ADVANCED TOPICS IN SOFTWARE ENGINEERING

AR- Anchoring, Scanning and World Understanding

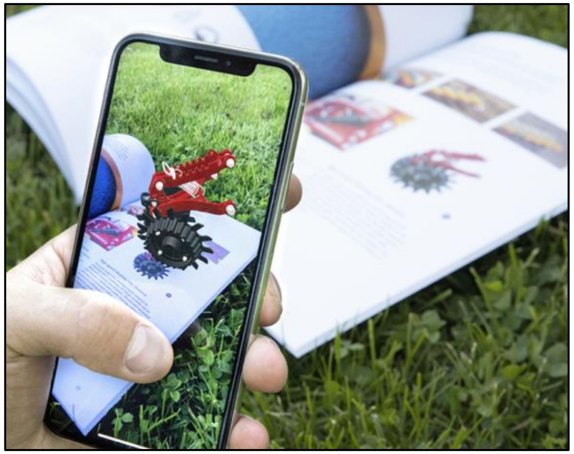
Sonal Lakhotia

Table Of Contents

Table Of Contents

What is Augmented Reality? (AR)

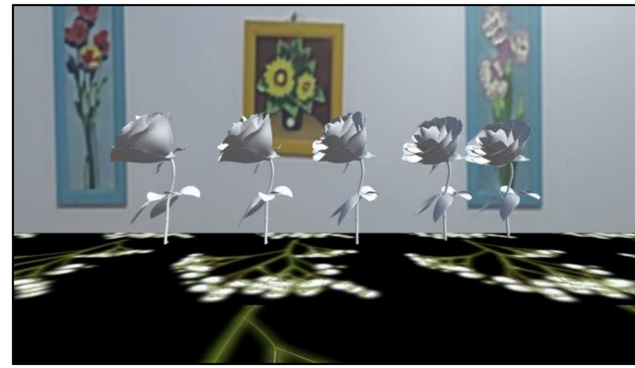
- Combination of :
  - Real scene viewed by a user
  - Virtual entity generated by computer
- Adds additional information and augments the real scene.



Augments the image from the book [1]

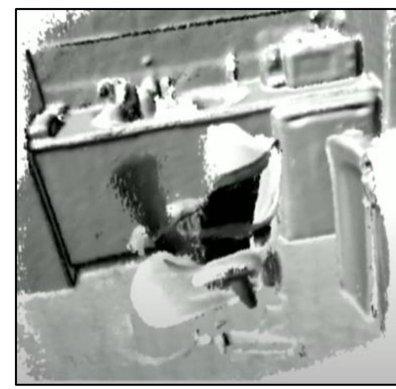
Motivation

- Physically integrated and interactive AR
- Environment-Aware AR
- Better world perception with Simultaneous Localization And Mapping (SLAM)
- Real-time tracking and detection of multiple objects in a scene



Motivation

- Most SLAM systems assume a static scene
- Assumption that the camera moves not objects
- Challenging to track and map objects in dynamic scene
- Inconsistent reconstruction
- Semantic information not available



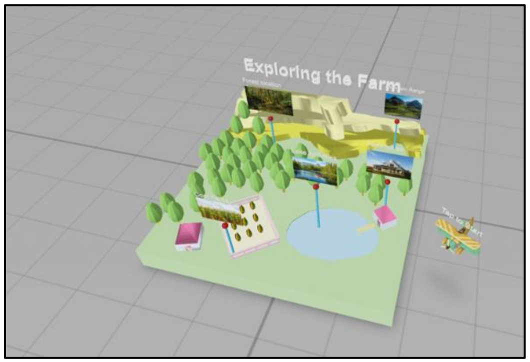
Failed reconstruction of the rotating chair[5]

