3D Object Detection with VI-SLAM Point Clouds: The Impact of Object and Environment Characteristics on Model Performance

Lin Duan[1], Tim Scargill[1], Ying Chen[1] and Maria Gorlatova[1]

***Abstract*— 3D object detection (OD) is a crucial element in scene understanding. However, most existing 3D OD models have been tailored to work with light detection and ranging (LiDAR) and RGB-D point cloud data, leaving their performance on commonly available visual-inertial simultaneous localization and mapping (VI-SLAM) point clouds unexamined. In this paper, we create and release two datasets: VIP500, 4772 VI-SLAM point clouds covering 500 different object and environment configurations, and VIP500-D, an accompanying set of 20 RGB-D point clouds for the object classes and shapes in VIP500. We then use these datasets to quantify the differences between VI-SLAM point clouds and dense RGB-D point clouds, as well as the discrepancies between VI-SLAM point clouds generated with different object and environment characteristics. Finally, we evaluate the performance of three leading OD models on the diverse data in our VIP500 dataset, revealing the promise of OD models trained on VI-SLAM data; we examine the extent to which both object and environment characteristics impact performance, along with the underlying causes.**

I. INTRODUCTION

In the realm of 3D object detection (OD) [1], [2], [3], [4], prevailing models have predominantly been trained on light detection and ranging (LiDAR) [5] and RGB-D [6], [7] point cloud data. In contrast, point cloud data generated through visual-inertial simultaneous localization and mapping (VI-SLAM) [8] has received much less attention. This type of point cloud is a vital consideration, given that VI-SLAM is widely used on resource-constrained mobile devices. For example, it is the standard mapping technique on augmented reality (AR) platforms, on which OD capabilities are of central importance [9], [10], and VI-SLAM point cloud sharing is integral to virtual content persistence in AR (e.g., [11]). However, the potential for OD models to extract information from VI-SLAM point clouds has not been explored until now.

VI-SLAM point cloud data introduces a specific set of challenges. Unlike LiDAR or RGB-D data, the accuracy, density, and spatial distribution of VI-SLAM data vary widely depending on object and background textures [12]. Consequently, even high-performing detection models, initially trained on LiDAR or RGB-D data, may exhibit poor performance when confronted with VI-SLAM data. Moreover, the diversity in VI-SLAM data, stemming from different object and environment characteristics, makes the development of models robust to all scenarios challenging. However, to the best of our knowledge, there is currently no work exploring the impact of object and environment characteristics on 3D OD using VI-SLAM point clouds.

To address this research gap, we conduct *the first systematic evaluation of 3D OD models on VI-SLAM point clouds, that includes the effect of object and environment characteristics on detection performance*. Our study focuses on OD in indoor environments, a common setting for VI-SLAM, and uncovers the influence of object shape and texture, as well as floor texture. Our main contributions are:

- We create and release two publicly available datasets[1]: **VIP500**, 4772 labeled VI-SLAM point clouds covering 500 object and environment configurations generated using a game engine, and the accompanying **VIP500-D**, 20 RGB-D point clouds generated from the object classes and shapes in VIP500.
- We quantify the fundamental differences between conventional RGB-D point clouds and VI-SLAM point clouds, as well as variations within VI-SLAM data, using the Density-aware Chamfer distance (DCD), a metric for calculating point cloud discrepancies [13].
- We assess the performance of three state-of-the-art (SOTA) 3D OD models on VI-SLAM point cloud data generated from diverse object and environment characteristics. We find that models trained on VI-SLAM data perform well with some object and environment characteristics but poorly with others, highlighting the potential for extracting information from VI-SLAM data, but also the need to develop more robust models.

Through these contributions, we shed light on the challenges and promise of 3D OD on VI-SLAM point cloud data, and reveal insights related to the impact of object and environment characteristics on OD performance.

II. BACKGROUND

A. Point Cloud Generation

Point cloud data can be acquired from 3D scanners that measure object-to-scanner distance, or generated from stereo- or multi-view 2D imagery [14]. Various types of 3D scanners, including LiDAR, laser stripe triangulators, and RGB-D cameras, produce point clouds with high spatial resolution and accuracy, and relatively uniform spatial distribution [15]. In contrast, the point clouds generated by VI-SLAM algorithms, including both SOTA open-source solutions (e.g., ORB-SLAM3 [16]) and 'black-box' commercial AR platforms (e.g., ARKit [17]), contain sparse points with nonuniform spatial distributions, and the accuracy of

[1]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. lin.duan@duke.edu, ts352@duke.edu, ying.chen151@duke.edu, maria.gorlatova@duke.edu

[1]https://github.com/timscargill/VIP-Datasets

(a) (b)

Fig. 1: Examples of the virtual environments we used to generate our VI-SLAM point cloud dataset: (a) chair object class example; (b) variations for this class (first row object shape, second row object texture, third row floor texture).

these points varies widely depending on the properties of surfaces present, e.g., their geometry and textures [12]. We generate a VI-SLAM point cloud dataset which covers diverse object and environment characteristics, detailed in Section III-A.

