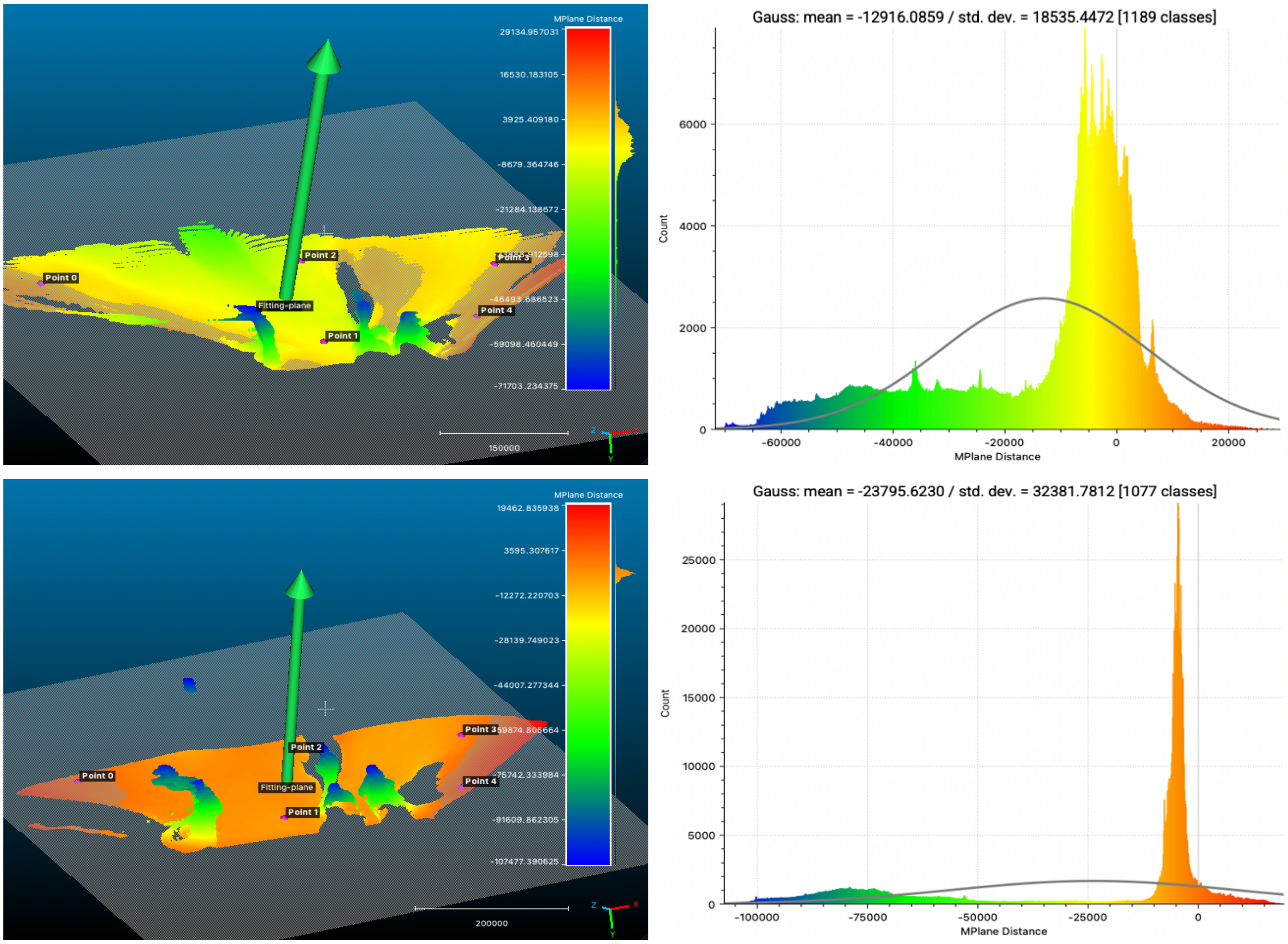


Figure 10: Experimental assessment of new perspectives. Annotated image (top) focuses on gaining pedestrian A and B real-life perspective. Both pedestrian A and B target (4) labeled and color-coded bounding boxes containing other pedestrians. Pedestrian A point-of-view (left). Pedestrian B point-of-view (right).