Table Of Contents

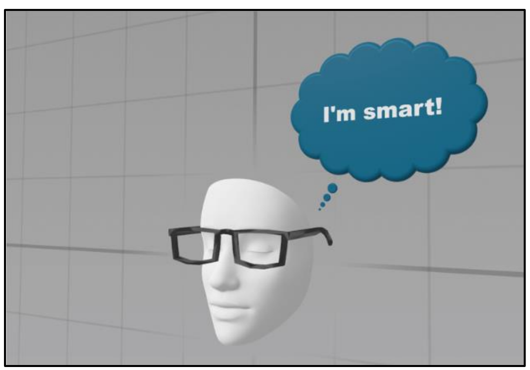
Anchoring and Scanning in AR

- Anchors – objects recognized by AR software
- They help to integrate the real and virtual worlds
- AR scanners initiate an augmented experience with virtual overlays on anchors
- AR platforms (such as ARCore, ARKit, and Vuforia) use sensors for environmental understanding
    - Detecting size and location – the main sensor is the camera
    - Position and orientation of surfaces
    - Real-world lighting conditions
    - Tracking motion

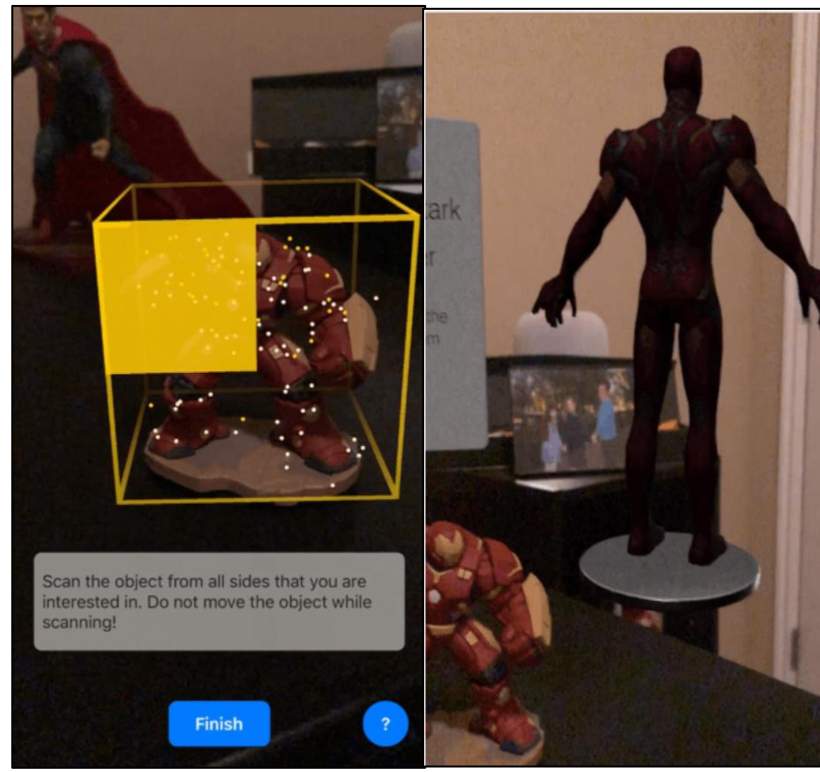
Anchoring and Scanning in AR



Plane Anchor [2]



Face Anchor [2]



3 D Object Anchor [2]

SLAM Augmented Reality

- Immersive AR is possible with SLAM technology
- It locates the device within the environment
- Builds the map of the environment
- Maps unknown environment
- Enables 3D reconstruction

[3]

[4]

Visual SLAM

- Visual SLAM for indoor localization and 3D reconstruction

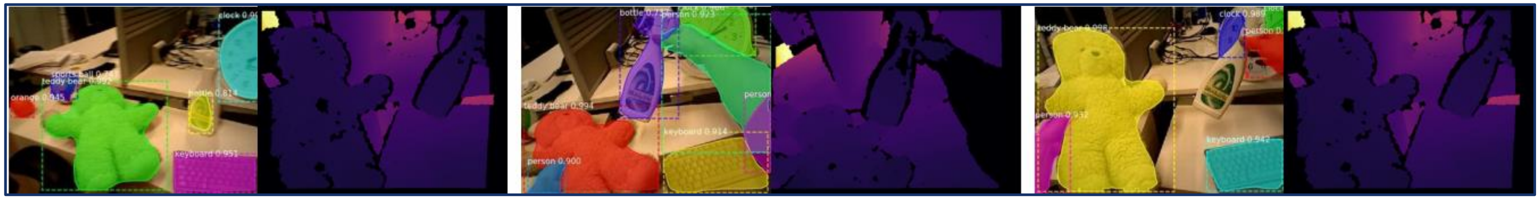
- Camera as a sensor – visual SLAM
  - Collects abundant information
  - Strong object recognition features
  - Senses at higher resolution
  - Lower cost
  - Easy to carry
  - Monocular vision SLAM, binocular vision SLAM, and RGB-D depth camera SLAM

