
FUSIONVISION: A COMPREHENSIVE APPROACH OF 3D OBJECT RECONSTRUCTION AND SEGMENTATION FROM RGB-D CAMERAS USING YOLO AND FAST SEGMENT ANYTHING

Safouane EL GHAZOUALI*
safouane.elghazouali@toelt.ai
TOELT LLC - Computer Vision
& Machine learning Lab
& Winterthur, Switzerland

Youssef MHIRIT
mm-youssef@protonmail.com
Independent Researcher
Paris, France

Ali OUKHRID
ali.oukhrid@gmail.com
Independent Researcher
Sonceboz, Switzerland

Umberto MICHELUCCI
umberto.michelucci@toelt.ai
TOELT LLC - Computer Vision
& Machine learning Lab
& Winterthur, Switzerland

Hichem NOUIRA
hichem.nouira@lne.fr
LNE - Laboratoire national de
& métrologie et d'essais
Paris, France

ABSTRACT

In the realm of computer vision, the integration of advanced techniques into the pre-processing of RGB-D camera inputs poses a significant challenge, given the inherent complexities arising from diverse environmental conditions and varying object appearances. Therefore, this paper introduces FusionVision, an exhaustive pipeline adapted for the robust 3D segmentation of objects in RGB-D imagery. Traditional computer vision systems face limitations in simultaneously capturing precise object boundaries and achieving high-precision object detection on depth map as they are mainly proposed for RGB cameras. To address this challenge, FusionVision adopts an integrated approach by merging state-of-the-art object detection techniques, with advanced instance segmentation methods. The integration of these components enables a holistic (unified analysis of information obtained from both color *RGB* and depth *D* channels) interpretation of RGB-D data, facilitating the extraction of comprehensive and accurate object information in order to improve post-processes such as object 6D pose estimation, Simultaneous Localization and Mapping (SLAM) operations, accurate 3D dataset extraction, etc. The proposed FusionVision pipeline employs YOLO for identifying objects within the RGB image domain. Subsequently, FastSAM, an innovative semantic segmentation model, is applied to delineate object boundaries, yielding refined segmentation masks. The synergy between these components and their integration into 3D scene understanding ensures a cohesive fusion of object detection and segmentation, enhancing overall precision in 3D object segmentation. The code and pre-trained models are publicly available at <https://github.com/safouaneelg/FusionVision/> (accessed on 28 February 2024).

Keywords RGBD · 3D reconstruction · Point-cloud · SAM · 3D object detection · 3D localization

1 Introduction

The significance of point-cloud processing has surged across various domains such as robotics [1, 2], medical field [3, 4], autonomous driving [5, 6], metrology [7, 8, 9], etc. Over the past few years, advancements in vision

sensors have led to remarkable improvements, enabling these sensors to provide real-time 3D measurements of the surroundings while maintaining decent accuracy [10, 11]. Consequently, point-cloud processing forms an essential pivot of numerous application by facilitating robust object detection, segmentation and classification operations.

Within the field of computer vision, two extensively researched pillars stand prominent: object detection and object segmentation. These sub-fields have captivated the research community for the past decades, helping computers understand and interact with visual data [12, 13, 14]. Object detection involves identifying and localizing one or multiple objects in an image or a video stream, often employing advanced deep learning techniques such as Convolutional Neural Networks (CNNs) [15] and Region-based CNNs (R-CNNs) [16]. The pursuit of real-time performance has led to the development of more efficient models such as Single Shot MultiBox Detector (SSD) [17] and You Only Look Once (YOLO) [18], which demonstrated a balanced performance between accuracy and speed. On the other hand, object segmentation goes beyond the detection process allowing delineating the precise boundaries of each identified object [19]. The segmentation process enables a finer understanding of the visual scene and a precise object localization in the given image. In the literature, two segmentation types are differentiated: semantic segmentation assigns a class label to each pixel [20], while instance segmentation distinguishes between individual instances of the same class [21].

One of the most popular object detection models is (YOLO). The latest known version of YOLO is YOLOv8 which is a real-time object detection system that uses a single neural network to predict bounding boxes and class probabilities simultaneously [22, 23]. It is designed to be fast and accurate, making it suitable for applications such as autonomous vehicles and security systems. YOLO works by dividing the input image into a grid of cells, each one predicts a fixed number of bounding boxes, which are then filtered using a defined confidence threshold. The remaining bounding boxes are then resized and repositioned to fit the object they are predicting. The final step is to perform non-maximum suppression [24] on the remaining bounding boxes to remove overlapping predictions. The loss function used by YOLO is a combination of two terms: the localization loss and the confidence loss. The localization loss measures the difference between the predicted bounding box coordinates and the ground truth coordinates, while the confidence loss measures the difference between the predicted class probability and the ground truth class.

SAM [25], on the other hand, is a recent popular deep learning model for image segmentation tasks. It is based on the U-Net architecture commonly selected for medical applications [26, 27, 28]. U-Net is a CNN that is specifically designed for image segmentation, it consists of an encoder and a decoder, which are connected by a skip connection [29]. The encoder is responsible for extracting features from the input image, while the decoder handles the generation of the segmentation mask. The skip connection allows the model to use the features learned by the encoder at different levels of abstraction, which helps in generating more accurate segmentation masks. SAM gained its popularity because it achieves state-of-the-art performance on various image segmentation benchmarks and many fields such as medical [30], and additional known dataset such as the PASCAL VOC 2012 [31]. It is particularly effective in segmenting complex objects, such as buildings, roads, and vehicles, which are common in urban environments. The model's ability to generalize across different datasets and tasks has highly contributed to its popularity.

The use of YOLO and SAM is still extensively studied and field-applied by the scientific community for 2D computer vision tasks [32, 33, 34]. However, in this paper, we focus the study on the involvement possibility of both state-of-the-art algorithms on RGB-D images. RGB-D cameras are depth sensing cameras that capture both RGB-channel (Red, Green, Blue) and D-map (depth information) of a scene (example shown in Figure 1). These cameras use infrared (IR) projectors and sensors to measure the distance of objects from the camera, providing an additional depth dimension to the RGB image with sufficient accuracy. For example, according to F. Pan *et al.* [35], an estimated accuracy of 0.61 ± 0.42 mm has been assessed on RGB-D camera for facial scanning. Compared to traditional RGB cameras, RGB-D cameras offer several advantages, including:

- (1) Improved object detection and tracking [36]: The depth information provided by RGB-D cameras allows for more accurate object detection and tracking, even in complex environments with occlusions and varying lighting conditions.
- (2) 3D reconstruction [37, 38]: RGB-D cameras can be used to create 3D models of objects and environments, enabling applications such as augmented reality (AR) and virtual reality (VR).
- (3) Human-computer interaction [39, 40]: The depth information provided by RGB-D cameras can be used to detect and track human movements, allowing more natural and intuitive human-computer interaction.

RGB-D cameras have a wide range of applications, including robotics, computer vision, gaming, and healthcare. In robotics, RGB-D cameras are used for object manipulation [41], navigation [42], and mapping [43]. In computer vision, they are used for 3D reconstruction [37], object recognition and tracking [44, 45]. All those algorithms take advantage of the depth information to work with 3D data instead of images. The point-cloud processing allows additional accuracy for the object tracking leading to improved knowledge about its position, orientation, and dimensions in 3D space. This offers distinct advantages compared to traditional image-based systems. Furthermore, RGB-D technologies are also able to surpass diverse lighting conditions [46] due to the use of IR lighting.

