2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Object-Oriented Material Classification and 3D Clustering for Improved Semantic Perception and Mapping in Mobile Robots

Siva Krishna Ravipati[1] Ehsan Latif[2] Ramviyas Parasuraman[1,2,*] Suchendra M. Bhandarkar[1,2]

***Abstract*— Classification of different object surface material types can play a significant role in the decision-making algorithms for mobile robots and autonomous vehicles. RGB-based scene-level semantic segmentation has been well-addressed in the literature. However, improving material recognition using the depth modality and its integration with SLAM algorithms for 3D semantic mapping could unlock new potential benefits in the robotics perception pipeline. To this end, we propose a complementarity-aware deep learning approach for RGB-D-based material classification built on top of an object-oriented pipeline. The approach further integrates the ORB-SLAM2 method for 3D scene mapping with multiscale clustering of the detected material semantics in the point cloud map generated by the visual SLAM algorithm. Extensive experimental results with existing public datasets and newly contributed real-world robot datasets demonstrate a significant improvement in material classification and 3D clustering accuracy compared to state-of-the-art approaches for 3D semantic scene mapping.**

***Index Terms*— Material Classification, Semantic Mapping, 3D Clustering, Mobile Robots, SLAM, RGB-D Data**

I. Introduction

Recent advancements in mobile robotics have underscored the importance of autonomous navigation and manipulation in unknown environments. A pivotal challenge in this domain is the accurate identification and classification of surface materials of objects, a capability crucial for effective decision-making and interaction within these environments, whether for exploration, manipulation, or clearing tasks [1]. Specifically, the deployment of robots in exploration and mapping applications [2] benefits from accurate perception and understanding of the environment, especially when integrated with SLAM (Simultaneous Localization and Mapping) algorithms [3]. Whether in domestic settings, firefighting, or logistics, understanding the material composition of surroundings is crucial for preplanning operations and navigating effectively [4]. For instance, distinguishing between concrete and black ice is essential for safely operating self-driving vehicles and service robots. Incorporating material recognition into conventional object recognition, scene understanding, and SLAM pipeline can significantly enhance robot performance, especially in use cases involving physical interactions and realistic renderings in virtual environments [5].

Current computer vision methods for material classification often focus on visual cues (shape and color) from RGB images. However, traditional RGB-based scene-level semantic (material) segmentation methods provide limited insights into the surface properties of the objects. While generally effective, they fall short in providing the nuanced material recognition necessary for robotic tasks, thus motivating the use of depth modality (i.e., RGB+D) for extracting rich features [6]. The challenge lies in effectively utilizing depth maps to predict material types from camera images, which becomes critical in complex and dynamic environments [7]. Therefore, creating a semantic material map using RGB-D images presents a valuable research avenue. Such maps enhance scene understanding and aid in robotic exploration and manipulation. Additionally, point cloud mapping, which creates a 3D representation of the environment, is invaluable for mapping and navigation, as well as in generating detailed models for mixed reality and architectural planning [3].

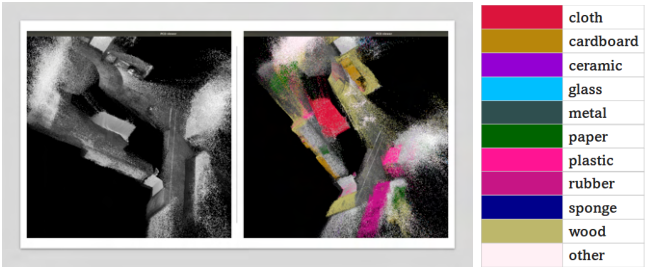


Fig. 1: Illustration of the objected-oriented 3D semantic mapping with material-level information (right) based on the RGB-D point clouds (left), shown along with the labels.

Motivated by the need to enhance robotic perception, particularly in the context of material classification, we introduce a novel approach to material classification and semantic mapping for mobile robots. Our framework presents a unique result, as illustrated in Fig. 1, which showcases the point cloud output combined with object-oriented material identification and clustering. We integrate our approach into the well-recognized ORB-SLAM2 algorithm [3] leveraging the benefits of visual odometry and SLAM, providing a real-time, accurate global map essential for mobile robots.

The key contributions in this paper include:

- A novel material classification network through complementarity-aware fusion of RGB and Depth-based convolutional neural networks, built on top of an RGB-based object detection pipeline for fast (real-time) and accurate material classification. The approach takes advantage of the extraction and fusion of distinctive and correlated features from RGB and depth modalities.
- Integration of the material classification outcome with the RGB-D SLAM using a voxel-based multiscale feature matching technique to obtain a precise metric-

[1] Institute for Artificial Intelligence, University of Georgia, Athens, GA 30602, USA
[2] School of Computing, University of Georgia, Athens, GA 30602, USA
[*] Corresponding author email: ramviyas@uga.edu

semantic mapping, a 3D environmental map consistently clustered by the material properties of the objects.
- An extensive experimental evaluation of our architecture with multiple real-world datasets (standard and custom) on material classification and 3D semantic mapping of complex environments showing significant improvement in accuracy over the state-of-the-art (SOTA).

Finally, we contribute new real-world mobile robot RGB-D datasets with meaningful object and material classes (as ROS bags) and open source the relevant codes[1] to benefit the community. Equipping robots with the ability to discern and cluster material properties accurately enables more reliable and efficient task execution across a range of applications, from domestic service to industrial automation [8].