- RGB-D SLAM system acquires colored models and is mostly used for AR

Table Of Contents

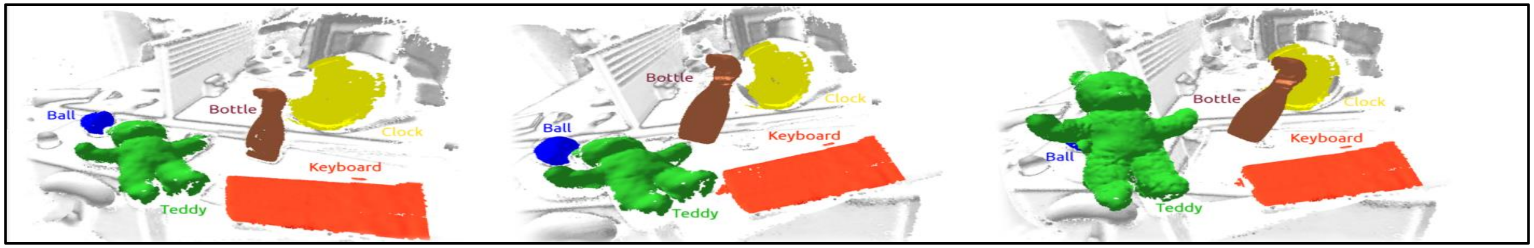
Maskfusion

- Real-time, object-aware, semantic, dynamic RGB-D SLAM system



(a) Frame 400 (b) Frame 700 (c) Frame 900

- Tracks, and reconstructs multiple moving objects along with background



(a) Frame 400 (b) Frame 700 (c) Frame 900

Related Work In Dynamic and Semantic SLAM

Single Non-Rigid Object	**Multiple rigid objects**	**Semantic information for 3D models available in advance**	**Fixed set semantic category (not differentiating object instances)**
Non-Rigid RGBD	Co-Fusion	2.5D is not enough	Convolutional Neural Network(CNN) SLAM
Fusion4D		Slam ++	Semantic-Fusion
Dynamic Fusion			

Comparison with related SLAM systems

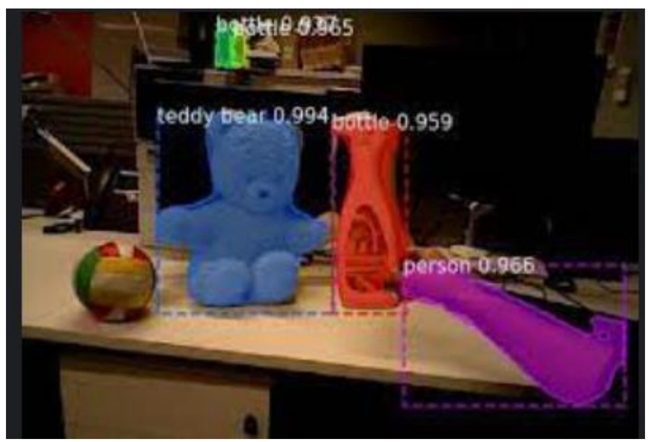
Exisitng related SLAM systems	Important properties considered for the related systems				
	Model-free	Scene Segmentation	Semantics	Multiple moving objects	Non-Rigid
Static-Fusion	✓	✓			
2.5D is not enough		✓	✓		
Slam++		✓	✓		
CNN-SLAM	✓	✓	✓		
Semantic-Fusion	✓	✓	✓		
Non-Rigid RGBD					✓
Dynamic-Fusion	✓				✓
Fusion4D	✓				✓
Co-Fusion	✓	✓		✓	
Mask-Fusion	✓	✓	✓	✓	

Maskfusion – Object-based SLAM

- **Instance–level** semantic segmentation (detects instances of objects and creates semantic object masks)