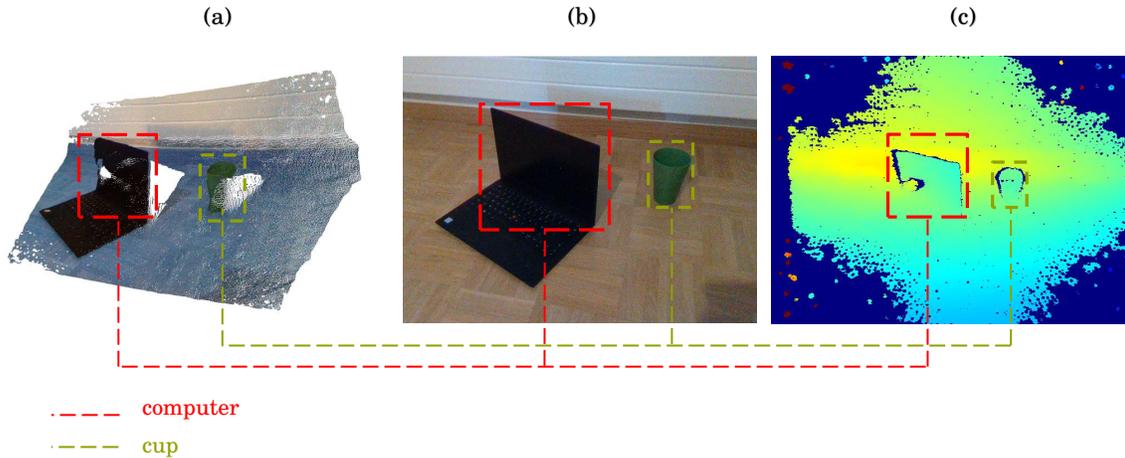


Figure 1: Example of RGB-D camera scene capturing and 3D reconstruction: (a) scene 3D reconstruction from RGB-D depth-channel. (b) RGB stream capture from RGB sensor. (c) Visual estimation of depth with the ColorMap JET (the closer objects are represented in green and far ones are the dark blue regions)

This paper presents a contribution in the fields of *RGB-D* and *object detection and segmentation*. The primary contribution lies in the development and application of FusionVision, a method that links models originally proposed for 2D images, with RGB-D types of data. Specifically, two known models have been implemented, validated and adjusted to work with RGB-D data through the use of both the Depth and RGB channels of an Intel Realsense camera. This combination has led to an enhancement in understanding scenes resulting in 3D object isolation and reconstructions without distortions or noises. Moreover, point-cloud post-processing techniques, including denoising and downsampling, have been integrated to remove anomalies and distortions caused by reflectivity or inaccurate depth measurements, as to improve the real-time performance of the proposed FusionVision pipeline.

The rest of the paper is organized as follows: Despite the uniqueness of the proposed pipeline and the scarcity of methods similar to the one proposed in this paper, few related works are discussed in Section 2. A detailed and comprehensive description of the FusionVision pipeline is given in Section 3 where the processes are discussed step-by-step. Following this, the implementation of the framework and results are presented and discussed in Section 4. Finally, the paper finds are summarized in Section 5.

2 Related work

The aforementioned YOLO and SAM models have been mainly proposed for 2D computer vision operations, lacking the adaptability for RGB-D images. The 3D detection and segmentation of the objects is therefore beyond their capabilities leading to a need for 3D object detection methods. Within this context, few methods have been studied for 3D object detection and segmentation from RGB-D Cameras. Tan Z. *et al.* [47] proposed an improved YOLO (version 3) for 3D object localization. The method aims to achieve real-time high-accuracy 3D object detection from point-clouds using a single RGB-D camera. The authors propose a network system that combines both 2D and 3D object detection algorithms to improve real-time object detection results and increase speed. The used combination of two state-of-the-art object detection methods are: [48] performing

object detection from RGB sensor, and Frustum PointNet [49], a real-time method that uses frustum constraints to predict a 3D bounding box of an object. The method framework could be summarized as follows (Figure 2):

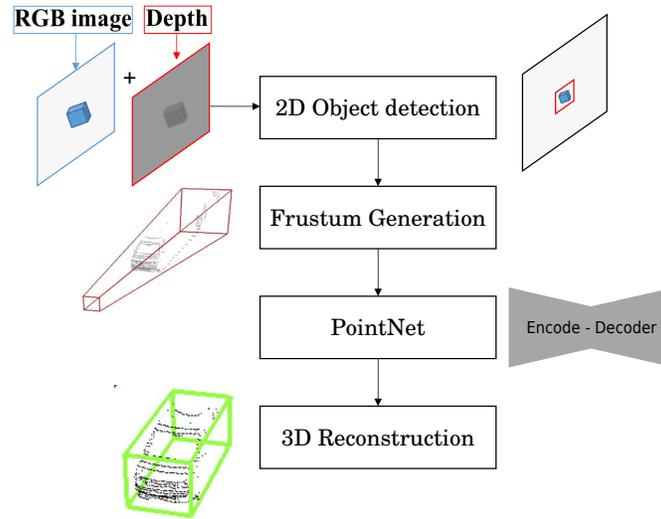


Figure 2: Complex YOLO framework for 3D object reconstruction and localization [47]

- (1) The system starts by obtaining 3D point-clouds from a single RGB-D camera along with the RGB stream.
- (2) The 2D object detection algorithm is used to detect and localize objects in the RGB images. This provides useful prior information about the objects, including their position, width, and height.
- (3) The information from the 2D object detection is then used to generate 3D frustums. A frustum is a pyramid-shaped volume that represents the possible location of an object in 3D space based on its 2D bounding box.
- (4) The generated frustums are fed into the PointNets algorithm, which performs instance segmentation and predicts the 3D bounding box of each object within the frustum.

By combining the results from both the 2D and 3D object detection algorithms, the system achieves real-time object detection performance, both indoors and outdoors. For the method evaluation, the author stated achieving real-time 3D object detection using an Intel realsense D435i RGB-D camera with the algorithm running on a GTX 1080 ti GPU-based system. However this proposed method has limitations and is subject to noise usually due to bad estimation of depth and object reflectivity.

3 FusionVision Pipeline

The implemented FusionVision pipeline could be summarized in six steps in addition to the first step of data acquisition (Figure 3):

- (1) **Data acquisition & Annotation:** This initial phase involves obtaining images suitable for training the object detection model. This image collection can include single- or multi-class scenarios. As part of preparing the acquired data, splitting into separate subsets designated for training and testing purposes is required. If the object of interest is within the 80 classes of Microsoft COCO (Common Objects in Context) dataset [50], this step may be optional, allowing the utilization of existing pre-trained models. Otherwise, if the special object is to be detected, or object shape is uncommon or different from the ones in the datasets, this step is required.
- (2) **YOLO model training:** Following data acquisition, the YOLO model undergoes training to enhance its ability to detect specific objects. This process involves optimizing the model's parameters based on the acquired dataset.

