Perception and Learning in Robotics & Augmented Reality

Depth-aware Mixed Reality: Capture the AR-Flag

Final Presentation

Students:
Muhammad Faizan
Ihsan Berkan Balaban
Jeremias Neth

Tutors:
Patrick Ruhkamp
Benjamin Busam
Hyunjun Junq

24th July, 2020

Technische Universität München

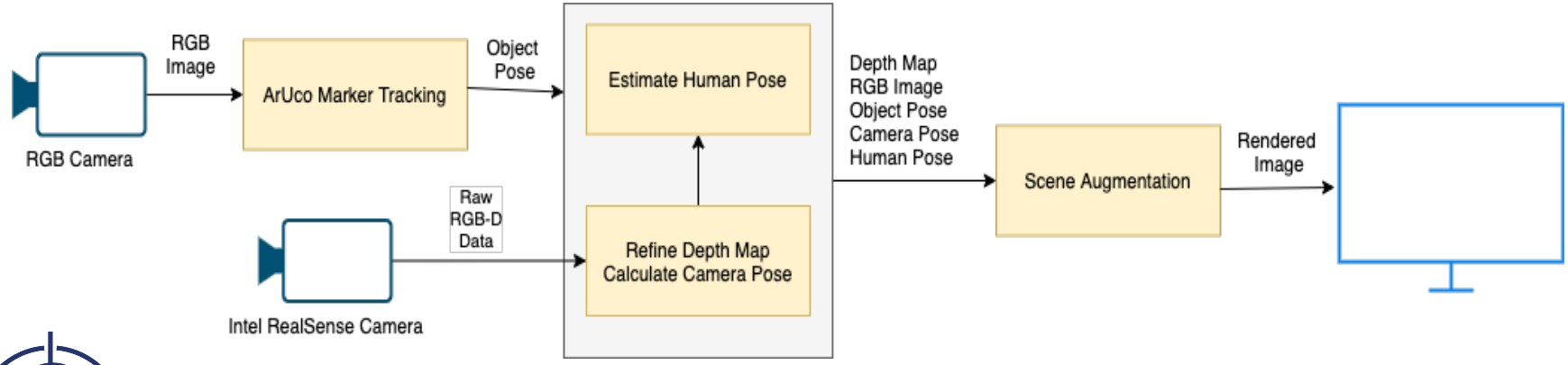
Johns Hopkins Whiting School of Engineering

Project Overview

- **Objective:** Interactively manipulate a scene, whose 3D structure is estimated with an RGB-D camera. Realistic object augmentation and occlusion awareness is explored, and the results are implemented in a simple AR game.

Augmentation

- Four main parts:
  1. Depth Estimation and Camera Pose
  2. User Interaction and Scene Manipulation
  3. Realism and Embedding
  4. Capture the AR Plane



Depth-aware Mixed Reality: Capture the AR-Flag

Depth Estimation and Camera Pose

Technische Universität München

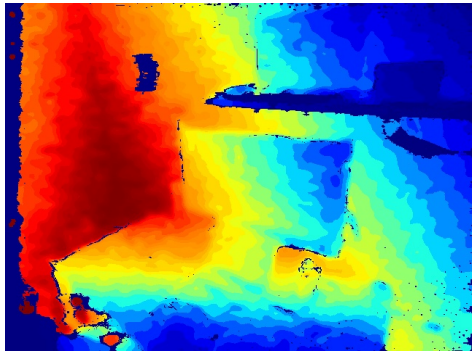
Johns Hopkins Whiting School of Engineering

Depth Estimation

- Intel RealSense Camera.