- Enables real-time object recognition

- Creates object-level representation for the world map

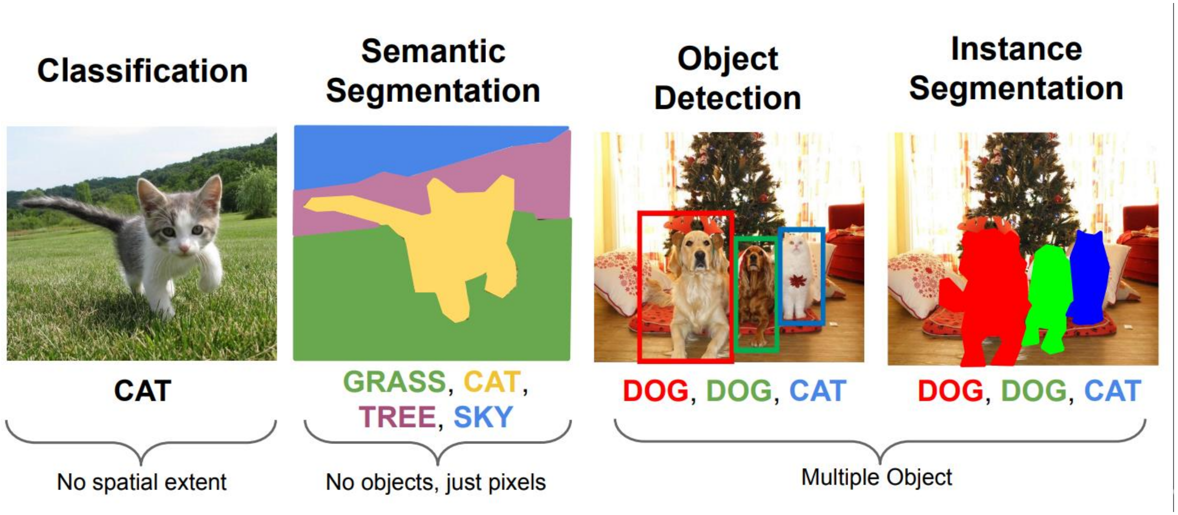


object boundaries, object masks, and semantic labels[5]

Instance level Semantic Segmentation

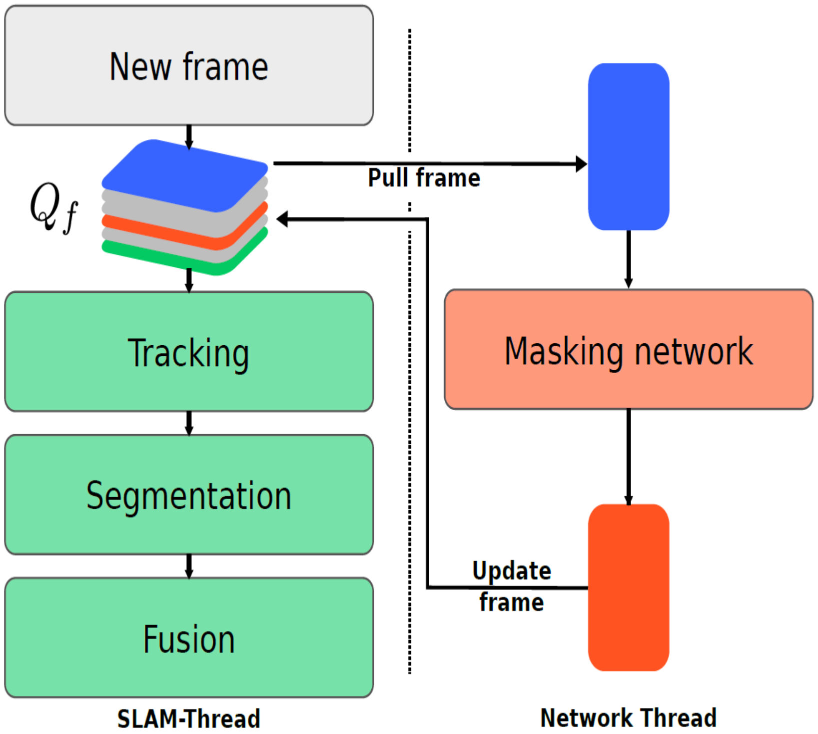
- Each instance is segmented and assigned semantic labels – MaskRCNN (Mask Region-Based Convolutional Neural Network)

- Semantic labels could be used for AR experiments



[6]