B. 3D Indoor OD

3D indoor OD, a crucial component of indoor scene understanding [18], localizes object instances and recognizes their categories within a scene. The performance of 3D indoor OD has garnered significant interest, particularly in the realm of service robots [19], [20] and AR applications [9], [10]. SOTA models include VoteNet [1] and H3DNet [2], which predict 3D bounding boxes and semantic classes based on point groups, GroupFree [3], which computes object features from all the points instead of grouping them, and FCAF3D [4] and TR3D [21], which leverage the synergy between geometric point cloud data and RGB inputs. However, none of these SOTA 3D OD models have been evaluated on VI-SLAM point cloud data. Here we evaluate their performance on our new VI-SLAM point cloud dataset (Section III-A), and examine the impact of object and environment characteristics.

C. Impact of Environment Characteristics on 3D OD

The inherent variability of environment conditions is known to influence model performance [22]. However, to our knowledge, there is currently no work investigating the impact of object and environment characteristics on 3D OD with VI-SLAM data. Prior research [22], [23], [24] has studied the effects of environment conditions on 3D OD with LiDAR data. Mai et al. [22] and Do et al. [24] delve into the effects of foggy and snowy scenes on LiDAR data, revealing substantial distortions that lead to diminished detection accuracy. Piroli et al. [23] emphasize how cold weather disrupts object size and orientation estimation. In contrast to the focus on LiDAR data, our work specifically examines 3D OD performance on VI-SLAM data, across diverse objects and environments.

III. OUR DATASETS

In this section, we introduce two new datasets that we have created as part of this work, one VI-SLAM point cloud dataset and one RGB-D point cloud dataset. Both datasets are publicly available in our GitHub repository[1].

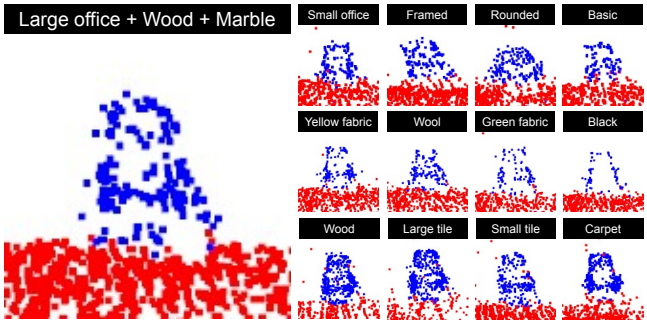


Fig. 2: Examples of the VI-SLAM data generated from the visual data shown in Fig. 1. The size, accuracy and distribution of an object's point clouds (blue points) are impacted by the object's shape and texture, and environment (floor) texture.

A. VI-SLAM Point Cloud Dataset

To study the effect of object and environment characteristics, we require a dataset in which these characteristics are varied in a systematic and controlled manner. To this end, we use a game engine-based methodology to generate *semi-synthetic data* [25], [26], [27], [28], comprising virtual visual data and real inertial data. As in [28], we create virtual environments in which we apply specific object and environment characteristics, use the ground truth trajectory from existing VI-SLAM datasets (e.g., [29], [30]) to generate camera images in those environments, then combine this new visual data with the inertial data from the original dataset.

As illustrated in Fig. 1, we created virtual environments in Unity 2020.3.14f1 consisting of a single object in a 8m×6m×4m room with blank walls and a textured floor. For each type of object (e.g., a chair) we generated different configurations in which we varied the object shape by using different 3D models (Fig. 1b, top row), the object texture (Fig. 1b, middle row), and the floor texture (Fig. 1b, bottom row). We used the A4 trajectory in the SenseTime VI-SLAM dataset [30] to generate a new sequence for each environment variant, then ran them on a SOTA open-source VI-SLAM algorithm, ORB-SLAM3 [16]; we modified the ORB-SLAM3 software to save the generated point cloud to a text file. Finally, because the exact transform between the ground truth game engine coordinate frame and the SLAM point cloud coordinate frame is not known (so the game engine coordinates cannot be used to segment the SLAM point cloud), we segmented and labeled the generated point clouds using the Open3D Python library [31]. We applied plane detection and outlier removal to identify points not part of the object, and appended a new column to each line in the point cloud file indicating the object class for that point. Examples of the VI-SLAM point clouds generated from the source data in Fig. 1 are shown in Fig. 2.

Using the above process, we created a dataset of 4772 labeled VI-SLAM point clouds, which we name **VIP500**. VIP500 covers 500 different environment configurations:

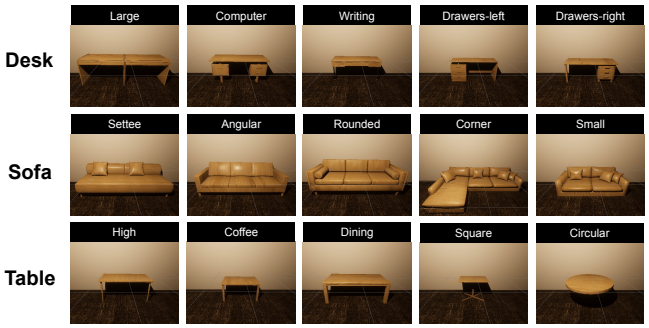


Fig. 3: Our Unity virtual environments show the five shapes we used for the desk, sofa, and table object classes (with the wood object texture and carpet floor texture).