IR Left Image


IR Right Image

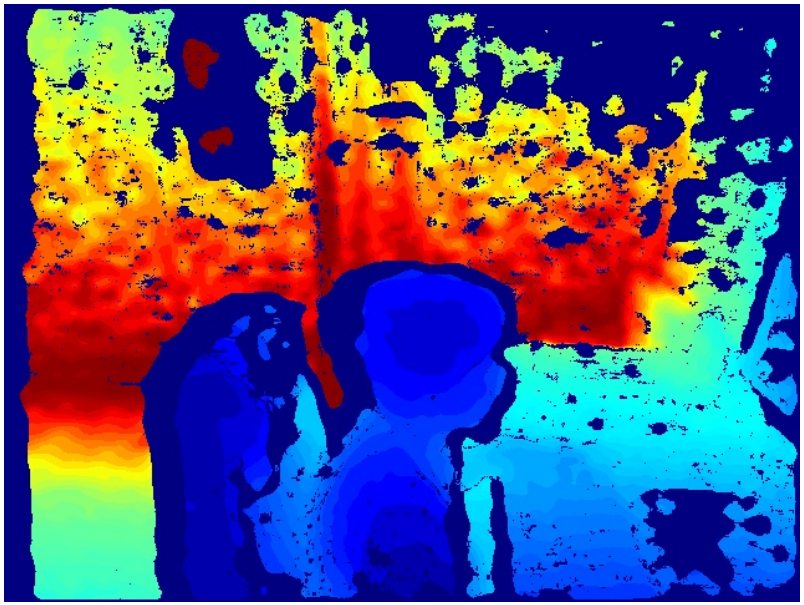

Depth Map

[1] Intel RealSense D435: https://www.intelrealsense.com/depth-camera-d435/, Accessed: 2020-06-24

Depth Estimation

- Raw depth map from RealSense:

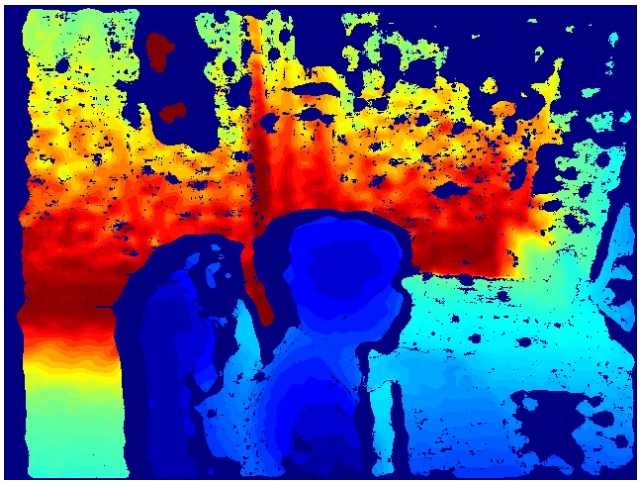
Scene



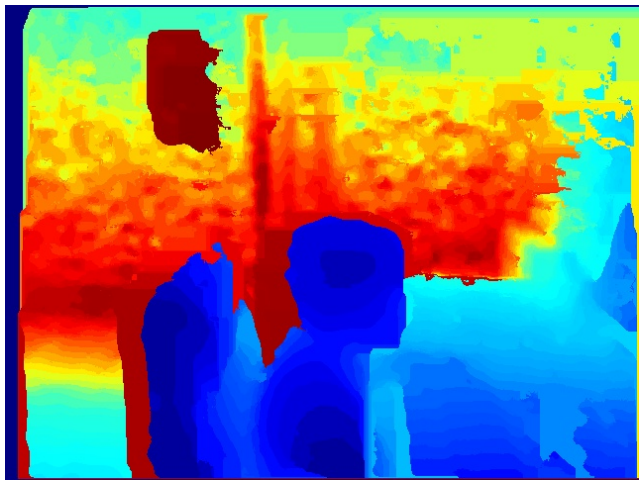
Depth Map

Depth Estimation

- Post-processing on depth map.
- Filters:
  - Spatial Edge-Preserving filter [2]
  - Temporal filter [3]
  - Holes Filling filter [3]



Depth Map (Without filtering)



Depth Map (With filtering)