MaskFusion Workflow

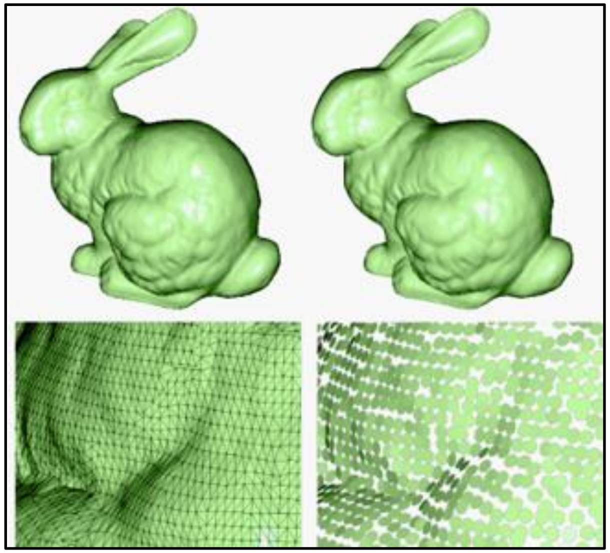


[7]

- MaskRCNN runs parallel to SLAM
- The masking network pulls frames, and updates frames to queue as soon as semantic masks are available
- **Tracking, Segmentation, and Fusion** are performed with each acquired camera frame
- The system runs in real-time

Tracking

- Scene representation using surface element (surfel) map
  - Only the surface of an object is represented
  - An object is represented by a dense set of points (surfel cloud)
  - Each surfel is a disk with data – position, color, radius, timestamps, normal



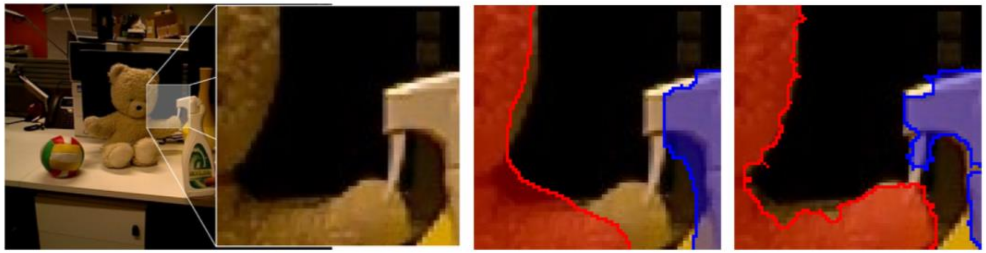
[5]

Tracking

- 3D geometry of each object is represented as a set of surfels
- Tracking jointly optimizes two costs
  - **Photometric cost** – ensuring photometric consistency between the surface objects being looked at and 3D models
  - **Iterative Closest Point (ICP) Geometric cost** - by minimizing the distance between the model and newly reconstructed surfels
  - A rigid transformation is obtained that tells the current position of the camera and each of the independently segmented objects

Segmentation

- Instance-level semantic segmentation on images is not enough
- Boundaries are not precise and bleed into the background
- Depth images are utilized for geometric segmentation – (over-segmentation)
- Geometric + Semantic segmentation gives very precise object boundaries