four common indoor object classes from the ModelNet10 dataset [32] (chair, desk, sofa, and table) $\times$ five object shapes $\times$ five object textures $\times$ five floor textures. We ran 10 trials for each configuration; some configurations resulted in the loss of tracking in some trials and invalid point clouds, which were excluded from the dataset. The object and floor textures are consistent across classes, and shown in Fig. 1; the object shapes for the desk, sofa and table classes are shown in Fig. 3 (chair in Fig. 1), and the corresponding point clouds in Fig. 5 (chair in Fig. 2). The size, accuracy, and distribution of object point clouds (blue points) are not only impacted by the shape and texture of the object, but also by the floor texture.

B. RGB-D Point Cloud Dataset

We also created an RGB-D dataset that accompanies VIP500, to study the differences between the VI-SLAM point clouds and point clouds generated from 3D scanners. We generated the dataset using the same virtual environments with the same object shapes as those used in VIP500. We exported virtual environments built in Unity to FBX files and imported them to Unreal Engine 4.27.2. To generate our point clouds, we leveraged the Unreal plugin AirSim [33], which facilitates the creation of RGB-D point clouds after capturing RGB camera images and depth sensor readings.

Using this process, we created a dataset of 20 RGB-D point clouds, which we name **VIP500-D**. This dataset contains four object classes (chair, desk, sofa, and table), each with five object shapes. We do not consider different object and floor textures because these characteristics have minimal influence on RGB-D point clouds. Examples of the RGB-D point clouds in VIP500-D are shown in Fig. 4.

IV. METHODS

A. VI-SLAM and RGB-D Point Clouds Comparisons

We use the DCD introduced in [13] as a metric for quantifying point cloud discrepancies, to assess the differences between VI-SLAM and RGB-D point clouds, and the differences between VI-SLAM point clouds generated with

Fig. 4: RGB-D point clouds for different chair shapes used in our datasets, generated from the visual data shown in Fig. 1.

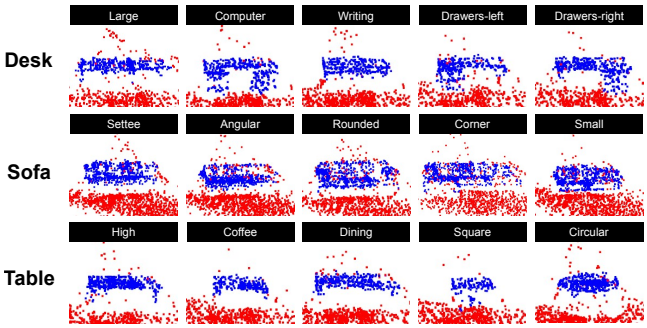


Fig. 5: Examples of the VI-SLAM point clouds generated from the visual data shown in Fig. 3, the five shapes used for the desk, sofa and table object classes.

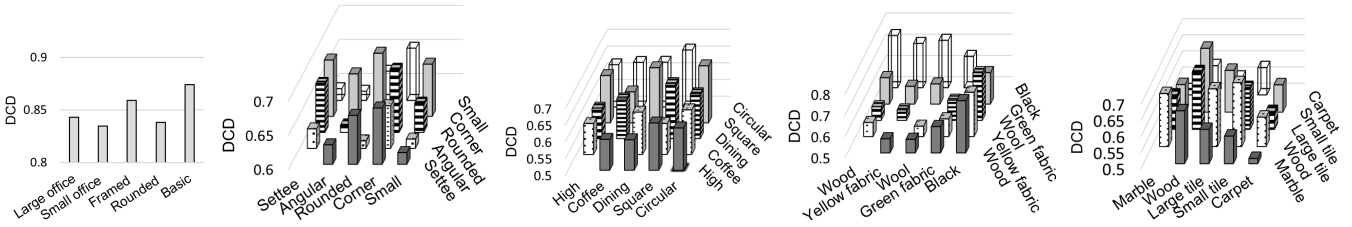
varied object and environment characteristics. DCD extends the conventional Chamfer distance by detecting the disparity of density distributions and removing outliers, making it a more comprehensive and reliable measure of similarity. Since the VI-SLAM and RGB-D point clouds may not share the same coordinate system, the first step is to use the iterative closest point (ICP) algorithm to rectify the translational and rotational discrepancies within the point clouds during comparisons. We do not adjust for the scale discrepancies, as VI-SLAM removes scale ambiguity through the integration of inertial measurements [16]. The ICP algorithm iteratively finds corresponding points between two point clouds and looks for a rigid transformation minimizing the alignment error [34]. We obtain two aligned point clouds S_A and S_B and calculate the DCD between them, denoted as $\mathcal{D}_c(S_A, S_B)$.

B. Performance of OD Models on VI-SLAM Point Clouds

3D OD Models: To conduct a comprehensive analysis, we evaluate the performance of three SOTA indoor 3D OD models: (1) VoteNet [1], (2) H3DNet [2] and (3) GroupFree [3]. These models are exclusively geometry-based, so that they only use the geometry information of the point cloud data.