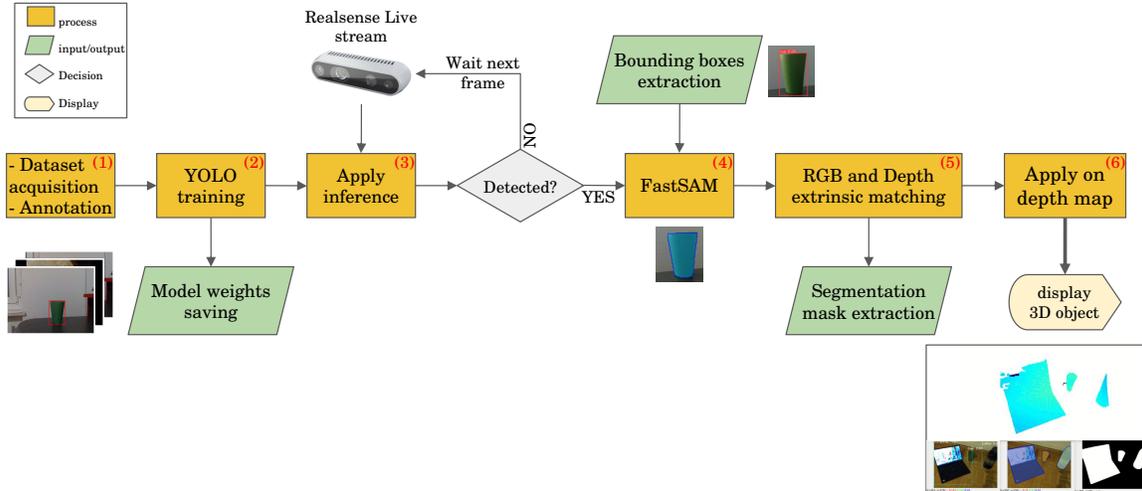


Figure 3: Proposed Pipeline for Real-Time 3D Object Segmentation Using Fused YOLO and FastSAM Applied on RGB-D Sensor.

- (3) **Apply model inference:** Upon successful training, the YOLO model is deployed on the live stream of the RGB sensor from the RGB-D camera to detect objects in real-time. This step involves applying the trained model to identify objects within the camera’s field of view.
- (4) **FastSAM application:** If any object is detected in the RGB stream, the estimated bounding boxes serve as input for the FastSAM algorithm, facilitating the extraction of object masks. This step refines the object segmentation process by leveraging FastSAM’s capabilities.
- (5) **RGB and Depth matching:** The estimated mask generated from the RGB sensor is aligned with the depth map of the RGB-D camera. This alignment is achieved through the utilization of known intrinsic and extrinsic matrices, enhancing the accuracy of subsequent 3D object localization.
- (6) **Application of 3D reconstruction from depth map:** Leveraging the aligned mask and depth information, a 3D point-cloud is generated to facilitate the real-time localization and reconstruction of the detected object in three dimensions. This final step results in an isolated representation of the object in the 3D space.

3.1 Data Acquisition

For applications requiring the detection of specific objects, the data acquisition consists of collecting a number of images using the camera of the specific object at different angles, positions and varying lighting conditions. The images need to be annotated afterwards with the bounding boxes corresponding to the location of the object within the image. Several annotators could be used for this step such as Roboflow [51] LabelImg [52] or VGG Image Annotator [53].

3.2 YOLO training

Training the YOLO model for robust object detection forms a strong backbone of FusionVision pipeline. The acquired data is split into 80% for training and 20% for validation. To further enhance the model’s generalization capabilities, data augmentation techniques were employed by horizontally and vertically flipping images, as well as applying slight angle tilts [54].

In the context of object detection using YOLO, several key loss functions are used in training the model to accurately localize and classify objects within an image. The Objectness Loss (O_L), defined by the Eq. (1), employs binary cross-entropy to assess the model’s ability to predict the presence or absence of an object in a given grid cell, where y_i represents the ground truth objectness label for a given grid cell in the image. The Classification Loss (CLS_L), as outlined in Eq. (2), utilizes cross-entropy to penalize errors in predicting the class labels of detected objects across all classes (C the class number). To refine the localization

accuracy, the Bounding Box Loss ($Bbox_L$), described in Eq. (3), leverages mean squared error to measure the disparity between predicted \hat{y}_i and ground truth y_i bounding box coordinates. Where c_x, c_y refer to the center coordinates of the bounding box and w, h are its width and height. Additionally, the Center Coordinates Loss (C_L), detailed in Eq. (4), incorporates focal loss, including parameters α and γ , to address the imbalance in predicting the center coordinates of objects. These loss functions collectively guide the optimization process during training, steering the YOLOv8 model towards robust and precise object detection performance across diverse scenarios.

$$O_L = -(y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)) \quad (1)$$

$$CLS_L = - \sum_{c=1}^C y_{i,c} \cdot \log(\hat{y}_{i,c}) \quad (2)$$

$$Bbox_L = \sum_{p \in \{c_x, c_y, w, h\}} (y_{i,p} - \hat{y}_{i,p})^2 \quad (3)$$

$$C_L = -\alpha \cdot (1 - \hat{y}_{i,center})^\gamma \cdot y_{i,center} \cdot \log(\hat{y}_{i,center}) \quad (4)$$

Throughout the training process, images and their corresponding annotations are fed into the YOLO network [22]. The network, in turn, generates predictions for bounding boxes, class probabilities, and confidence scores. These predictions are then compared to the ground-truth data using the aforementioned loss functions. This iterative process progressively improves the model’s object detection accuracy until reaching a minimal value of total loss.

3.3 FastSAM deployment

Once the YOLO model is trained, its bounding boxes serve as input for the subsequent step involving the FastSAM model. When processing the complete image, FastSAM estimate instance segmentation mask for all the viewed objects. Therefore, instead of processing the entire image, the YOLO estimated bounding box are used as input information to focus the attention on the relevant region where the object is, significantly reducing computational overhead. Its Transformer-based architecture then delves into this cropped image patch to generate a pixel-wise mask.

3.4 RGB and Depth matching

RGB-D imaging devices typically incorporate an RGB sensor, responsible for capturing traditional 2D color images, and a depth sensor integrating left and right cameras alongside an infrared (IR) projector positioned in the middle. The project IR patterns onto the physical object are distorted by its shape, then get captured by the left and right cameras. Afterwards, the disparity information between corresponding points in the two images is used to estimate the depth of each pixel in the scene. The extracted segments resulting as an output of the FastSAM are represented through binary masks in the RGB channel of the cameras. The identification of the physical object in the DS is carried out by aligning both binary masks and depth frames (Figure 4).