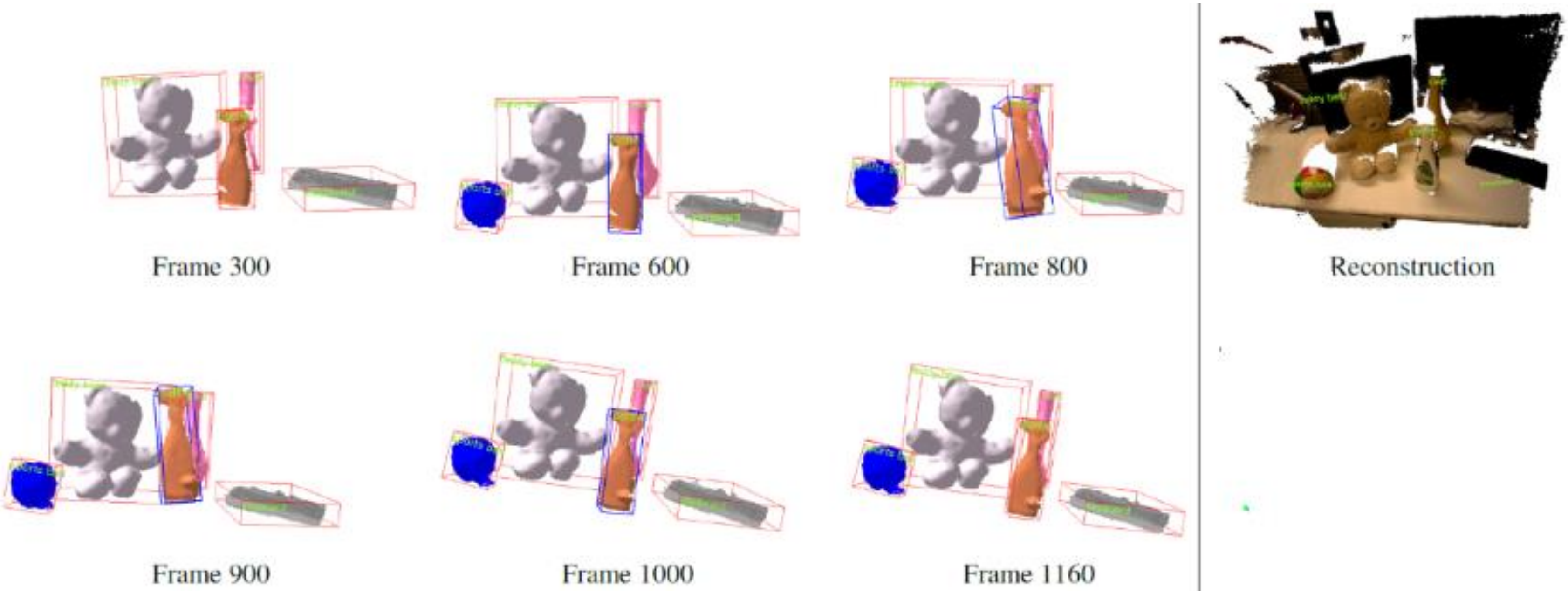
(a) Segment of interest (b) Semantic only (c) With geometric

[7]

Fusion

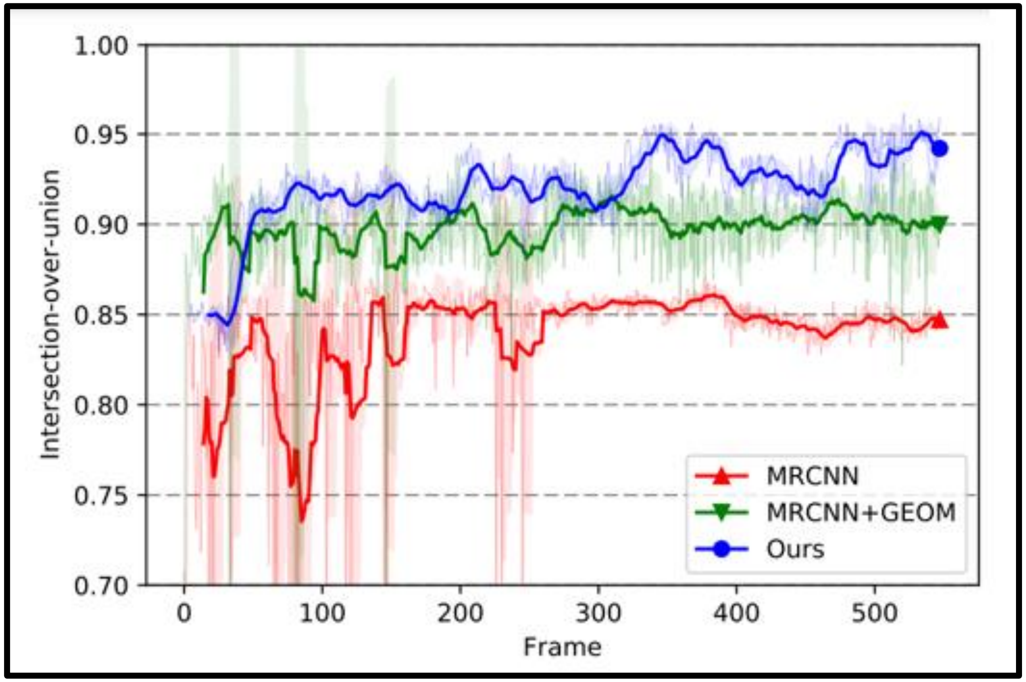
- Individual surfel map for each detected object
- 2D segmentation defines the models to fuse with
- Depth map provides new 3D points with normal
- These are weighted with existing surfels
- Object labels used to associate surfels with the correct model