VIP500 Dataset Grouping: To investigate the impact of object and environment characteristics on OD performance, we organize the VIP500 dataset into a total of 30 groups. To study the effect of object texture, we divide our VIP500 dataset into the five object texture types (wood, yellow fabric, wool, green fabric, black), and to study the effect of environment texture we divide it into the five floor texture types (marble, wood, large tile, small tile, carpet). These groups contain instances from all classes. To study the effect of object shapes, we divide the VIP500 dataset into 20 groups (4 classes $\times$ 5 object shapes), because object shapes cannot be compared across classes (so each object shape group only contains instances from one class).

Performance of Pre-trained OD Models: We start by examining the performance of SOTA 3D OD models pre-trained on ScanNetV2 [6], a commonly used OD benchmark dataset comprising 3D reconstructed indoor scene meshes, from which dense point clouds can be generated. ScanNetV2 contains $\sim$ 1.2K training examples derived from diverse rooms, and has 18 object categories. In line with the standard practice of deploying OD models in specialized environ-



(a) Chair shapes (b) Sofa shapes (c) Table shapes (d) Object textures (e) Floor textures

Fig. 6: (a) DCD between the RGB-D point clouds and the VI-SLAM point clouds for different chair shapes; (b–e) DCD between VI-SLAM point clouds for different sofa and table shapes, object textures, and floor textures.

ments, we source our pre-trained OD models from a well-known and widely-used GitHub repository [35].

To assess the performance of these pre-trained models on our VIP500 dataset, we employ the standard OD evaluation metric: mean average precision (mAP), reported as a percentage. We present mAP with an intersection over union (IoU) threshold of 0.25, rather than the conventional threshold of 0.5 [36], since the substantial domain gap between ScanNetV2 and VIP500 leads to nearly zero $mAP@.5$ across all test groups. To comprehensively evaluate model performance, we also report the mean average recall (mAR) with an IoU threshold of 0.25.

Performance of OD Models Trained on VI-SLAM Data: As fine-tuning is a well-established domain adaptation technique, we fine-tune all layers of SOTA 3D OD models with VI-SLAM data and evaluate their performance. This process illustrates the potential of using VI-SLAM data for scene understanding and forms a foundational exploration, paving the way for future investigations into more sophisticated adaptation methods specifically tailored for VI-SLAM data. We first define the object shape, object texture, and floor texture training sets; to limit the number of trained models to a tractable amount, for each of these three characteristics we choose one group (e.g., one texture) to train the models on, and test the models on the remaining data groups. To maximize fairness and prevent extreme cases where models are trained on the best- or worst-performing characteristics, we opt to train the OD models on the middle-performing VI-SLAM characteristics. These middle-performing characteristics are selected based on the results we obtain using the pre-trained models. Specifically, for the object and floor texture we train the models on the middle-performing texture and test it on the other four textures. For object shape, we train the model on the middle-performing chair, sofa, table, and desk shapes, and test it on the remaining 16 object shapes. To train the models, we use the AdamW optimizer with a 0.01 weight decay coefficient, and set the learning rate as 0.008, as provided in the GitHub repository [35]. To measure the performance of models trained on VI-SLAM data, we employ the standard $mAP@.5$.

V. RESULTS

We quantify the geometric disparities of point clouds in Section V-A. We then train three leading OD models on ScanNetV2 data and VI-SLAM data, and evaluate the OD performance on our VIP500 dataset in Sections V-B and V-C.

A. VI-SLAM and RGB-D Point Clouds Comparisons

We report the DCD $\mathcal{D}_c$ between the RGB-D and VI-SLAM point clouds in Fig. 6a, and among VI-SLAM point clouds with different object shapes, object textures, and floor textures in Figs. 6b-6e.

We observe that the discrepancies between RGB-D and VI-SLAM point clouds are large ($\mathcal{D}_c$ is 0.85 on average), especially compared with discrepancies among VI-SLAM point clouds with different object and environment characteristics ($\mathcal{D}_c$ is 0.63 on average). This highlights *the substantial domain gap between RGB-D and VI-SLAM point clouds*. This finding aligns with a visual comparison of the RGB-D point clouds with higher spatial resolution and relatively uniform spatial distributions, shown in Fig. 4, and the VI-SLAM point clouds characterized by sparser points with nonuniform distributions, shown in Figs. 2 and 5.

Figs. 6b-6c show $\mathcal{D}_c$ for pairs of VI-SLAM point clouds generated with different sofa shapes and table shapes, representative examples of different object shapes. We observe that $\mathcal{D}_c$ reflects the levels of visual discrepancies observed in Fig. 5. For example, the corner sofa models have the greatest average $\mathcal{D}_c$ to other sofa models; this finding aligns with our observations from Figure 5, in which corner sofa point clouds notably stand out in size compared to other models. Another example is that the square table has the largest $\mathcal{D}_c$ to the dining table; this also corresponds with Figure 5, in which the square and dining tables exhibit a large size discrepancy.