Within this alignment process, the transformation between the coordinate systems of the RGB camera and the depth sensor needs to be estimated either using the calibration process or based on the default factory values. Few calibration techniques can be used for the improvement of the matrices estimation such as [55, 56] This transformation is represented mathematically in Eq. (5).

$$Z_0 \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} = K_c T_{cd} K_d^{-1} \begin{bmatrix} Z \\ u \\ v \\ 1 \end{bmatrix} \quad (5)$$

Where:

- Z_0, u_0, v_0 represent the depth value and pixel coordinates in the aligned depth image,
- Z, u, v is the depth value and pixel coordinates in the original depth image

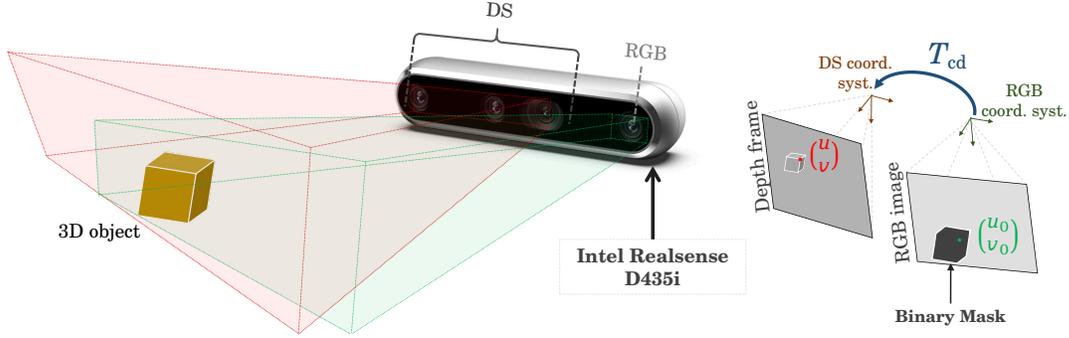


Figure 4: Visual representation of RGB camera alignment with the depth sensor

- K_c is the RGB camera intrinsic matrices
- K_d is the DS intrinsic matrices
- T_{cd} represents the rigid transformation between RGB and DS.

3.5 3D Reconstruction of the physical Object

Once FastSAM mask is aligned with the depth map, the identified physical objects could be reconstructed in 3D coordinates, taking into account only the region of interest (ROI). This process involves several key steps, including: (1) downsampling, (2) denoising, and (3) generating the 3D bounding boxes for each identified object in the point-cloud.

The downsampling process is applied to the original point-cloud data allows the reduction of computational complexity while retaining essential object information. The selected downsampling technique involves voxelization, where the point-cloud is divided into regular voxel grids, and only one point per voxel is retained [57]. Following downsampling, a denoising procedure based on statistical outliers removal [58] is implemented to enhance the quality of the generated point-cloud. Outliers, which may arise from sensor noise are identified and removed from the point-cloud. Finally, for each physical object detected in the aligned FastSAM mask, a 3D bounding box is generated within the denoised point-cloud. The bounding box generation involves creating a set of lines connecting the min and max coordinates along each axis. This set of lines is aligned with the object's position in the denoised point-cloud. The resulting bounding box provides a spatial representation of the detected object in 3D.

4 Results and discussion

4.1 Setup configuration

For the experimental study, the proposed framework is tested on the detection of 3 commonly used physical objects: cup, computer and bottle. The setup configuration that has been used is summarized in Table 1.

Table 1: Setup configuration for realtime FusionVision pipeline

Name	Version	Description
Linux	22.04 LTS	Operating system
Python	3.10	Baseline programming language
Camera	D435i	Intel RealSense RGB-D camera
GPU	RTX 2080 TI	GPU for data parallelization
OpenCV	3.10	Open source Framework for computer vision operations
CUDA	11.2	Platform for GPU based processing

4.2 Data acquisition and annotation

For the data acquisition step a total of 100 images featuring common objects, namely a cup, computer, and bottle, were captured using the RGB stream of a RealSense camera. The recorded images include several poses of the selected 3D physical objects and lighting conditions, as to ensure robust and comprehensive dataset for the model training. The images were annotated using the Roboflow annotator for the YOLO object detection model. Additionally, data augmentation techniques were then applied to enrich the dataset, involving horizontal and vertical flipping, as well as angle tilting (Figure 5).

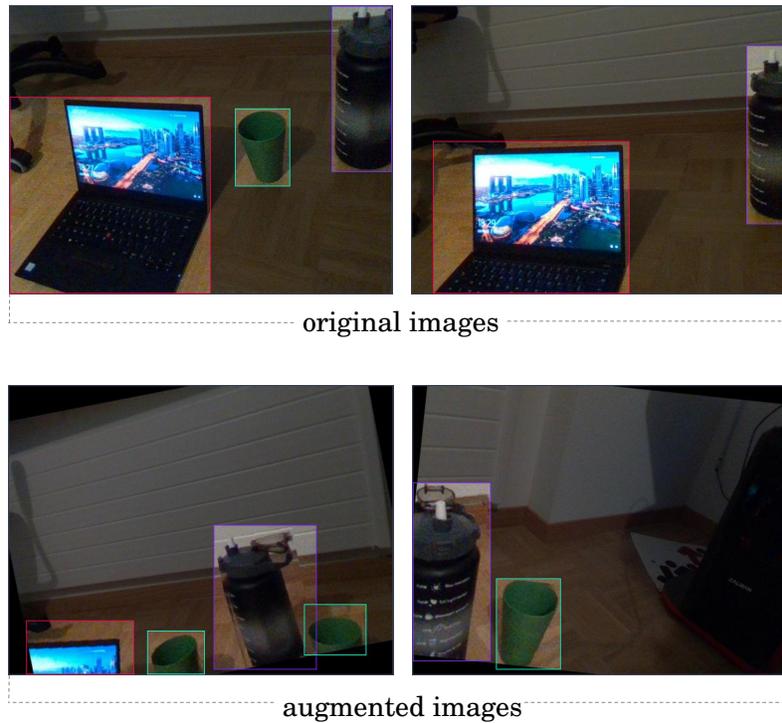


Figure 5: Example of acquired images for YOLO training: the top two images are original, the bottom ones are augmented images

4.3 YOLO training and FastSAM deployment

4.3.1 Quantitative analysis

A comprehensive evaluation of the trained object detection YOLO model has been conducted to assess the robustness and generalization capabilities across diverse environmental conditions. The evaluation process involves three distinct sets of images. Each set contains between 20 and 30 images, designed to represent different scenarios encountered in real-world deployment. (1) The first set of images comprises similar environmental and lighting conditions to those used during model training. These images serve as a baseline for assessing the model's performance in familiar settings and providing insights into its ability to handle variations within its training domain. (2) The second set of images introduces variability in object positions, orientations and lighting conditions compared to the training data. By capturing a broader range of scenarios, this set enables to evaluate the model's adaptability to changes in object positions, orientations and lighting, while simulating real-world challenges such as occlusions and shadows. (3) The third set of images presents a more significant departure from the training data by incorporating entirely different backgrounds, surfaces and lighting conditions. This set aims to test the model's generalization capabilities beyond its training domain, such as to assess its ability to detect objects accurately in novel environments with diverse visual characteristics.

Table 2 presents a comprehensive analysis of the YOLO model’s performance in terms of *Intersection over Union* (IoU) and precision metrics across the different test subsets.

Table 2: Summary of YOLO’s performance in bounding box estimation compared to ground truth annotated 3 test subsets

Metrics	Test sets	cup	bottle	computer	overall
IoU	1	0.96	0.96	0.95	0.95
	2	0.93	0.90	0.91	0.92
	3	0.83	0.52	0.72	0.70
Precision	1	0.99	0.96	0.98	0.98
	2	0.91	0.77	0.85	0.87
	3	0.6	0.31	0.54	0.49

Across these scenarios, the "cup" class consistently demonstrates superior performance, achieving high IoU and Precision scores across all test sets (0.96 and 0.99 for IoU and Precision, respectively). This performance suggests robustness in the model’s ability to accurately localize and classify instances of cups, regardless of environmental factors or object configurations. Conversely, the "bottle" class exhibits the lowest IoU and Precision scores, particularly for test set (3) with respective values of 0.52 and 0.31. It indicates additional challenges in accurately localizing and classifying bottle instances under more complex environmental conditions or object orientations.