II. Related Work

Recent advancements in vision-based SLAM and material recognition have leveraged deep learning to achieve notable progress. The complexity of indoor environments and the diverse material composition of objects therein make this an especially challenging and relevant problem [9], [7]. Mur-Artal and Tardós [3] introduced ORB-SLAM2, an efficient SLAM pipeline for monocular, stereo, and RGB-D cameras. It provides high accuracy but lacks semantic understanding. In the domain of material recognition, Qi et al. [10] proposed PointNet++, which directly processes point clouds for 3D classification and segmentation. Although they process point clouds efficiently, these models do not consider the material properties of objects. Our approach complements this by providing material-level semantic information.

Chen et al. [11] developed a progressively complementarity-aware (CA) fusion network for RGB-D salient object detection. Their method effectively fuses RGB and depth features but is not tailored for material classification. Our proposed CA fusion module extends this idea to material recognition, enhancing the accuracy of object-material association. Schwartz and Nishino [12] focused on material recognition from a global context, emphasizing the role of local features. Their work, while insightful, does not incorporate depth information, which is critical for distinguishing materials with similar textures. By incorporating depth data, our method provides a more robust solution for material classification in complex scenes.

In robot mapping, Zhao et al. [9] implemented a deep learning method for 3D reconstruction and material recognition, yet their approach does not effectively handle dynamic environments. Similarly, the study in [13] achieves simultaneous material segmentation and 3D reconstruction, primarily in static industrial settings, but it falls short in adapting to changing environments that are integral to mobile robots. Our method addresses these limitations with a voxel-based matching component, significantly enhancing the SLAM system's adaptability in dynamic scenarios. This is a critical improvement over existing methods, as it enables accurate real-time mapping and material recognition in environments where conditions and object placements are constantly evolving. Furthermore, works in [5], [14], [15] contributed to integrating object-level semantic information with SLAM to improve localization accuracy through graph-based pose corrections, but these methods do not differentiate different materials. While we do not focus on improving the localization, our approach fills the gap by adding material-level semantic maps with a vision to extend the capabilities of robotic systems in complex, variable settings by providing a detailed and dynamic semantic material map essential for interaction tasks such as grasping and manipulation.

In a recent work [16], the authors proposed an online 2D and 3D semantic, modular map representation and object detection framework using RGB-D data over-refinement and likelihood maintenance to avoid false detection. However, the object-based likelihood maintenance mechanism may miss out on semantically important objects that are occluded by other objects and have a low hit-to-miss ratio in likelihood calculation. In our approach, we use a voxel-based feature matching technique, which considers all the objects irrespective of their occurrence frequency.

In summary, our approach departs from the above SOTA methods by integrating RGB-D data fusion and clustering object-level results with point clouds for consistent semantic mapping in 3D environments. Leveraging a complementarity-aware (CA) fusion module, our system synergistically combines RGB and depth data, enabling more refined material classification than traditional models that treat these modalities in isolation. This method captures unique visual characteristics and incorporates depth information, crucial for discerning materials with similar appearances under varying lighting conditions. Unlike existing methods that might overlook material attributes in environmental mapping, our approach recognizes and classifies materials with improved accuracy. Additionally, using RGB-D images can improve the system's robustness to lighting changes and provide more accurate representations of the environment in low-light conditions.

Another key aspect of our method is incorporating voxel-based matching within the SLAM framework. This component ensures both dynamic and static objects in the point cloud map are correctly identified and associated with their material types. Such an advancement is vital in robotics, where accurate material recognition influences tasks like manipulation or navigation. Our approach extends the capabilities of existing systems like ORB-SLAM2 [3], PointNet [10], and SemanticFusion [5] by addressing their limitations in depth and semantic understanding and maintaining the consistency of semantic mapping in 3D, a feature not fully realized in prior works. Consequently, our approach offers a more robust, real-time semantic material map, greatly enhancing a robot's perception and interaction abilities in complex and dynamically changing environments. As a result, combining object detection and point cloud mapping with RGB-D images can provide a rich and detailed representation of the environment, enhancing performances in various tasks such as localization, navigation, and object manipulation.