Figs. 6d-6e show $\mathcal{D}_c$ for pairs of VI-SLAM point clouds generated using different object textures or floor textures. We calculate the means and standard deviations of $\mathcal{D}_c$ for pairs of different object textures and different floor textures, denoted as $E(\mathcal{D}_c^o)$, $E(\mathcal{D}_c^f)$, $\sigma(\mathcal{D}_c^o)$ and $\sigma(\mathcal{D}_c^f)$. We observe that although $E(\mathcal{D}_c^f)$ (0.61) is smaller than $E(\mathcal{D}_c^o)$ (0.63), $\sigma(\mathcal{D}_c^f)$ (0.072) is larger than $\sigma(\mathcal{D}_c^o)$ (0.057). $\sigma(\mathcal{D}_c^f)$ being larger can be attributed to the influence of floor textures on VI-SLAM tracking accuracy, as well as their impact on the distribution of point clouds across objects and environments (e.g., floors). This observation suggests that *certain factors, such as floor textures, which have minimal impact on RGB-D data, can have a significant influence on VI-SLAM data*.

The geometric disparity of point clouds quantified by the DCD is one of the factors leading to performance degradation when training OD models on point clouds with specific object and environment characteristics, and subsequently testing them on point clouds with differing characteristics.

TABLE I: The $mAP@.25$ and $mAR@.25$ of three models pre-trained on ScanNetV2, across 30 groups. 71% of $mAP@.25$ values are $\leq 1\%$, showing the limited utility of pre-trained models on VI-SLAM point clouds. Best in bold, worst underlined.

Metrics	mAP@.25 (%)					mAR@.25 (%)				
Chair shapes	Large office	Small office	Framed	Rounded	Basic	Large office	Small office	Framed	Rounded	Basic
VoteNet	**0.01**	0	**0.01**	**0.01**	0	10.70	4.53	**12.45**	10.59	7.38
H3DNet	0.07	0.23	**0.24**	0.06	0.03	44.44	**79.42**	58.09	40.68	29.92
GroupFree	0.03	**0.05**	0.04	0.01	0.02	**53.50**	44.86	50.21	32.63	40.57
Desk shapes	Large	Computer	Writing	Drawers-left	Drawers-right	Large	Computer	Writing	Drawers-left	Drawers-right
VoteNet	2.75	1.92	1.00	1.49	**4.47**	**79.75**	71.60	54.77	57.02	75.10
H3DNet	0.07	**0.80**	0.10	0.01	0.63	28.10	**72.40**	42.32	14.88	19.50
GroupFree	0.01	0.01	0	**0.03**	0	5.60	4.96	1.24	**6.61**	3.73
Sofa shapes	Settee	Angular	Rounded	Corner	Small	Settee	Angular	Rounded	Corner	Small
VoteNet	0.11	0.10	0.74	0.02	**1.42**	25.13	23.48	**47.95**	7.66	18.37
H3DNet	0	0.09	0.02	**1.01**	0	6.15	**39.68**	22.13	22.58	0.82
GroupFree	0	0	0	0	0	0	**2.43**	1.64	0.81	0.82
Table shapes	High	Coffee	Dining	Square	Circular	High	Coffee	Dining	Square	Circular
VoteNet	18.2	19.18	**33.44**	5.26	17.38	78.51	59.29	76.15	49.12	**88.19**
H3DNet	**6.90**	2.47	1.77	0.15	0.56	45.45	**62.39**	45.19	14.16	24.89
GroupFree	0.02	0	0	0.01	**0.06**	4.13	7.08	2.51	**15.49**	15.19
Object textures	Wood	Yellow fabric	Wool	Green fabric	Black	Wood	Yellow fabric	Wool	Green fabric	Black
VoteNet	**8.66**	6.82	5.48	3.07	0.03	**58.31**	53.23	53.69	39.75	5.57
H3DNet	0.71	0.58	**3.92**	0.02	0.03	35.98	37.42	**86.60**	3.48	6.25
GroupFree	**0.03**	0.03	0.02	0.01	0	**23.02**	17.96	15.28	10.59	3.77
Floor textures	Marble	Wood	Large tile	Small tile	Carpet	Marble	Wood	Large tile	Small tile	Carpet
VoteNet	4.33	5.36	2.25	2.82	**9.09**	40.45	42.53	**45.02**	41.70	44.91
H3DNet	**6.02**	0.23	0.08	0.13	3.50	**71.31**	12.98	21.05	12.92	68.01
GroupFree	0.01	0.01	0.02	0	**0.03**	13.45	11.38	16.40	6.85	**24.42**

We investigate this aspect below.

B. Performance of Pre-trained OD Models

We conduct evaluations on three pre-trained 3D OD models using the 30 data groups detailed in Section IV-B. The $mAP@.25$ and $mAR@.25$ of pre-trained models for these groups are shown in Table I. We observe that the $mAP@.25$ of the three OD models is low ($< 34\%$ in all cases, $< 10\%$ for all except VoteNet on the table class) compared to the $mAP@.25$ of models trained on the ScanNetV2 dataset (62.34% for VoteNet, 66.07% for H3DNet and 66.17% for GroupFree). This result reveals *the substantial domain gap between VI-SLAM point cloud data and RGB-D point cloud data*. In addition to $mAP@.25$, we note that better performance is achieved for the $mAR@.25$ metric. This implies that while numerous objects are detected by the pre-trained models, a significant proportion of these detections are incorrect.