Recognition, Tracking, and Mapping in MaskFusion



[7]

Quantitative Evaluation

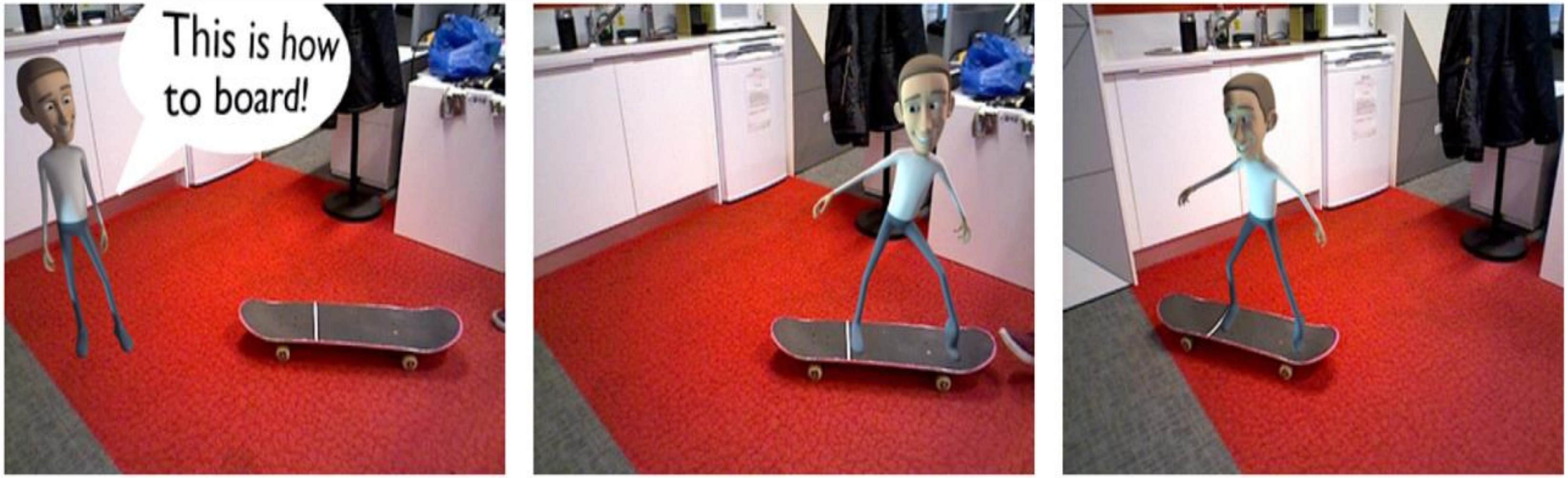


IOU graphs comparing the labeling performance over time [7]

- Red – Mask RCNN
- Green – Mask RCNN + geometric segmentation
- Blue – MaskFusion

MaskFusion maintains temporally consistent 3D models through tracking and fusion and gives better results