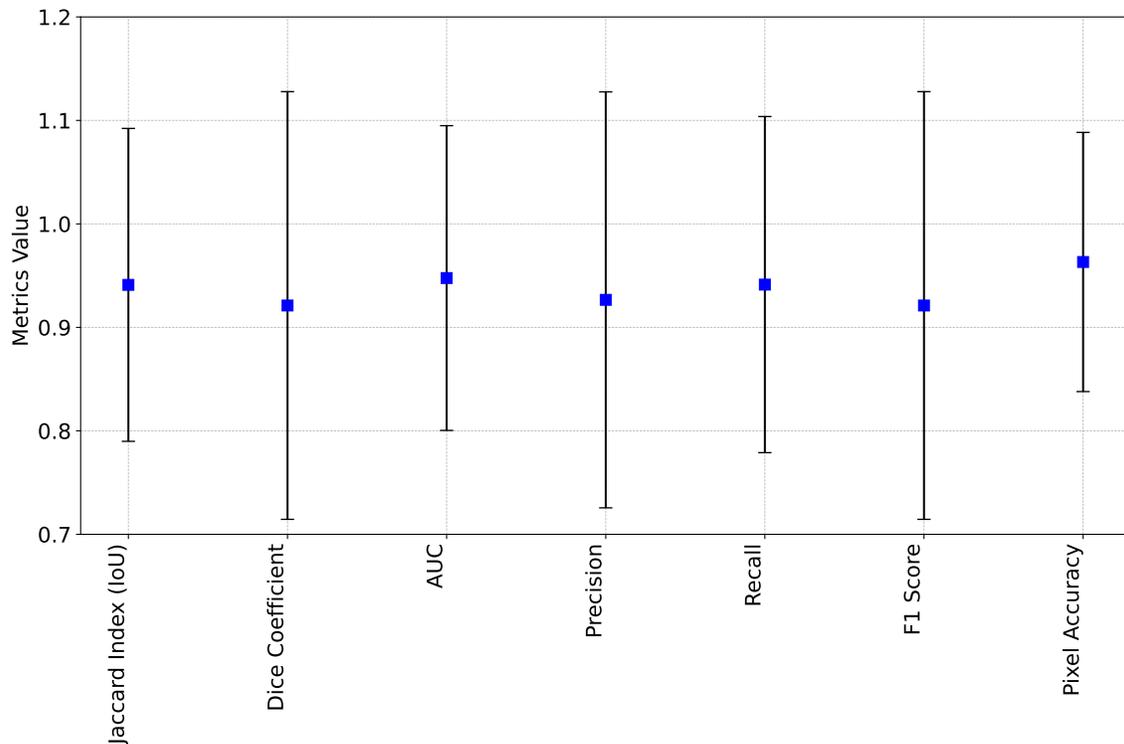


Figure 6: Overall evaluation metrics of FastSAM applied on extracted YOLO bounding boxes and compared to ground truth annotation. The blue points refers to the values of the metrics and black segments are standard deviations.

In addition to YOLO evaluation, FastSAM have been analysed through the annotation of one subset to create a set of ground truth instance segmentation masks. These masks have been overlapped and grouped in a single array, followed by a conversion to binary image, which allow an overall assessment of mask prediction quality. Afterwards, FastSAM has been applied to predict the objects masks by considering the predicted

YOLO bounding boxes as inputs. The resulting mask is also converted to binary image then compared to the ground truth one. The evaluation of segmentation algorithms involves assessing various metrics to gauge their performance. The Jaccard Index (also known as Intersection over Union) and Dice Coefficient [59] are key measures that evaluate the overlap between the predicted and ground truth masks, with higher values indicating better agreement. Precision quantifies the accuracy of positive predictions, while recall measures the ability to identify all relevant instances of the object [60]. The F1 Score balances precision and recall, offering a single metric that considers both false positives and false negatives. The Area under the ROC curve (AUC) assesses segmentation performance across different threshold settings by plotting the true positive rate against the false positive rate [61]. Pixel-wise Accuracy (PA) provides an overall measure of segmentation accuracy at the pixel level [62].

Upon evaluating a segmentation algorithm, the obtained results are summarized in the Figure 6. The mean metrics demonstrate high values across various evaluation criteria: Jaccard Index (IoU) at 0.94, Dice Coefficient at 0.92, AUC at 0.95, Precision at 0.93, Recall at 0.94, F1 Score at 0.92, and Pixel Accuracy at 0.96. However, considering the standard deviation of the metrics helps in understanding the variability in the results. Despite generally favorable mean metrics, standard deviations shows some variability across evaluations (ranging from 0.12 for Pixel-wise Accuracy to 0.20) and indicates areas for potential improvement or optimization in the algorithm. Upon evaluating a segmentation algorithm, the obtained results are summarized in the Figure 6. Since standard deviation analysis assumes a Gaussian distribution of the data, any disturbance (outliers due to inaccurate FastSAM mask estimation at certain sensor’s poses) can cause a mis-estimation (Example in Figure 7). In such cases, the median absolute deviation values, ranging from 0.0029 to 0.0097, provide further insight into the spread of the data and complement the standard deviation analysis.

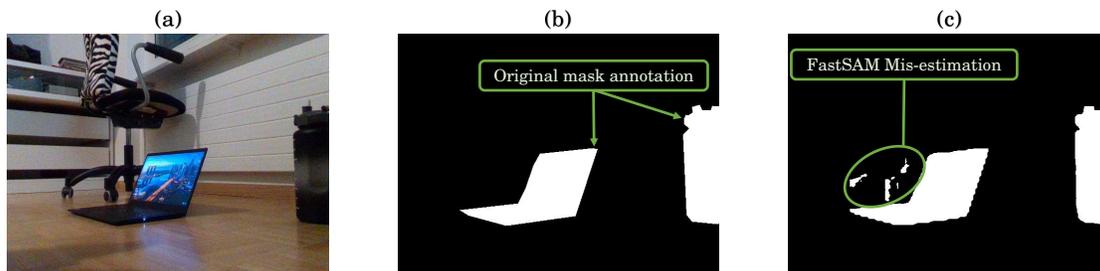


Figure 7: Example of FastSAM mis-estimation of segmentation mask: (a) Original image, (b) Ground truth annotation mask, (c) FastSAM estimated mask.

4.4 3D object reconstruction and discussion

The resulting mask is then aligned with depth frame using the default realsense parameters *rs.align_to* and K matrices [63]. The selected native resolution for both RGB and depth images are 640×480 , which results into approximately 300k 3D points in the full-view reconstructed point-cloud. When applying the FusionVision pipeline, the background has been removed decreasing the number of points to around 32k and focusing the detection on the region of interest only, which leads to more accurate object identification.