Considering both $mAP@.25$ and $mAR@.25$, we find that the object shapes which result in the best performance are the framed chair, the angular sofa, the circular table, and the computer and drawers-left desk, while wood and carpet are the best-performing object and floor textures respectively. The characteristics which result in the worst performance are the basic chair, the settee, corner and small sofas, the square table, and the writing desk, along with the black object texture and small tile floor texture. From the corresponding point cloud data in Fig. 2 and Fig. 5, we observe that VI-SLAM data generated with the best-performing characteristics frequently bear a closer resemblance to the RGB-D data (e.g., the dense point cloud generated with the wood object texture or the circular table shape), while VI-SLAM data generated with the worst-performing characteristics exhibit greater dissimilarity from RGB-D data (e.g., the sparse point cloud generated with the black object texture or the square table shape). Although the performance of pre-trained models varies across different VI-SLAM data groups, the overall performance is low (e.g., 71% of the $mAP@.25$ values are $\leq 1\%$), which indicates that *directly applying OD models pre-trained on dense data to VI-SLAM point clouds provides limited utility, underscoring a critical need for domain adaptation techniques [37], [38].*

C. Performance of OD Models Trained on VI-SLAM Data

As explained in Section IV-B, we fine-tune three OD models on the middle-performing characteristics, defined according to the test performance presented in Table I. The middle-performing data characteristics are the large office chair shape, the rounded sofa shape, the dining table shape, the computer desk shape, the wool object texture, and the marble floor texture.

The $mAP@.5$ values across the remaining 24 groups are shown in Table II. Performance is better with the OD models trained on VI-SLAM data than the pre-trained OD models (Section V-B), due to the differences among VI-SLAM data being smaller than the differences between VI-SLAM data and RGB-D data, as shown in Fig. 6. Importantly though, *OD performance on VI-SLAM point clouds varies greatly across different object and environment characteristics, even when models are trained on VI-SLAM data*. As we discuss below, in some scenarios SOTA OD models may provide informative outputs from VI-SLAM point clouds, while in others they perform much more poorly.

The OD models generally perform better on object shapes

TABLE II: The mAP@.5 of OD models trained on VI-SLAM data across 24 groups varies from 0.51% to 76.66%. Best in bold, worst underlined.

Metric	mAP@.5 (%)			
Chair shapes	Small office	Framed	Rounded	Basic
VoteNet	14.48	2.95	**24.57**	10.03
H3DNet	5.19	3.97	3.36	**6.21**
GroupFree	37.68	34.28	21.77	**43.29**
Desk shapes	Large	Writing	Drawers-left	Drawers-right
VoteNet	52.86	**58.73**	33.36	13.10
H3DNet	9.36	**12.30**	7.51	4.07
GroupFree	13.80	**49.71**	25.59	30.95
Sofa shapes	Settee	Angular	Corner	Small
VoteNet	69.77	**76.66**	31.41	45.76
H3DNet	**25.14**	17.45	0.51	0.78
GroupFree	62.64	**66.07**	59.31	54.92
Table shapes	High	Coffee	Square	Circular
VoteNet	**23.12**	16.51	5.68	12.57
H3DNet	**8.41**	2.39	4.8	6.41
GroupFree	**33.92**	22.94	9.75	25.32
Object textures	Wood	Yellow fabric	Green fabric	Black
VoteNet	66.78	**68.69**	44.20	6.08
H3DNet	**76.24**	75.83	51.17	7.27
GroupFree	**75.71**	73.53	44.23	5.32
Floor textures	Wood	Large tile	Small tile	Carpet
VoteNet	35.55	14.67	**45.59**	32.00
H3DNet	46.16	14.62	**54.30**	44.71
GroupFree	40.62	29.01	**51.59**	46.47

that are more similar to the shapes in the training data. For example, point clouds from the best-performing shape, the angular sofa, align closely with those from the rounded sofa used for training, evidenced by a low DCD of 0.61. In contrast, the DCD between the rounded sofa and the worst-performing shape, the corner sofa, is higher at 0.69. This highlights the need to consider how well object shapes are represented in training data.

For the object textures, the $mAP@.5$ values are highest with the feature-rich wood and yellow fabric textures ($65-80\%$ for all OD models), and by far the lowest with the plain black texture ($< 10\%$ for all OD models). While the DCD between the point clouds from the black texture and the wool texture used for training is not the largest, the point clouds generated with the black texture exhibit a high level of sparsity (Fig. 2). In fact, *of all the characteristics we studied, object texture had the most consistent effect on OD performance*, intuitive given that feature-based VI-SLAM algorithms like ORB-SLAM3 are designed to store information about recognizable textures, and thus an object's texture directly affects how much information is available about it in VI-SLAM point clouds. This result mirrors those from studies of the effect of texture on VI-SLAM pose tracking, in that textureless environment regions degrade tracking performance [28], [39], [30]. Moreover, black textureless regions (e.g., TVs, monitors) have also been shown to result in poor performance of the infrared time-of-flight depth sensors sometimes incorporated into mobile devices [40]. Given the prevalence of these types of surfaces in built environments, the development of OD models robust to sparse VI-SLAM point cloud data is an important topic for future work.