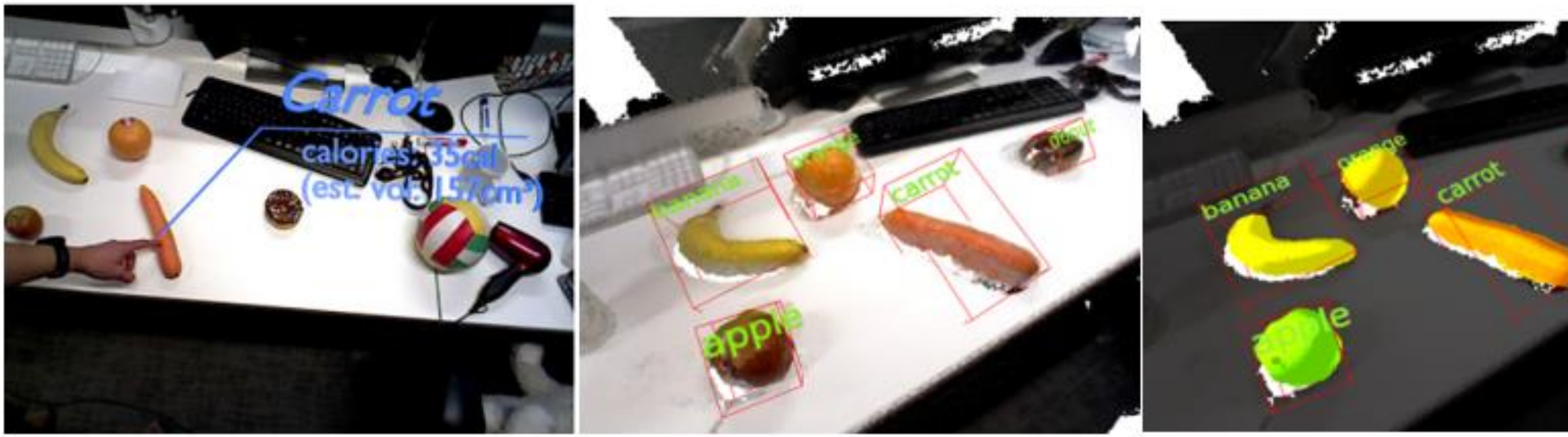
Qualitative Evaluation – AR experiment



(a) Semantic reaction (b) Object interaction (c) Obeying dynamics

An augmented reality experiment [7]

Qualitative Evaluation – Calories Demo



Estimated calories for a carrot | 3D reconstruction | Object labels in 3D

[7]

Limitations

- Limited object recognition, on 80 classes of the MS-COCO dataset
- Miss-classification of object labels not accounted
- Tracking and reconstruction are limited to rigid objects
- Challenging to track small objects with less geometric information

Future Work

- Including more classes for object recognition
- Enabling tracking of small objects with limited features
- Extending the system to track and reconstruct non-rigid objects such as humans in a dynamic environment
- Virtual characters could be aware of the background model along with the object it interacts with

Table Of Contents

Summary

- Interactive, environment-aware AR enables better world understanding
- Static SLAM systems do not achieve this
- MaskFusion used to develop novel AR applications
- AR with MaskFusion - instance-aware, semantic, dynamic
- MaskFusion is limited to object classes on which MaskRCNN is trained
- It can be extended to observe human-object interactions

References for Images

[1] https://www.apptunix.com/blog/augmented-reality-apps-industries-uses-ar/

[2] https://www.learningguild.com/articles/understanding-anchors-in-augmented-reality-experiences/

[3] https://augmentedpixels.com/pokemon-go-make-truly-augmented-reality-game-slam-sdk/

[4] Snapfeet. https://snapfeet.io/en

[5]] https://www.youtube.com/watch?v=GbD3OMOk2DE

[6] http://cs231n.stanford.edu/slides/2020/lecture_12.pdf

[7] M. Runz, M. Buffier, and L. Agapito, 'MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects', in 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Oct. 2018, pp. 10–20. doi: 10.1109/ISMAR.2018.00024.

[

References

- **Primary**: M. Runz, M. Buffier, and L. Agapito, 'MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects', in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2018, pp. 10–20. doi: 10.1109/ISMAR.2018.00024.

- Y.-C. Wu, L. Chan, and W.-C. Lin, 'Tangible and Visible 3D Object Reconstruction in Augmented Reality', in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2019, pp. 26–36. doi: 10.1109/ISMAR.2019.00-30.

- X. Xiang *et al.*, 'Mobile3DScanner: An Online 3D Scanner for High-quality Object Reconstruction with a Mobile Device', *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 11, pp. 4245–4255, Nov. 2021, doi: 10.1109/TVCG.2021.3106491.

- P. Stotko, S. Krumpen, M. Weinmann, and R. Klein, 'Efficient 3D Reconstruction and Streaming for Group-Scale Multi-Client Live Telepresence', in *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2019, pp. 19–25. doi: 10.1109/ISMAR.2019.00018.

- K. Wang, H. Zheng, G. Zhang, and J. Yang, 'Parametric Model Estimation for 3D Clothed Humans from Point Clouds', in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2021, pp. 156–165. doi: 10.1109/ISMAR52148.2021.00030.