[1] https://github.com/herolab-uga/matsee

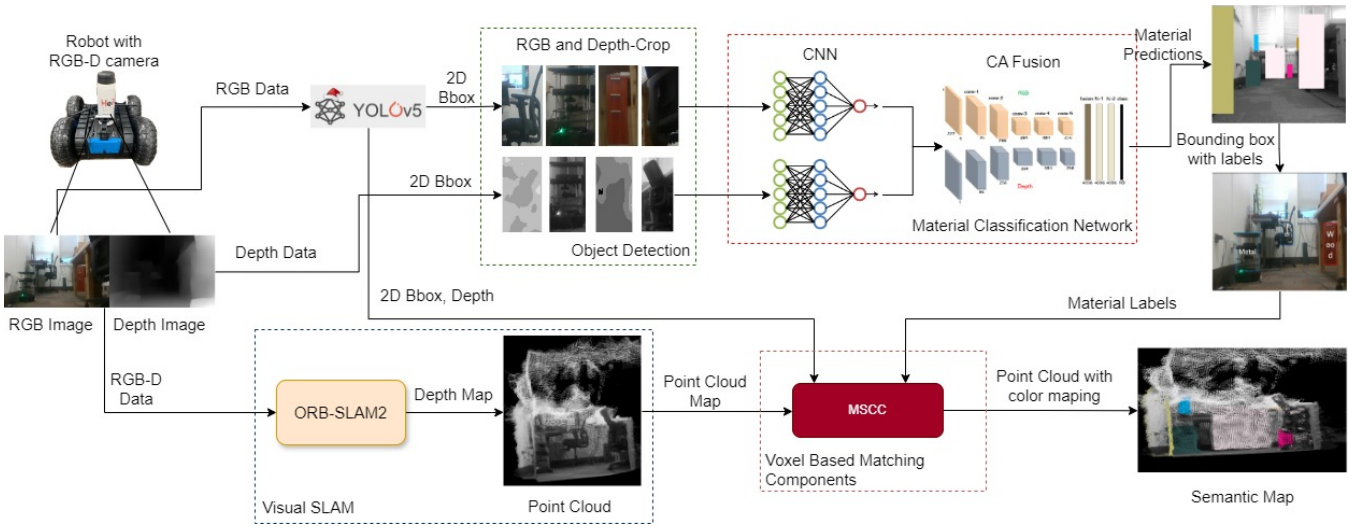


Fig. 2: Architectural overview of the proposed object-oriented 3D semantic mapping with material labels. The visual SLAM (VSLAM) component generates the point cloud map, and the YOLO component (OBJ) detects objects and locates the bounding boxes of the objects in the images. The material classification network (MCN) classifies the objects in the bounding boxes into different material classes. The voxel-based matching component (VOXM) uses the point cloud map generated by visual SLAM and the material labels obtained from the material classification component to match the 3D coordinates of the bounding boxes with the 3D coordinates of the point cloud and propagate the material labels to the points in the point cloud.

III. Proposed Approach

The core problem we address is the semantic mapping of materials in a 3D environment captured by a mobile robot. Mathematically, this involves identifying and classifying various materials within the robot's field of view, represented by a 3D point cloud. Let $\mathcal{P}$ denote the point cloud, where each point $p_i \in \mathcal{P}$ has associated f_{RGB} color and f_{Depth} depth feature map obtained through fusion mechanisms. The goal is to assign a material label $l_i \in \mathcal{L}$ to each point p_i associated with fused map f_{fuse}, forming clusters of points (map) with similar material properties.

The proposed solution involves three main pillars. It first detects objects in the environment using a fast object detection pipeline (Sec. III-A), followed by material classification of these detected objects (Sec. III-B). Using these classified features with a 3D point cloud map generated by a visual SLAM (VSLAM) algorithm (e.g., ORB-SLAM2), a feature matching technique described in Sec. III-D is applied to create a comprehensive semantic map of the environment. An architectural overview of the proposed approach is shown in Fig. 2. This integrated approach is designed to enhance the perception capabilities of mobile robots by enabling them to recognize and accurately classify materials in their environment. Algorithm 1 provides the pseudocode description of our semantic mapping of materials from the RGB-D data.

A. Object Detection (OBJ) Pipeline

Our approach uses an object-oriented pipeline, meaning that the object detections trigger the material classifications. Input RGB images (from an RGB-D camera) are used for object detection, and we build our architecture on top of the SOTA YOLO (You Only Look Once) model [17] because of its fast, robust, accurate, and versatile real-time object detection capabilities. We specifically used the YOLOv5 version, as it had proven to be robust in the ROS framework[2] (for SLAM integrations) with excellent performance on several benchmarks, including the COCO object detection dataset [18]. Regardless, the research community has consistently upgraded the YOLO-based models. For instance, the newer YOLOv8 [19] can replace the YOLOv5 in our pipeline. In our approach, we have utilized feature pyramid networks, which allow the model to process inputs at numerous scales and produce multiscale predictions. This process enables the model to detect small and large objects in the same image, which can be challenging for other object detection models. The output of YOLO (bounding boxes of all detected objects along with their labels) is used further for material classification and localization in a 3D semantic map. We have customized the YOLO model (building on top of a pre-trained model with the COCO dataset [18]) to add five new object classes: *board, door, mat, robot, and trash bin*, as these new classes are repetitive in our academic office/lab settings, and will be used in the later mapping procedure.

B. Material Classification Network (MCN)

Once the objects are detected, each object is cropped from the RGB and the aligned Depth images using the object's bounding boxes. Because of the different image generation mechanisms between RGB and depth images, fusing cross-modal features effectively is a key issue for RGB-D-based material classification. In our work, we exploit the concept of complementarity-aware (CA) fusion proposed in [11] as it effectively merges the distinct features from RGB and depth data, addressing the limitations of relying solely on

[2] https://ros.org/

Algorithm 1: 3D Semantic Mapping of Materials

Input: RGB-D data stream;
Models: Object detection model (OBJ), material classification network (MCN), visual SLAM framework (VSLAM), and voxel-based matching component (VOXM);
Return: 3D semantic map $\mathcal{SM}$ (point cloud clusters) with object and material labels;
while *robot is navigating the environment* **do**
    Capture current RGB frame and depth data;
    Detect objects using OBJ on the RGB frame and generate 2D bounding boxes;
    for *each detected object* **do**
        Crop corresponding RGB and Depth sections;
        Classify the material using MCN;
        Assign material label l_i to the object;
    Generate a sparse 3D point cloud map ($\mathcal{P}$) of the environment using VSLAM;
    Divide the point cloud $\mathcal{P}$ into a voxel grid $\mathcal{V}$;
    for *each voxel in* $\mathcal{V}$ **do**
        Find the closest 3D bounding box from OBJ output;
        Propagate the corresponding l_i outputs from MCN to points within the voxel;
    Apply the 3D clustering algorithm (VOXM) on $\mathcal{P}$ to obtain segmented point clouds $\mathcal{PS}$;
    Propagate material labels l_i to the clusters in $\mathcal{PS}$;
    Update semantic map $\mathcal{SM}$ with material l_i (and optionally, object labels) for each point;

the visual appearance in RGB images. By processing RGB and depth images independently, our CA fusion for material classification extracts diverse features, capturing various aspects of the material types and enhancing accuracy by fusing distinctive and correlative features from the two modalities (color and depth). While recent works such as [20] have explored the concept of *late fusion* of classification outcomes from multiple modalities, they might yield some specific patterns in multiple modalities instead of finding shared common modal patterns. In a late fusion, the classifications are obtained from two parallel networks adopted to learn saliency maps from the high-level features of RGB and depth images separately. These are then concatenated to obtain a final prediction map. In contrast, the CA fusion mechanism encourages the determination of complementary information from the different modalities at different abstraction levels.