5.4 Point-Based Analysis



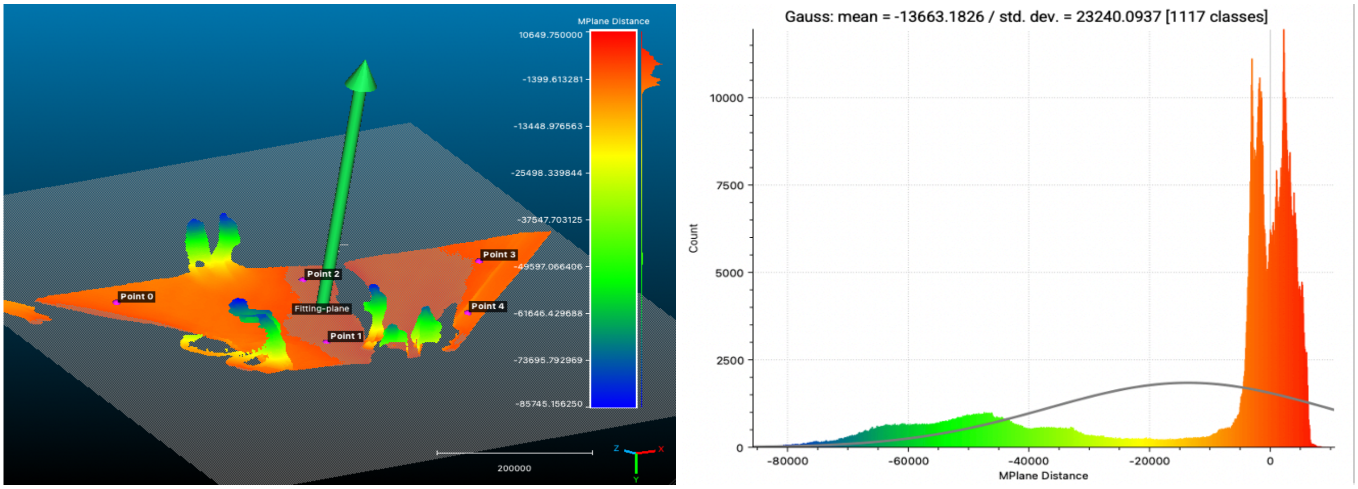


Figure 11. MPlane implementation using (5) points on Y-axis to measure distribution of pedestrian perpendicularity to the defined MPlane (X-axis). MiDaS Small-CNN (top), MiDaS Hybrid-DPT (middle), MiDaS Large-DPT (bottom).

VI. DISCUSSION OF RESULTS

6.1 Production-Rate Analysis

Production-rate analysis, defined by the speed and memory required of producing a point cloud from a single image, is an important factor to consider because 3D data requires a considerably longer time to analyze than 2D data. Figure 6 looks at how each model performed in production at varying levels of resolution. CNNs computational efficiency stands out compared to dense prediction transformers, allowing quicker time to produce PLY files and also at a fraction of memory.

6.2 Outlier-Based Analysis

Outlier-based analysis, defined by total outlier points divided by total points in the point set, aims to distinguish the difference of outlier detection performance. A lower percent of outliers detected indicates a higher probability that 3D geometries were compactly arranged. A higher percentage of outliers may indicate larger skewness in 3D geometries. Outliers were determined by a statistical measurement from two inputs: nearby neighbors (50 point sampled for every point) and standard deviation (2.0).

6.3 Visual Point Cloud Comparison

Provided with input image, Figure 9 assesses the overall point cloud shape from (3) different viewpoint angles to determine the differences from using the different MiDaS depth estimation models. CNN (top row) struggled in the Y-axis shape by producing fairly flat pedestrians in the background while the foreground pedestrians remained relatively slanted toward the ground. The Hybrid-DPT performed better by increasing the height in Y-axis for the pedestrians in the front, but ultimately did not perform well for the two pedestrians in the background as only one pedestrian's floating head was visible. Large-DPT scored highest in all viewpoint assessments with only one drawback being the ground remained slightly slanted than the Hybrid-DPT structure of the floor plane.

Figure 10 was generated from MiDaS Large-DPT as an experimental assessment of gaining new perspectives not directly shown in the input image. Here, the goal was to capture labeled pedestrian A and B's perspective in 3D, which is the exact opposite angle of the camera's position. The experimental assessment of gaining new perspective provides an interesting further discussion for future work to see which pedestrians view which objects.

6.4 Point-Based Analysis

Point-based analysis focused on the Y-axis, height, of the point cloud from the pedestrians on a flat plane, see Figure 11. This analysis aimed to assess two areas important to 3D geometric shapes: 1) flatness of the surface, 2) perpendicularity of the pedestrians from the flat plane. MPlane, a plugin to perform normal distances on a defined plane consisting of (5) points on the flat-plane surface (i.e. ground), was used to assess these qualities. Histograms on the corresponding point clouds are used as a way to assess the differences in the Y-axis structure. The assumption is that the most accurate histogram will display two peaks where one corresponds with the flatness of the surface and the other peak corresponding with the pedestrians.

Small-CNN: The figure and histogram show little delineation between ground and pedestrian levels. The histogram does not show two clear distinct peaks, which furthers the idea that the Small-CNN performs poorly in differentiating 3D structures. Pedestrians in the back of the point cloud are limited while pedestrians toward the front are more apparent, but still slope downward.