For the floor textures, the highest $mAP@.5$ values ($45-55\%$) are consistently achieved with the small tile texture, and the the lowest $mAP@.5$ values ($< 30\%$) with the large tile texture. *This reveals the surprising extent to which environment texture, as well as object texture, impacts OD performance on VI-SLAM data*. However, a notable exception arises: the worst-performing texture, large tile texture, has a small DCD with the marble texture used for training (Fig. 6e). We note that with the large tile texture, VI-SLAM-based tracking results in less accurate point clouds, degrading OD performance. As illustrated in Fig. 2, different environment textures can also result in different object point cloud densities (due to the selection of a fixed number of keypoints in each frame in the VI-SLAM algorithm), and the extent to which this occurs in a wider range of environments is worthy of further investigation.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we assess the performance of 3D OD models on VI-SLAM point clouds, including the impact of object and environment characteristics. To facilitate this, we create and release a dataset of VI-SLAM point clouds covering 500 different object and environment configurations, along with an accompanying dataset of RGB-D point clouds. Leveraging these datasets, we systematically quantify the disparities between VI-SLAM and RGB-D point clouds, as well as variations between VI-SLAM point clouds generated with different object and environment characteristics. We evaluate three SOTA 3D OD models on VI-SLAM data, both when the models are trained on dense point clouds, and when they are trained on VI-SLAM data.

Overall, our results demonstrate the promise of OD on the type of VI-SLAM point clouds that are readily available on a wide range of mobile devices. Despite SOTA OD models trained on dense point clouds providing little useful information, we show that training them on VI-SLAM point clouds dramatically improves performance, especially when objects of interest are textured. Indeed, given that VI-SLAM point cloud sharing is an integral part of current techniques for virtual content persistence in AR (e.g., [11]), these results raise important concerns about the potential for attackers to infer information about an environment using OD models. There is also significant scope to explore how OD performance can be improved for the object and environment characteristics we identified as more challenging. This improvement could be achieved by collecting more diverse training data, up-sampling sparse point clouds, or using self-supervised learning [41], contrastive learning [42], and other domain adaptation methods.

ACKNOWLEDGMENT

This work was supported in part by NSF grants CSR-1903136, CNS-1908051, CNS-2312760, and CNS-2112562, NSF CAREER Award IIS-2046072, a CISCO Research Award, and a Meta Research Award.

REFERENCES

[1] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep Hough Voting for 3D Object Detection in Point Clouds," in *Proceedings of IEEE/CVF ICCV*, 2019.

[2] Z. Zhang, B. Sun, H. Yang, and Q. Huang, "H3DNet: 3D Object Detection using Hybrid Geometric Primitives," in *Proceedings of ECCV*, 2020.

[3] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3D Object Detection via Transformers," in *Proceedings of IEEE/CVF ICCV*, 2021.

[4] D. Rukhovich, A. Vorontsova, and A. Konushin, "FCAF3D: Fully Convolutional Anchor-free 3D Object Detection," in *Proceedings of ECCV*, 2022.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[6] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Proceedings of IEEE/CVF CVPR*, 2017.

[7] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite," in *Proceedings of IEEE/CVF CVPR*, 2015.

[8] G. Jia, X. Li, D. Zhang, W. Xu, H. Lv, Y. Shi, and M. Cai, "Visual-SLAM Classical Framework and Key Techniques: A Review," *Sensors*, vol. 22, no. 12, p. 4582, 2022.

[9] Y. Guan, X. Hou, N. Wu, B. Han, and T. Han, "DeepMix: Mobility-aware, Lightweight, and Hybrid 3D Object Detection for Headsets," in *Proceedings of ACM MobiSys*, 2022.

[10] L. Liu, H. Li, and M. Gruteser, "Edge Assisted Real-time Object Detection for Mobile Augmented Reality," in *Proceedings of ACM MobiCom*, 2019.

[11] Apple, "ARWorldMap," 2024, https://developer.apple.com/documentation/arkit/arworldmap.

[12] K. Sartipi, T. Do, T. Ke, K. Vuong, and S. I. Roumeliotis, "Deep Depth Estimation from Visual-inertial SLAM," in *Proceedings of IEEE/RSJ IROS*, 2020.

[13] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, and D. Lin, "Density-aware Chamfer distance as a comprehensive metric for point cloud completion," in *Proceedings of NeurIPS*, 2021.

[14] Y. Xie, J. Tian, and X. X. Zhu, "Linking Points with Labels in 3D: A Review of Point Cloud Semantic Segmentation," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 38–59, 2020.

[15] 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*, vol. 43, no. 12, pp. 4338–4364, 2020.

[16] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-source Library for Visual, Visual–inertial, and Multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[17] Apple, "More to Explore with ARKit 6," 2024, https://developer.apple.com/augmented-reality/arkit/.