To excavate the complementarity of two modalities and maintain the discriminability of cross-modal features, we use a Complementarity-aware Fusion Network (CAFN). We first model the distinctive features from two modalities, then select complementary information of two modality features in spatial dimension with two symmetry gates. Finally, an element-wise weighting mechanism is conducted to fuse them to capture more discriminative cross-modal features. The fused features retain not only information existing in both modalities but also modality-specific information. This reduces fusion ambiguity and increases fusion efficiency. In principle, CAFN can be extended to include other modalities as well (e.g., depth-aligned LIDAR data).

In our implementation, CAFN includes two symmetric backbones for RGB and depth feature extraction and five cascaded fusion modules. We use ResNet-101 as unimodal symmetric backbones similar to [21]. We remove the average pooling and the fully connected layers of the backbone. The last two stages are modified with dilated convolution to maintain feature resolution for more spatial information. Then we use hierarchical features from RGB and depth branches respectively, i.e., $\{F^i_{RGB} \mid \text{i=1,2,3,4,5}\}$ and $\{F^i_{Depth} \mid \text{i=1,2,3,4,5}\}$ with five cascade layers. Two unimodal features $F^i_{RGB} \in R^{C_i \times H_i \times W_i}$ and $F^i_{Depth} \in R^{C_i \times H_i \times W_i}$ extracted from corresponding backbones are sent to fusion modules, where C_i, H_i and W_i refer to the channel, height, and width number of the i^{th} layer respectively. As a result, CAFN can select complementary information from two modalities and then fuse enhanced unimodal features for accurate cross-modal features with the help of multiple cascaded layers.

Let $\mathbf{I}_{RGB}$ and $\mathbf{I}_{Depth}$ be the RGB and Depth input images, respectively, and f_{RGB} and f_{Depth} be their corresponding CNN feature maps which can be obtained by element-wise multiplication with their unimodal features:

$$f_{RGB} = \mathbf{I}_{RGB} \odot F_{RGB},\ f_{Depth} = \mathbf{I}_{Depth} \odot F_{Depth},$$

herein, $\odot$ is element-wise multiplication. Further, at each level, the feature maps are fused as $f_{fuse} = f_{RGB} \odot f_{Depth}$.

Next, a CA attention mechanism is used to highlight the complementary regions of the two feature maps:

$$\alpha_{RGB} = \sigma(\mathbf{W}_{RGB} * f_{RGB}),\ \alpha_{Depth} = \sigma(\mathbf{W}_{Depth} * f_{Depth}),$$
$$\alpha_{fuse} = \alpha_{RGB} \odot \alpha_{Depth} \odot \sigma(\mathbf{W}_{fuse} * f_{fuse}),$$
$$\alpha = \frac{\alpha_{fuse}}{\alpha_{RGB} + \alpha_{Depth} - \alpha_{fuse}},$$

where $*$ denotes the convolution operation, $\mathbf{W}_{RGB}$, $\mathbf{W}_{Depth}$ and $\mathbf{W}_{fuse}$ are learnable convolutional filters, σ is the sigmoid activation function, and α_{RGB}, α_{Depth}, α_{fuse} are the attention maps for the RGB, Depth and fused feature maps, respectively, the attention map α is the normalized attention map. α provides a material prediction map for the given scene, which we use to create the bounding box with labels $l_i \in \mathcal{L}$ on the RGB map for the given label set $\mathcal{L}$.

C. Visual SLAM (VSLAM) Pipeline

We build our 3D semantic mapping on top of a visual SLAM solution. We employ the ORB-SLAM2 [3] algorithm for VSLAM due to its real-time performance, scalability, and portability advantages. Furthermore, ORB-SLAM2 is designed to handle monocular, stereo, and RGB-D cameras, and it is based on the ORB (Oriented FAST and Rotated BRIEF) feature descriptors and uses a combination of point features and line features for robust and accurate localization and mapping, producing sparse 3D reconstruction. In our pipeline, the SLAM module uses the RGB-D data to obtain the point cloud map of the environment in real time.

D. Voxel Based Matching (VOXM) and 3D Clustering

The voxel-based point cloud matching component is crucial in our architecture. This module uses the depth information from the RGB-D sensor to estimate the 3D coordinates of the bounding boxes obtained through the OBJ module in the camera coordinate system. Note that both the object detections and the SLAM algorithm's outputs are in the same coordinate system as the RGB-D camera frame. Specifically, this component creates a voxel grid to divide a point cloud into smaller voxels during voxelization. It associates each point in the point cloud with the voxel to which it belongs. Then, it iterates over all the voxels, and for each voxel, it finds the closest bounding box. This step is essential as it enables the efficient processing of the point cloud by reducing the number of points that need to be considered for segmentation and material label propagation.