Hybrid-DPT: Clearer display of distinct peaks than the Small-CNN model. The floor surface remains relatively flat and highly concentrated on the left-side of “0” on the x-axis, which indicates that the model consistently gauged the ground level appropriately. The peaks of pedestrians perpendicularity to the ground level (green-blue) showed signs of appropriate 3D geometric relationships, but remained relatively flat compared to the Large-DPT model.

Large-DPT: Clearest display of the two peaks, ground and pedestrian levels, with larger distinction between peaks. The ground plane (red-orange) features a double-peak shape in the histogram which indicates larger variance in gauging the flatness of the ground.

VII. CONCLUSION

In this review, the research aimed to assess the quality of inferring point clouds from single-view monocular images by comparing two deep learning-based monocular depth estimation models. Convolutional neural networks reigned superior in computation efficiency, yet performed poorly in generating appropriate 3D geometries, specifically in regard to Y-axis perpendicularity of the objects and range in depth. The convolutional neural network also suffered from translational and rotational errors when viewing the same object (e.g. pedestrian) from different viewing angles, where the object stretches due to the misalignment. Dense prediction transformers are a recent architecture for computer vision and proved to be more accurate in 3D representations of full-scene images, but comes at the expense of time and memory. Translational and rotational errors when viewing objects from different viewpoints were minimal, which proved vision transformers to be a better approach than convolutional neural networks in terms of accuracy. In the end, the research also featured experimental results in gaining original perspectives from the opposite location of the camera’s direction in efforts to understand objects’ perspective within the full-scene environment.

For future work, I propose point cloud registration to build complete 3D reconstruction of a scene from multi-view datasets and implement MoveNet or MediaPipe, both human pose detection techniques, to obtain each pedestrian’s face landmarks in order to set the point cloud virtual camera to the exact position of the pedestrian’s nose or eyes. Reconstructing 2D images to full-fledge 3D models with this application can serve numerous industries by obtaining unique viewpoints and classifying objects within those viewpoints. Such implementation can serve retailers by mining which objects people are looking at the most or for the longest duration, law enforcement by deducing plausible witnesses for first person testimony to an event of interest, marketers by identifying hotspots for physical ad locations, and other areas where scanning viewpoints leads to the mining of data from a unique viewpoint that is not explicitly provided with 2D images.

REFERENCES

[1] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, “Deep Learning for 3D point clouds: A survey,” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 23-Jun-2020.

[2] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D object detection network for autonomous driving,” in CVPR, 2017.

[3] W. Liu, J. Sun, W. Li, T. Hu, and P. Wang, “Deep learning on point clouds and its application: A survey,” *Sensors (Basel, Switzerland)*, 26-Sep-2019.

[4] V. Pajić, M. Govedarica, and M. Amović, “Model of Point Cloud Data Management System in big data paradigm,” *MDPI*, 09-Jul-2018.

[5] B. Peasley, “[PDF] 1 3 D point cloud construction from stereo images: Semantic scholar,” *Semantic Scholar*, 2008.

[6] W. Zeng, S. Karaoglu, and T. Gevers, “Inferring point clouds from single monocular images by depth intermediation,” *ResearchGate*, Dec-2018.

[7] G. Kordelas, P. Daras, P. Klavdianos, E. Izquierdo, and Q. Zhang, “Accurate stereo 3D point cloud generation suitable for multi-view ...,” *ResearchGate*, Dec-2014.

[8] C. Benedek, A. Majdik, B. Nagy, Z. Rozsa, and T. Sziranyi, “Positioning and perception in Lidar Point Clouds,” *Science Direct*, Dec-2021.

[9] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe and E. Ricci, "Structured Attention Guided Convolutional Neural Fields for Monocular Depth Estimation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3917-3925, doi: 10.1109/CVPR.2018.00412.

[10] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision Transformers for dense prediction,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[11] “SceneScan/SceneScanPro User Manual,” (v1.7), Nerian Vision GmbH, Stuttgart Germany, 2019, pp 15-18.

[12] T. Chavdarova *et al*., "The WILDTRACK Multi-Camera Person Dataset," *IEEE,* 18-Jul-2017, pp. 5030-5039, doi: 10.1109/CVPR.2018.00528.

[13] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, D. Scaramuzza. The Event-Camera Dataset and Simulator: Event-based Data for Pose Estimation, Visual Odometry, and SLAM. International Journal of Robotics Research, Vol. 36, Issue 2, pages 142-149, Feb. 2017.

[14] T. Sercu and V. Goel, “Dense prediction on sequences with time-dilated convolutions for speech recognition,” *arXiv;1611.09288*, 14-Dec-2016.

[15] D. Elton, “Geoffrey Hinton on what's wrong with CNNs,” *Dan Elton*, 30-Sep-2017.

[16] N. Polanksy and Y. Hasbany, “Intelitouch Turning any surface into a touch surface,” *Technion Israel Institute of Technology*.