[18] M. Naseer, S. Khan, and F. Porikli, "Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey," *IEEE Access*, vol. 7, pp. 1859–1887, 2018.

[19] C. M. Sánchez, M. Zella, J. Capitán, and P. J. Marrón, "Semantic Mapping with Low-Density Point-Clouds for Service Robots in Indoor Environments," *Applied Sciences*, vol. 10, no. 20, p. 7154, 2020.

[20] Z. Yan, T. Duckett, and N. Bellotto, "Online Learning for 3D LiDAR-based Human Detection: Experimental Analysis of Point Cloud Clustering and Classification Methods," *Autonomous Robots*, vol. 44, pp. 147–164, 2020.

[21] D. Rukhovich, A. Vorontsova, and A. Konushin, "TR3D: Towards Real-time Indoor 3D Object Detection," *arXiv preprint arXiv:2302.02858*, 2023.

[22] N. A. M. Mai, P. Duthon, L. Khoudour, A. Crouzil, and S. A. Velastin, "3D Object Detection with SLS-fusion Network in Foggy Weather Conditions," *Sensors*, vol. 21, no. 20, p. 6711, 2021.

[23] A. Piroli, V. Dallabetta, M. Walessa, D. Meissner, J. Kopp, and K. Dietmayer, "Robust 3D Object Detection in Cold Weather Conditions," in *Proceedings of IEEE IV*, 2022.

[24] A. T. Do and M. Yoo, "LossDistillNet: 3D Object Detection in Point Cloud under Harsh Weather Conditions," *IEEE Access*, vol. 10, pp. 84 882–84 893, 2022.

[25] T. Sayre-McCord, W. Guerra, A. Antonini, J. Arneberg, A. Brown, G. Cavalheiro, Y. Fang, A. Gorodetsky, D. McCoy, S. Quilter, *et al.*, "Visual-inertial Navigation Algorithm Development using Photorealistic Camera Simulation in the Loop," in *Proceedings of IEEE ICRA*, 2018.

[26] W. Guerra, E. Tal, V. Murali, G. Ryou, and S. Karaman, "FlightGoggles: Photorealistic Sensor Simulation for Perception-driven Robotics using Photogrammetry and Virtual Reality," in *Proceedings of IEEE IROS*, 2019.

[27] A. Antonini, W. Guerra, V. Murali, T. Sayre-McCord, and S. Karaman, "The Blackbird Dataset: A Large-Scale Dataset for UAV Perception in Aggressive Flight," in *Proceedings of ISER*, 2018.

[28] T. Scargill, Y. Chen, N. Marzen, and M. Gorlatova, "Integrated Design of Augmented Reality Spaces using Virtual Environments," in *Proceedings of IEEE ISMAR*, 2022.

[29] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, "The TUM VI Benchmark for Evaluating Visual-inertial Odometry," in *Proceedings of IEEE/RSJ IROS*, 2018.

[30] L. Jinyu, Y. Bangbang, C. Danpeng, W. Nan, Z. Guofeng, and B. Hujun, "Survey and Evaluation of Monocular Visual-inertial SLAM Algorithms for Augmented Reality," *Virtual Reality & Intelligent Hardware*, vol. 1, no. 4, pp. 386–410, 2019.

[31] Open3D, "Open3D - A Modern Library for 3D Data Processing," 2024, http://www.open3d.org/.

[32] 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 IEEE/CVF CVPR*, 2015.

[33] Microsoft, "AirSim," 2021, https://microsoft.github.io/AirSim/point_clouds/.

[34] V. Kubelka, M. Vaidis, and F. Pomerleau, "Gravity-constrained Point Cloud Registration," in *Proceedings of IEEE/RSJ IROS*, 2022.

[35] OpenMMLab, "MMDetection3D," 2024, https://github.com/open-mmlab/mmdetection3d.

[36] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," in *Proceedings of the IEEE*, 2023.

[37] T. Wang, X. Zhang, L. Yuan, and J. Feng, "Few-shot adaptive faster r-cnn," in *Proceedings of IEEE/CVF CVPR*, 2019.

[38] G. Li, Z. Ji, and X. Qu, "Stepwise domain adaptation (SDA) for object detection in autonomous vehicles using an adaptive centernet," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 729–17 743, 2022.

[39] T. Scargill, M. Hadziahmetovic, and M. Gorlatova, "Invisible Textures: Comparing Machine and Human Perception of Environment Texture for AR," in *Proceedings of ACM ImmerCom*, 2023.

[40] Y. Zhang, T. Scargill, A. Vaishnav, G. Premsankar, M. Di Francesco, and M. Gorlatova, "InDepth: Real-time Depth Inpainting for Mobile Augmented Reality," in *Proceedings of ACM IMWUT*, 2022.

[41] J. Xu, L. Xiao, and A. M. López, "Self-supervised Domain Adaptation for Computer Vision Tasks," *IEEE Access*, vol. 7, pp. 156 694–156 706, 2019.

[42] M. Thota and G. Leontidis, "Contrastive Domain Adaptation," in *Proceedings of IEEE/CVF CVPR*, 2021.