Let $p_i \in \mathcal{P}$ be the set of points in the point cloud (obtained from VSLAM), and $b_i \in \mathcal{B}$ be the set of 3D bounding boxes obtained from OBJ, and $l_i \in \mathcal{L}$ be the label of classified material obtained through MCN. The Voxel-Based Matching Component aims to find a mapping between b_i, l_i, and the voxels of p_i. Let $\mathcal{V}$ be the voxel grid created by dividing p_i into a 3D grid of small cubes or voxels. Each voxel is assigned a 3D coordinate and contains a set of points from the point cloud that fall within its boundaries.

The mapping between b_i and $\mathcal{V}$ can be represented as a function $\mathcal{M} : b_i \rightarrow \mathcal{V}$, which maps each bounding box in b_i to the voxel in $\mathcal{V}$ that contains the majority of its points. Once the voxel grid is created, point cloud segmentation is applied in such a way that a static color map $\mathcal{L} \rightarrow \mathcal{C}$ is used to color the material associated with each classified material label in the output semantic map $\mathcal{SM}$ to separate it into different cluster points. To achieve this objective, we leverage the multi-scale connected components (MSCC) algorithm [22], which provides an efficient way to propagate material labels in the point cloud. The MSCC algorithm is applied on the point cloud after it has been matched with the 3D bounding boxes obtained from object detection using the voxel-based matching component. The algorithm is applied at multiple scales, starting from a large scale and gradually reducing it to capture more fine-grained details in the point cloud. At each scale, the algorithm applies connected component labeling to group points that are spatially close and similar. These groups of points are then assigned a unique label, and the scale is reduced until the minimum scale is reached.

Let $S = s_1, s_2, \ldots, s_n$ be the set of scales. At each scale, s_i, the MSCC algorithm performs the following steps:

1) Divide $\mathcal{V}$ into a set of larger voxels at scale s_i.
2) Apply connected component labeling to group points that are spatially close and similar to one another within each larger voxel.
3) Merge the resulting clusters across adjacent voxels, considering their spatial proximity and similarity.
4) Assign a unique label to each resulting cluster.

The resulting segmentation at scale s_i can be represented as a function $\text{Seg}(s_i) : \mathcal{V} \rightarrow \mathcal{L}$, which maps each voxel in V to its assigned label in the segmentation at scale s_i. The final segmentation of $\mathcal{V}$ is obtained by merging the segmentations at all scales: $\text{Seg} = \text{merge}(\text{Seg}(s_1), \text{Seg}(s_2), \ldots, \text{Seg}(s_n))$. Where merge is a function that merges the labels of overlapping voxels across scales.

Next, the material labels l_i obtained from CAFN are propagated to each cluster obtained from the MSCC algorithm. This is done by finding the closest bounding box for each cluster and assigning the material label of that bounding box to the cluster. This step is important as it enables the creation of a point cloud map with material labels.

This process can be repeated for all segments in the point cloud, resulting in the accurate propagation of material labels to the corresponding segments in the point cloud. The output of the voxel-based matching component is a point cloud map with material labels, which can be used for various applications such as robot navigation, object recognition, and scene understanding. The final output is a semantic map $\mathcal{SM}$ represented as a 3D point cloud, where each point is associated with semantic (object type) and material labels. This approach provides a detailed and informative perception of the environment, which is crucial for efficient robot navigation and interaction with its physical surroundings.

IV. Experimental Validation

We train and evaluate our approach using publicly available RGB and RGB-D datasets and demonstrate the accuracy of the semantic map through real-world mobile robot experiments in our lab setting. We compare various components of our material classification and clustering pipeline with relevant SOTA methods from the literature.

A. Datasets and Model Training

Given the fact that most object and material classification datasets available are RGB images, we used the benchmark MINC-2500 [23] (RGB) and Flickr Material Database (FMD) [24] (RGB). For RGB and depth network fusion, we used the Washington RGB-D [25] dataset. In our material classification experiments, we grouped the objects together based on their material type. We categorized them into ten material classes: {*Cardboard, Ceramic, Cloth, Glass, Metal, Paper, Plastic, Rubber, Sponge, Wood*}. The material type predictions with max probability less than 0.5 are considered as an additional "*other*" type. Furthermore, for the semantic mapping objective, we used the TUM RGB-D dataset [26], which is a comprehensive collection of 39 real-world indoor sequences categorized into Handheld SLAM, Robot SLAM, and Dynamic Objects. The YOLOv5 network for object detection has been pre-trained on the COCO dataset [18] and can distinguish between 80 different classes of objects. In addition, we created a custom YOLOv5 model built on top of the pre-trained model by adding new object classes, such as board, door, mat, robot, and trash bin, that are not commonly available in existing datasets (but useful in robotics and navigation context [15]). We collected 500 images from the public internet for each of these additional classes, labeled using the publicly-available LabelImg annotation tool.

To verify the effectiveness of our object detection pipeline on these new classes, we compared the "door" and "trash bin" object detections with [15] in two of their datasets (sequence1-Kinect and sequence3-astra), where we found our pipeline providing superior detection accuracies (e.g., ours 85.9 % compared to their's 53.3% for "door" detections).

The experiments are performed on a PC with 3 NVIDIA GeForce GPUs. The learning rate, weight decay, and mini-batch size are set to 1e-5, 0.0005, and 4, respectively. The training procedure used 50 epochs. We evaluate our method on the material classification task using five cross-validation splits. Each split consists of roughly 12,440 training images and 3,920 images for testing. During the test, the task of the model is to assign the correct class label to a previously unseen object instance. During the inference step in our pipeline applied to real-time RGB-D data, the profiling of average per-keyframe processing times at various stages is: {`OBJ`: 56 ms, `MCN`: 3.27 ms, `VSLAM`: 6.34 ms, `VOXM`: 92 ms, Final label propagation: 8.26 ms, Total: 165.87 ms}. As expected, point cloud segmentation along with YOLOv5 takes significant time, both of which can be upgraded or improved for strict real-time applications.