Before performing 3D object reconstruction, the point-cloud undergoes downsampling and denoising procedures for enhanced visualization and accuracy. The downsampling is achieved using Open3D’s voxel-downsampling method with a voxel size of 5 units. Subsequently, statistical outlier removal is applied to the downsampled point-cloud with parameters: *neighbors* = 300 and standard deviation *ratio* = 2.0. These processes result in a refined and denoised point-cloud, addressing common issues such as noise and redundant data points. This refined point-cloud serves as the basis for precise 3D object reconstruction. The real-time performance of the YOLO and FastSAM has been approximated to $\frac{1}{30.6 \text{ ms}} \approx 32.68 \text{ fps}$ as the image processing involves three main components: preprocessing (1 ms), running the inference (27.3 ms), and post-processing the results (2.3 ms). When incorporating 3D processing and visualization of the raw, non-processed obtained 3D objects’ point-clouds, the real-time performance decreases to 5 fps. Thus, the need to additional point-cloud post-processing including downsampling and denoising. The results are presented in Figure 8.

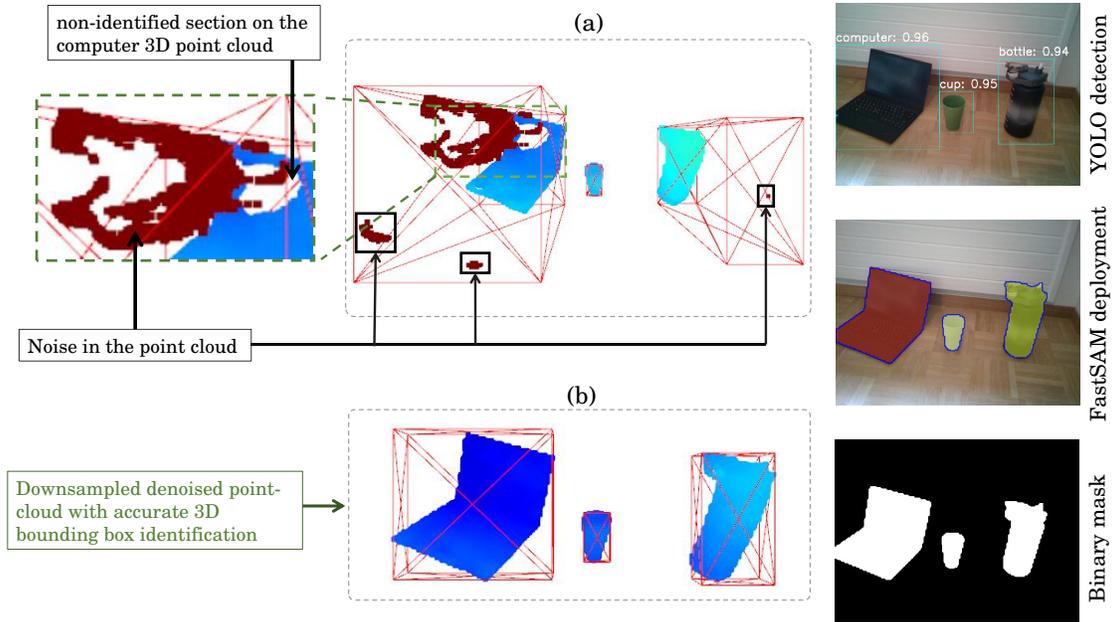


Figure 8: 3D object reconstruction from aligned FastSAM mask: (a) raw point-cloud, (b) post-processing point-cloud by voxel downsampling and statistical denoiser technique. The left images visualizing the YOLO detection, FastSAM mask extraction and Binary mask estimation at specific positions of the physical objects within the frame.

In Figure 8-(a), we can distinguish the presence of noises and wrong depth estimations, mainly due to the object reflectance and inaccurate calculation of disparity. Therefore, the post-processing increases the accuracy of 3D bounding box detection as shown in Figure 8-(b) while maintaining an accurate representation of the 3D object.

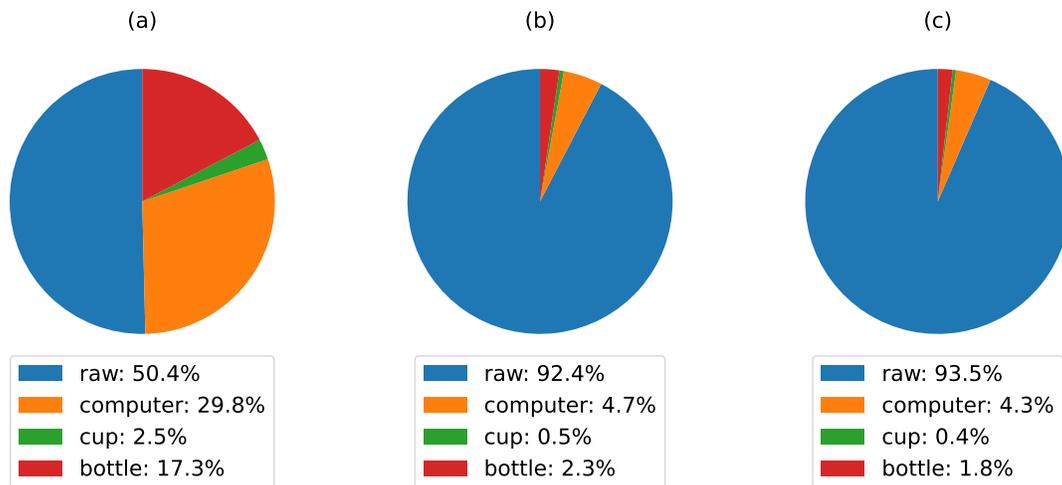


Figure 9: post-processing impact on 3D object reconstruction: (a) raw point-clouds, (b) Downsampled point-clouds, and (c) Downsampled + denoised point-clouds.

The impact of different processing techniques on the distribution of points and object reconstructions derived from a raw point-cloud is illustrated in Figure 9: (a) Raw point-cloud, (b) Downsampled point-cloud, and (c) Downsampled + Denoised point-cloud:

- In 9-(a), the raw point-cloud displays a relatively balanced distribution among different object categories. Notably, the computer and bottle categories contribute significantly, comprising 29.8% and 17.3% of the points, respectively. Meanwhile, the cup and other objects make up smaller proportions. This point-cloud presented several noise and inaccurate 3D estimation.
- In 9-(b), where the raw point-cloud undergoes downsampling with $voxel = 5$ without denoising, a substantial reduction in points assigned to the computer and bottle categories (4.7% and 2.3%, respectively) is observed which improves the real-time performance while maintaining a good estimation of the object 3D structure.
- In 9-(c), the downsampled point-cloud is further subjected to denoising. The distribution remains relatively similar to 9-(b) with a minor decreases in the computer and bottle categories (4.3% and 1.8%, respectively) while eliminating the point-cloud noise for each detected object.

Table 3 summarizes the frame rate evolution when applying the FusionVision Pipeline step by step.