B. Material Classification

First, we validate the performance of our `MCN` against GoogLeNet [27] and VGG_CNN_M [28] networks on the benchmark MINC-2500 [23] and FMD [24] datasets, based only on the RGB images. The comparison of the accuracy results for the common materials (available in the specific dataset) can be seen in Table I. Full confusion matrices are not shown due to brevity. Our approach demonstrates robust and competitive performance in material classification, excelling particularly in recognizing cloth, plastic, and wood across different RGB datasets.

TABLE I: Comparative analysis of the material classification accuracy (%) on the RGB datasets.

Material	**MINC-2500 Dataset** [23]			**FMD Dataset** [24]		
	GoogLeNet [27]	**VGG** [28]	**Ours**	**GoogLeNet** [27]	**VGG** [28]	**Ours**
Ceramic	**89.49**	88.7	87.1	N/A	N/A	N/A
Cloth	81.39	79.85	**82.6**	71.07	68	**84.35**
Glass	84.97	84.76	**86.3**	94.03	**96.54**	89.2
Metal	86.51	84.76	**87.23**	86.3	**92.28**	88.4
Paper	**92.87**	90.15	74.26	**90.57**	82.69	77.3
Plastic	71.04	61.39	**77.82**	87.25	87.01	**89.46**
Wood	**92.52**	86.47	86.38	86.87	85.35	**88.52**

Next, we present the accuracy of our material classification network on the Washington RGB-D dataset [25], with comparisons to the late fusion scheme [20] as well as schemes that use single-modality (RGB or Depth) inputs. Results in Table II show that our multi-modal CA fusion network outperforms the late fusion scheme with up to 12% improvement in accuracy over all material classes. The late fusion scheme fails to utilize the depth data effectively and, in some cases, performs more poorly than the RGB-only models. In contrast, the CA fusion effectively fused diverse and contrastive features from both the depth and RGB images and thus provides superior accuracy in all classes. Also, we can observe the advantages of adding the depth modality, which provided consistent improvements in the classification accuracy by our CA fusion method on the RGB-D data compared to the RGB-only modality.

TABLE II: Accuracy of object-oriented material classification methods on the Washington RGB-D dataset [25].

Material	**RGB Only**	**Depth Only**	**RGB + D Fusion**	
			Late Fusion [20]	**Ours**
Cardboard	77.40%	73.80%	73.10%	**83.40%**
Ceramic	87.30%	84.10%	90.80%	**95.20%**
Cloth	82.10%	79.80%	82.50%	**87.80%**
Glass	86.30%	83.70%	86.20%	**94.20%**
Metal	87.30%	85.30%	89.10%	**93.80%**
Paper	69.90%	68.30%	64.20%	**76.10%**
Plastic	90.00%	87.30%	91.20%	**95.60%**
Rubber	86.30%	85.40%	87.20%	**94.00%**
Sponge	84.80%	82.50%	86.50%	**92.70%**
Wood	84.30%	82.60%	83.50%	**88.70%**

C. Point Cloud Segmentation and 3D Clustering

We used the TUM RGB-D dataset [26] to evaluate the performance of our 3D clustering of material labels with the SLAM integration. Multiple objects are successfully detected and represented as semantic labels on the map with their corresponding material type. We were able to estimate the size of objects accurately using their point clouds, which gives an idea of their general dimensions. As a preliminary evaluation, we qualitatively compare our work to the closest relevant work [14] that performs 3D object segmentation on the point clouds from the SLAM algorithm to refine the localization accuracy. The corresponding examples are shown in Fig. 3. As we see, the segmentation in [14] detects only a few objects and naively clusters them as 3D spheres without consistency checks. In contrast, our pipeline detects various objects and estimates their material type to achieve a smoother and finer clustering consistency across the voxels.

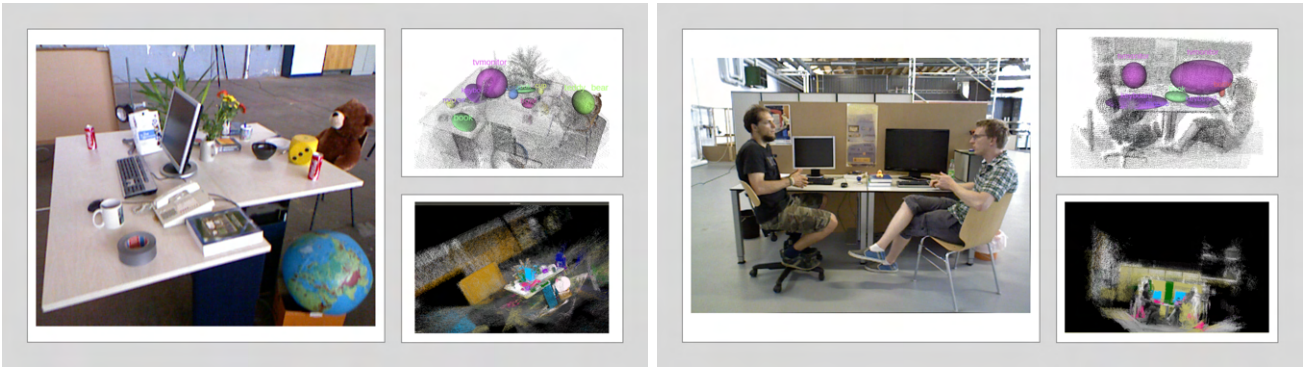
To formally evaluate, we compare our object-oriented clustering results against [16], where a similar semantic object-level 3D clustering is proposed (without material classification). To allow this comparison, we disregarded material classification and applied the 3D clustering based on the object semantics. Table III shows the performance metrics in terms of mean average precision (mAP) (for object/material detection accuracies), Intersection over Union (IoU) (for clustering accuracy), and the number of object detections (for object-oriented effectiveness) on two different TUM dataset sequences. Our model has outperformed [16] in all the metrics due to the robust multiscale clustering.

TABLE III: Results of 3D clustering compared with [16].

TUM RGB-D dataset [26]	**#Objects**	**3D Object Segmentation [16]**			**Ours**		
		#Detections	**IoU**	**mAP**	**#Detections**	**IoU**	**mAP**
fr2_desk	10	581	0.567	0.671	**598**	**0.776**	**0.734**
fr3_sitting_xyz	5	356	0.615	0.652	**367**	**0.765**	**0.718**

D. Real-world Mobile Robot Experiments

We used a mobile robot platform (4WD Rover Zero V3) equipped with a D435i Intel Realsense RGB-D camera, an RPLidar-v3, and an NVIDIA Jetson Nano. We conducted teleoperated mobile robot trajectories in multiple rooms (consisting of diverse scenes and different object compositions) in controlled indoor conditions and recorded them as ROSbag datasets. The experiments were run on an Intel Core i7



(a) fr2_desk sequence. (b) fr3_sitting_xyz sequence.

Fig. 3: RGB Image, point cloud, and semantic material map comparison with [14] (bottom colored point cloud) on two different sequences and their corresponding color representations in the TUM RGB-D dataset [26]: (a) fr2_desk sequence, (b) fr3_sitting_xyz sequence. The colored labels of the respective material classes are shown in Fig. 1.

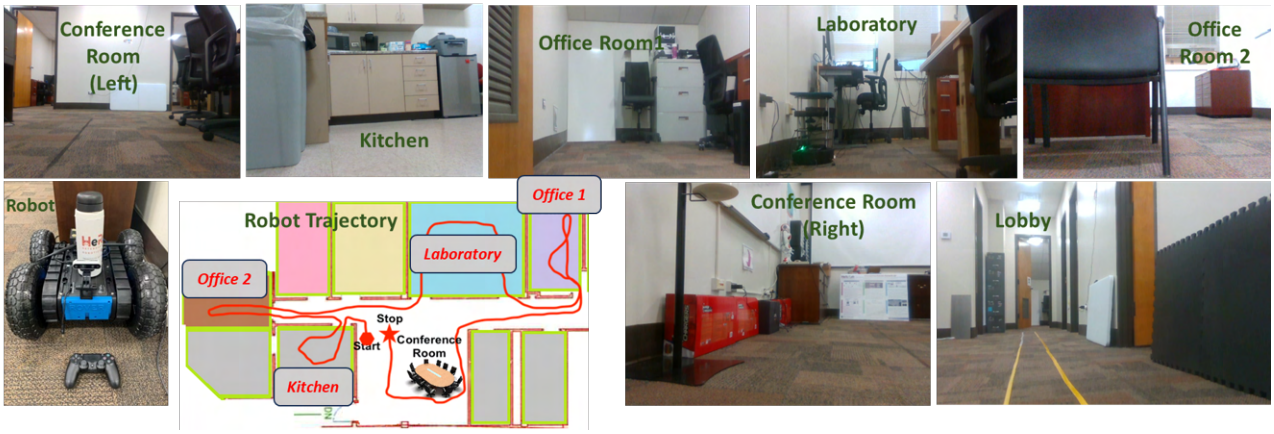


Fig. 4: Robot setup for the real-world mobile robot RGB-D experiments and dataset collection.

laptop with Nvidia GeForce GTX 1050 Ti running on Ubuntu 18.04 with ROS-Melodic. The goal was to obtain enhanced 3D maps with object material information that does not change over time, such as doors, desks, chairs, and other objects. We intentionally prepared the settings such that the datasets consist of multiple objects made of similar materials (e.g., cabinets, doors, desks made of wood) and the same object types made of different materials (e.g., wooden, plastic, and metal chairs). A manually labeled ground truth map specified the static objects' location. Each sequence included raw data from multiple sources, including two RGB-D cameras, LiDAR, and odometry. Fig. 4 shows the robot setup, experiment trajectory, and visuals for each room.

We organized the results based on the material types. Fig. 5 showcases example outputs from multiple stages in our object-oriented pipeline, and Table IV shows the comprehensive accuracy results in different rooms along with the number of objects representing each material type in each room as per the ground truth. As we can observe, the proposed approach has effectively clustered the material properties in the map with acceptable classification and 3D localization accuracies (approx. 0.8 IoU and 0.65 mAP on average). This is an acceptable and significant result (e.g., the IoU reported in [13] for 3D reconstruction is 0.39 for an industrial scenario). The attached video provides a demonstration and in-depth information on these results.

V. Conclusion

We proposed an object-oriented pipeline for 3D semantic mapping of surface material properties in a mobile robotics environment. The proposed approach significantly advances the SOTA in mapping and perception by integrating semantic objects and material identification into a cohesive and effective mapping system. The resulting semantic map associates each point in a 3D point cloud with semantic and material labels. Our extensive experiments, conducted with public and in-house datasets, have demonstrated the robustness and accuracy of our method in creating detailed semantic maps better than comparable approaches. These maps have been validated with qualitative results, showing successful object detection and material classification, which are crucial for robots operating in dynamic and unknown environments. This dual-label (object and material) mapping can prove to be a valuable asset for static landmark identification, facilitating more precise trajectory planning.