Table 3: Summary of frame rate improvement when applying FusionVision pipeline for 3D objects isolation and reconstruction

Process	Processing time (ms)	Frame rate (fps)	Point-cloud density
Raw point-cloud visualization	~16	up to 60	~302.8 k
RGB + Depth map (Without point-cloud visualization)	~11	up to 90	-
+ YOLO	~31.7	~34	-
+ FastSAM	~29.7	~33.7	-
+ Raw 3D Object visualization	~189	~5	~158.4 k
complete FusionVision Pipeline	~30.6	~27.3	~20.8 k

The fusion of 2D image processing and 3D point-cloud data has led to a significant improvement in object detection and segmentation. By combining these two disparate sources of information, we have been able to eliminate over 85% of combined non-interesting and the noisy point-cloud, resulting in a highly accurate and focused representation of the objects within the scene. This allows the enhancement of scene understanding and enables reliable localization of individual objects, which can then be used as input for 6D object pose identification, 3D tracking, shape and volume estimation, and 3D object recognition. The accuracy and efficiency of the FusionVision pipeline make it particularly well-suited for real-time applications such as autonomous driving, robotics, and augmented reality.

5 Conclusion

FusionVision stands as a comprehensive approach in the realm of 3D object detection, segmentation, and reconstruction. The outlined FusionVision pipeline, encompasses a multi-step process, involving YOLO-based object detection, FastSAM model execution, and subsequent integration into the three-dimensional space using point-cloud processing techniques. This holistic approach not only amplifies the accuracy of object recognition but also enriches the spatial understanding of the environment. The results obtained through experimentation and evaluation underscore the efficiency of the FusionVision framework. First, the YOLO model has been trained on a custom-created dataset then deployed on real-time RGB frames. FastSAM model has been subsequently applied on the frame while considering the detected objects bounding boxes to estimate their masks. Finally, point-cloud processing techniques have been added to the pipeline to enhance the 3D segmentation and scene understanding. This has led to the elimination of over 85% of unnecessary point-cloud for the 3D reconstruction of specific physical objects. The estimated 3D bounding boxes of the objects defines well the shape of the 3D object in the space. The proposed FusionVision method showcases high real-time performances particularly in indoor scenarios, which could be adopted in several applications including robotics, augmented reality and autonomous navigation. Through the deployment of FusionVision (NVIDIA GPU RTX 2080 Ti with 11GB memory), it allows reaching a real-time performance of about 27.3 fps (frames per second) while accurately reconstructing the objects in 3D from the RGB-D view. Such performance underscores the scalability and versatility of the proposed framework for real-world

deployment. As perspectives, the continuous evolution of FusionVision could involve leveraging the latest zero-shot detectors to enhance its object recognition capabilities. Additionally, the investigation of Language Model (LLM) integration for operation such as prompt-based specific object identification and real-time 3D reconstruction stands as a promising avenue for future enhancements.

Acknowledgments

This work has received funding from the EURAMET programme (22DIT01-ViDiT and 23IND08-DiVision) co-financed by the Participating States and from the European Union’s Horizon 2020 research and innovation program.

References

- [1] Ming Liu. Robotic online path planning on point cloud. *IEEE Transactions on Cybernetics*, 46(5):1217–1228, 2016.
- [2] Zifeng Ding, Yuxuan Sun, Sijin Xu, Yan Pan, Yanhong Peng, and Zebing Mao. Recent advances and perspectives in deep learning techniques for 3d point cloud data processing. *Robotics*, 12(4):100, 2023.
- [3] Damian Krawczyk and Robert Sitnik. Segmentation of 3d point cloud data representing full human body geometry: A review. *Pattern Recognition*, page 109444, 2023.
- [4] Fan Wu, Yumeng Qian, Haozhun Zheng, Yan Zhang, and Xiawu Zheng. A novel neighbor aggregation function for medical point cloud analysis. In *Computer Graphics International Conference*, pages 301–312. Springer, 2023.
- [5] Xing Xie, Haowen Wei, and Yongjie Yang. Real-time lidar point-cloud moving object segmentation for autonomous driving. *Sensors*, 23(1):547, 2023.
- [6] Yan Zhang, Kang Liu, Hong Bao, Ying Zheng, and Yi Yang. Pmpf: Point-cloud multiple-pixel fusion-based 3d object detection for autonomous driving. *Remote Sensing*, 15(6):1580, 2023.
- [7] Giulio D’Emilia, Luciano Chiominto, Antonella Gaspari, Stefano Marsella, Marcello Marzoli, and Emanuela Natale. Extraction of a floor plan from a points cloud: some metrological considerations. *Acta IMEKO*, 12(2):1–9, 2023.
- [8] Zhongyi Michael Zhang, Sofia Catalucci, Adam Thompson, Richard Leach, and Samanta Piano. Applications of data fusion in optical coordinate metrology: a review. *The International Journal of Advanced Manufacturing Technology*, 124(5–6):1341–1356, 2023.
- [9] Cihan Altuntas. Review of scanning and pixel array-based lidar point-cloud measurement techniques to capture 3d shape or motion. *Applied Sciences*, 13(11):6488, 2023.
- [10] Polina Kurtser and Stephanie Lowry. Rgb-d datasets for robotic perception in site-specific agricultural operations—a survey. *Computers and Electronics in Agriculture*, 212:108035, 2023.
- [11] Xinyang Zhao, Qinghua Li, Changhong Wang, Hexuan Dou, and Bo Liu. Robust depth-aided rgbd-inertial odometry for indoor localization. *Measurement*, 209:112487, 2023.
- [12] Mingqi Gao, Feng Zheng, James JQ Yu, Caifeng Shan, Guiguang Ding, and Jungong Han. Deep learning for video object segmentation: a review. *Artificial Intelligence Review*, 56(1):457–531, 2023.
- [13] Bingxin Hou, Ying Liu, Nam Ling, Yongxiong Ren, Lingzhi Liu, et al. A survey of efficient deep learning models for moving object segmentation. *APSIPA Transactions on Signal and Information Processing*, 12(1), 2023.
- [14] Ershat Arkin, Nurbiya Yadikar, Xuebin Xu, Alimjan Aysa, and Kurban Ubul. A survey: object detection methods from cnn to transformer. *Multimedia Tools and Applications*, 82(14):21353–21383, 2023.
- [15] Ravpreet Kaur and Sarbjeet Singh. A comprehensive review of object detection with deep learning. *Digital Signal Processing*, 132:103812, 2023.
- [16] Shet Reshma Prakash and Paras Nath Singh. Object detection through region proposal based techniques. *Materials Today: Proceedings*, 46:3997–4002, 2021. International Conference on Materials, Manufacturing and Mechanical Engineering for Sustainable Developments-2020 (ICMSD 2020).
- [17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