References

[1] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 681–687.

TABLE IV: 3D Material Mapping accuracy results for the real-world robot dataset experiments in multiple rooms.

Material Class	Kitchen			Office Room 1			Office Room 2			Laboratory Room			Conference Room		
	#Objects	IoU	mAP	#Objects	IoU	mAP	#Objects	IoU	mAP	#Objects	IoU	mAP	#Objects	IoU	mAP
Cardboard	4	0.674	0.652	5	0.782	0.626	1	0.863	0.678	4	0.768	0.61	8	0.784	0.622
Cloth	-	-	-	-	-	-	-	-	-	1	0.712	0.666	1	0.67	0.643
Glass	-	-	-	1	0.672	0.654	1	0.858	0.676	2	0.715	0.623	-	-	-
Metal	4	0.789	0.613	9	0.816	0.586	3	0.867	0.616	3	0.85	0.631	2	0.732	0.66
Paper	3	0.852	0.632	2	0.84	0.72	2	0.788	0.66	1	0.813	0.668	3	0.88	0.593
Plastic	5	0.849	0.645	-	-	-	-	-	-	2	0.833	0.659	1	0.872	0.646
Rubber	-	-	-	-	-	-	-	-	-	1	0.768	0.638	1	0.776	0.631
Wood	6	0.839	0.666	7	0.778	0.665	5	0.705	0.657	8	0.845	0.659	5	0.831	0.661
Other (e.g., Fiber)	-	-	-	-	-	-	3	0.786	0.622	4	0.816	0.586	4	0.844	0.943
Average	22 (total)	0.801	0.642	24 (total)	0.778	0.650	15 (total)	0.811	0.652	26 (total)	0.791	0.638	25 (total)	0.799	0.675

Fig. 5: Demonstration of sample outputs at various stages in our pipeline for the real-world robot RGB-D datasets. From left: RGB Image, ground truth material labels, segmented objects, material classification, and the final 3D clustering.

[2] E. Latif and R. Parasuraman, "Seal: Simultaneous exploration and localization for multi-robot systems," in *2023 International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 5358–5365.

[3] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[5] J. McCormac, A. Handa, A. Davison, and S. Leutenegger, "Semanticfusion: Dense 3d semantic mapping with convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4628–4635.

[6] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*. Springer, 2014, pp. 345–360.

[7] S. Lee, H. Lim, S. Ahn, and S. Lee, "Ir surface reflectance estimation and material type recognition using two-stream net and kinect camera," in *ACM SIGGRAPH 2019 Posters*, ser. SIGGRAPH '19. New York, NY, USA: Association for Computing Machinery, 2019.

[8] J. Degol, M. Golparvar-Fard, and D. Hoiem, "Geometry-informed material recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[9] C. Zhao, L. Sun, and R. Stolkin, "A fully end-to-end deep learning approach for real-time simultaneous 3d reconstruction and material recognition," *2017 18th International Conference on Advanced Robotics (ICAR)*, pp. 75–82, 2017.

[10] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[11] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for rgb-d salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3051–3060.

[12] G. Schwartz and K. Nishino, "Material recognition from local appearance in global context," *CoRR*, vol. abs/1611.09394, 2016.

[13] C. Zhao, L. Sun, and R. Stolkin, "Simultaneous material segmentation and 3d reconstruction in industrial scenarios," *Frontiers in Robotics and AI*, vol. 7, 2020.

[14] T. Hempel and A. Al-Hamadi, "An online semantic mapping system for extending and enhancing visual slam," *Engineering Applications of Artificial Intelligence*, vol. 111, p. 104830, 2022.

[15] R. Martins, D. Bersan, M. F. M. Campos, and E. R. Nascimento, "Extending maps with semantic and contextual object information for robot navigation: a learning-based framework using visual and depth cues," *Journal of Intelligent & Robotic Systems*, pp. 1–15, 2020.

[16] N. Dengler, T. Zaenker, F. Verdoja, and M. Bennewitz, "Online object-oriented semantic mapping and map updating," in *2021 European Conference on Mobile Robots (ECMR)*, 2021, pp. 1–7.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[19] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[20] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang, "Depth-aware saliency detection using convolutional neural networks," *J. Vis. Comun. Image Represent.*, vol. 61, no. C, p. 1–9, may 2019.

[21] W. Wu, T. Chu, and Q. Liu, "Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation," *Pattern Recognition*, vol. 131, p. 108881, 2022.

[22] J. Huang, L. Xie, W. Wang, X. Li, and R. Guo, "A multi-scale point clouds segmentation method for urban scene classification using region growing based on multi-resolution supervoxels with robust neighborhood," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, pp. 79–86, 2022.

[23] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," *CoRR*, vol. abs/1412.0623, 2014.

[24] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?" *Journal of Vision*, vol. 9, no. 8, pp. 784–784, 2009.

[25] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," *2011 IEEE International Conference on Robotics and Automation*, pp. 1817–1824, 2011.

[26] C. Keimel, A. Redl, and K. Diepold, "The tum high definition video datasets," in *2012 Fourth international workshop on quality of multimedia experience*. IEEE, 2012, pp. 97–102.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[28] G. Kalliatakis, G. Stamatiadis, S. Ehsan, A. Leonardis, J. Gall, A. Sticlaru, and K. D. McDonald-Maier, "Evaluating deep convolutional neural networks for material classification," *arXiv preprint arXiv:1703.04101*, 2017.