- [18] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.
- [19] Yuanbo Wang, Unaiza Ahsan, Hanyan Li, and Matthew Hagen. A comprehensive review of modern object segmentation approaches. *Foundations and Trends® in Computer Graphics and Vision*, 13(2–3):111–283, 2022.
- [20] Xiaolong Liu, Zhidong Deng, and Yuhan Yang. Recent progress in semantic image segmentation. *Artificial Intelligence Review*, 52(2):1089–1106, June 2018.
- [21] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International Journal of Multimedia Information Retrieval*, 9(3):171–189, July 2020.
- [22] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. 2023.
- [23] Shuang Cong and Yang Zhou. A review of convolutional neural network architectures and their optimizations. *Artificial Intelligence Review*, 56(3):1905–——–1969, 2023.
- [24] Zekun Luo, Zheng Fang, Sixiao Zheng, Yabiao Wang, and Yanwei Fu. Nms-loss: Learning with non-maximum suppression for crowded pedestrian detection. *CoRR*, abs/2106.02426, 2021.
- [25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.
- [26] Jiaqi Shao, Shuwen Chen, Jin Zhou, Huisheng Zhu, Ziyi Wang, and Mackenzie Brown. Application of u-net and optimized clustering in medical image segmentation: A review. *CMES—Computer Modeling in Engineering & Sciences*, 136(3), 2023.
- [27] Shanwen Zhang and Chuanlei Zhang. Modified u-net for plant diseased leaf image segmentation. *Computers and Electronics in Agriculture*, 204:107511, 2023.
- [28] Ehsan Khodapanah Aghdam, Reza Azad, Maral Zarvani, and Dorit Merhof. Attention swin u-net: Cross-contextual attention mechanism for skin lesion segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1——–5. IEEE, 2023.
- [29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.
- [30] Sheng He, Rina Bao, Jingpeng Li, Jeffrey Stout, Atle Bjornerud, P. Ellen Grant, and Yangming Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets, 2023.
- [31] Peng-Tao Jiang and Yuqi Yang. Segment anything is a good pseudo-label generator for weakly supervised semantic segmentation, 2023.
- [32] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124:103540, 2023.
- [33] Lingzhi Xu, Wei Yan, and Jiashu Ji. The research of a novel wog-yolo algorithm for autonomous driving object detection. *Scientific reports*, 13(1):3699, 2023.
- [34] Rizwan Qureshi, MOHAMMED GAMAL RAGAB, SAID JADID ABDULKADER, amgad muneer, ALAWI ALQUSHAIB, EBRAHIM HAMID SUMIEA, and Hitham Alhussian. A comprehensive systematic review of yolo for medical object detection (2018 to 2023). July 2023.
- [35] Fangwei Pan, Jialing Liu, Yueyan Cen, Ye Chen, Ruilie Cai, Zhihe Zhao, Wen Liao, and Jian Wang. Accuracy of rgb-d camera-based and stereophotogrammetric facial scanners: a comparative study. *Journal of Dentistry*, 127:104302, 2022.
- [36] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Ales Leonardis, and Joni-Kristian Kämäräinen. Depthtrack : Unveiling the power of RGBD tracking. *CoRR*, abs/2108.13962, 2021.
- [37] Kyriaki A. Tychola, Ioannis Tsimperidis, and George A. Papakostas. On 3d reconstruction using rgb-d cameras. *Digital*, 2(3):401——–421, 2022.
- [38] Jianwei Li, Wei Gao, Yihong Wu, Yangdong Liu, and Yanfei Shen. High-quality indoor scene 3d reconstruction with rgb-d cameras: A brief review. *Computational Visual Media*, 8(3):369——–393, 2022.

- [39] Cai Linqin, Cui Shuangjie, Xiang Min, Yu Jimin, and Zhang Jianrong. Dynamic hand gesture recognition using rgb-d data for natural human-computer interaction. *Journal of Intelligent & Fuzzy Systems*, 32(5):3495–3507, 2017.
- [40] Wei Gao and Peng Miao. Rgb-d camera assists virtual studio through human computer interaction. In *Institute of Management Science and Industrial Engineering. Proceedings of 2018 3rd International Conference on Materials Science, Machinery and Energy Engineering (MSMEE 2018)*, volume 6. Institute of Management Science and Industrial Engineering: Computer ..., 2018.
- [41] Max Schwarz, Anton Milan, Arul Selvam Periyasamy, and Sven Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5):437–451, 2018.
- [42] Young Hoon Lee and Gérard Medioni. Rgb-d camera based wearable navigation system for the visually impaired. *Computer vision and Image understanding*, 149:3–20, 2016.
- [43] Felix Endres, Jürgen Hess, Jürgen Sturm, Daniel Cremers, and Wolfram Burgard. 3-d mapping with an rgb-d camera. *IEEE transactions on robotics*, 30(1):177–187, 2013.
- [44] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Rgb-d object recognition: Features, algorithms, and a large scale benchmark. *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*, pages 167–192, 2013.
- [45] Johann Prankl, Aitor Aldoma, Alexander Svejda, and Markus Vincze. Rgb-d object modelling for object recognition and tracking. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 96–103. IEEE, 2015.
- [46] Jordi Gené-Mola, Jordi Llorens, Joan R. Rosell-Polo, Eduard Gregorio, Jaume Arnó, Francesc Solanelles, José A. Martínez-Casasnovas, and Alexandre Escolà. Assessing the performance of rgb-d sensors for 3d fruit crop canopy characterization under different operating and lighting conditions. *Sensors*, 20(24), 2020.
- [47] Ya Wang, Shu Xu, and Andreas Zell. Real-time 3d object detection from point clouds using an rgb-d camera. In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM*, pages 407–414. INSTICC, SciTePress, 2020.
- [48] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.
- [49] Charles Ruizhongtai Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from RGB-D data. *CoRR*, abs/1711.08488, 2017.
- [50] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
- [51] B. Dwyer, J. Nelson, J. Solawetz, and et al. Roboflow (version 1.0). [Software], 2022. <https://roboflow.com>.
- [52] Tzutalin. Labelimg. Free Software: MIT License, 2015.
- [53] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). <http://www.robots.ox.ac.uk/vgg/software/via/>, 2016. Version: 2.0.1, Accessed: 08.10.2018.
- [54] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022. International Conference on Intelligent Engineering Approach(ICIEA-2022).
- [55] Safouane El Ghazouali, Alain Vissiere, Louis-Ferdinand Lafon, Mohamed-Lamjed Bouazizi, and Hichem Nouira. Optimised calibration of machine vision system for close range photogrammetry based on machine learning. *Journal of King Saud University - Computer and Information Sciences*, 34(9):7406–7418, 2022.
- [56] V. Paradiso, A. Crivellaro, K. Amgarou, N. Blanc de Lanaute, P. Fua, and E. Liénard. A versatile calibration procedure for portable coded aperture gamma cameras and rgb-d sensors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 886:125–133, 2018.
- [57] Carlos Moreno. A comparative study of filtering methods for point clouds in real-time video streaming.
- [58] Haris Balta, Jasmin Velagic, Walter Bosschaerts, Geert De Cubber, and Bruno Siciliano. Fast statistical outlier removal based method for large 3d point clouds of outdoor environments. *IFAC-PapersOnLine*, 51(22):348–353, 2018. 12th IFAC Symposium on Robot Control SYROCO 2018.

- [59] Jeroen Bertels, Tom Eelbode, Maxim Berman, Dirk Vandermeulen, Frederik Maes, Raf Bisschops, and Matthew B. Blaschko. *Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice*, page 92–100. Springer International Publishing, 2019.
- [60] Rohit Jena, Lukas Zhornyyak, Nehal Doiphode, Pratik Chaudhari, Vivek Buch, James Gee, and Jianbo Shi. Beyond map: Towards better evaluation of instance segmentation, 2023.
- [61] Pablo Gimeno, Victoria Mingote, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. Generalizing auc optimization to multiclass classification for audio segmentation with limited training data. *IEEE Signal Processing Letters*, 28:1135–1139, 2021.
- [62] Juana Valeria Hurtado and Abhinav Valada. Semantic scene segmentation for robotics, 2024.
- [63] Intel Corporation. Intel RealSense SDK 2.0 - Python Documentation. <https://dev.intelrealsense.com/docs/python2>, 2022. Developer Documentation.