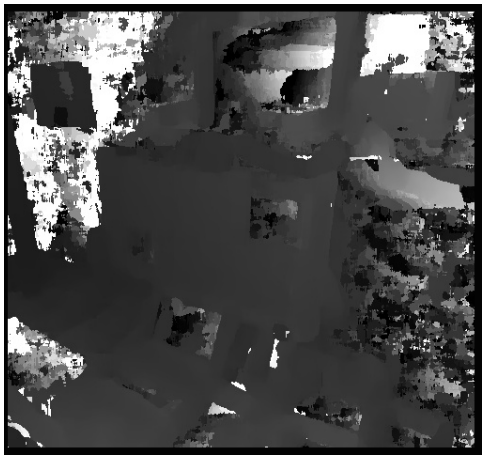
[2] Eduardo S.L. Gastal, Manuel M. Oliveira, Domain Transform for Edge-Aware Image & Video Processing (https://www.inf.ufrgs.br/~eslgastal/DomainTransform/)
[3] Intel RealSense SDK Post-Filters:https://github.com/IntelRealSense/librealsense/blob/master/doc/post-processing-filters.md

Depth Estimation

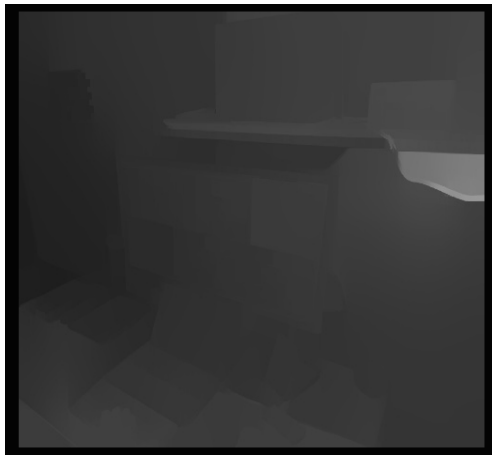
- Stereo matching
- Filters
  - WLS Disparity Filter

IR Left Image

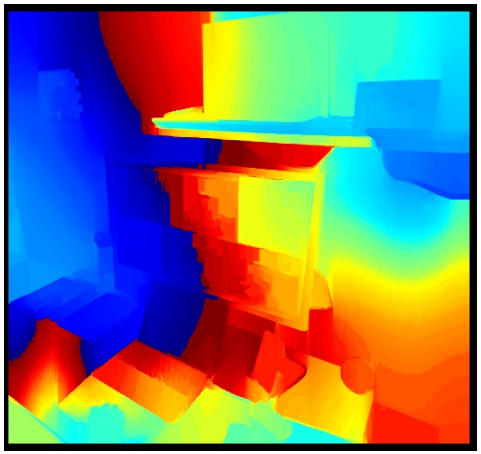
IR Right Image



Disparity Map



Disparity Map (Filtered)



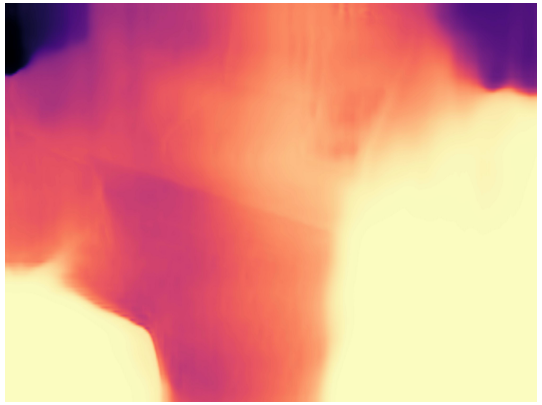
Depth Map

[4] Disparity map post-filtering: https://docs.opencv.org/trunk/d3/d14/tutorial_ximgproc_disparity_filtering.html

Depth Estimation using Deep NNs

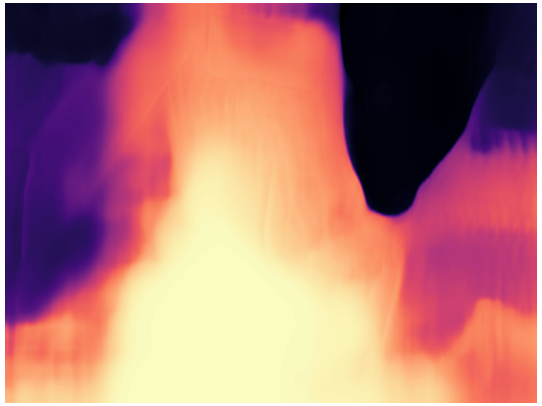
- Monodepth2:

Indoor scene



Depth Map

Indoor scene

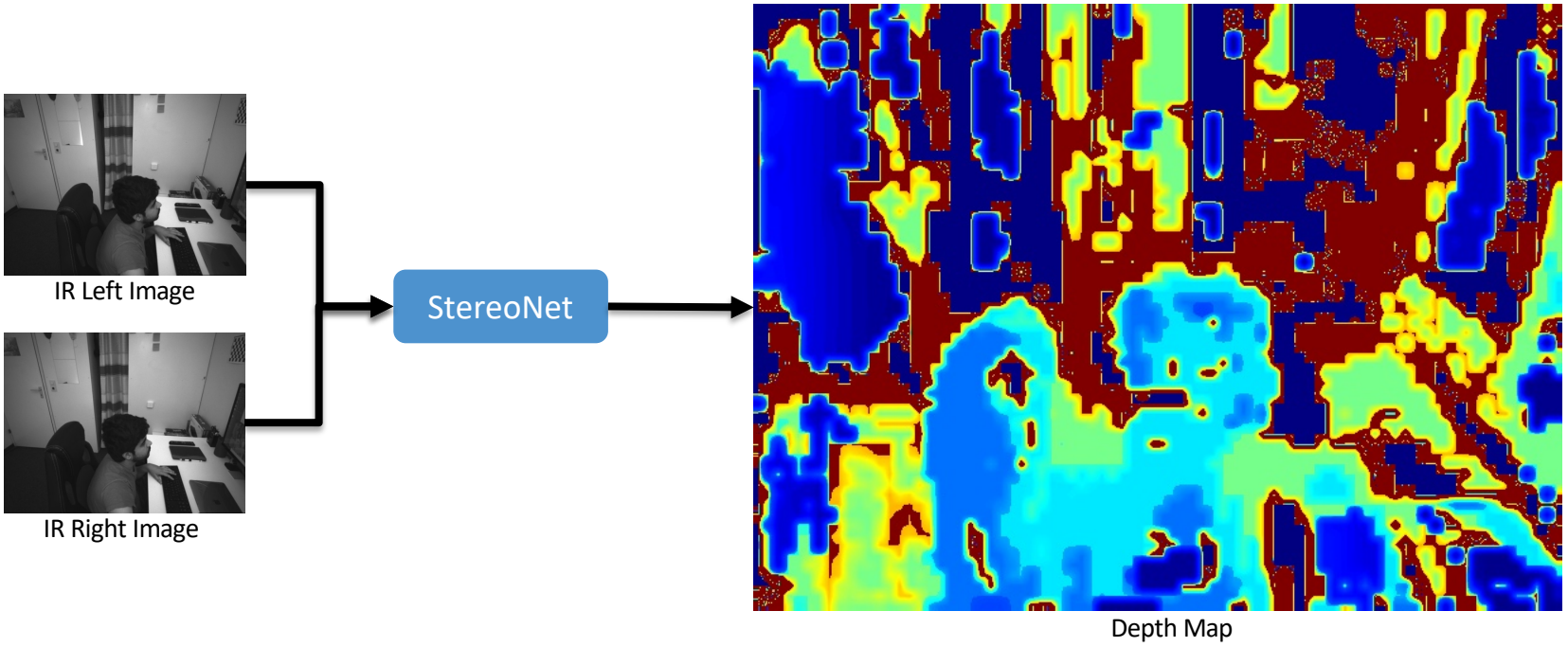


Depth Map

[5] Clément Godard, Oisin Mac Aodha, Michael Firman and Gabriel Brostow, Digging Into Self-Supervised Monocular Depth Estimation, ICCV 2019

Depth Estimation using Deep NNs

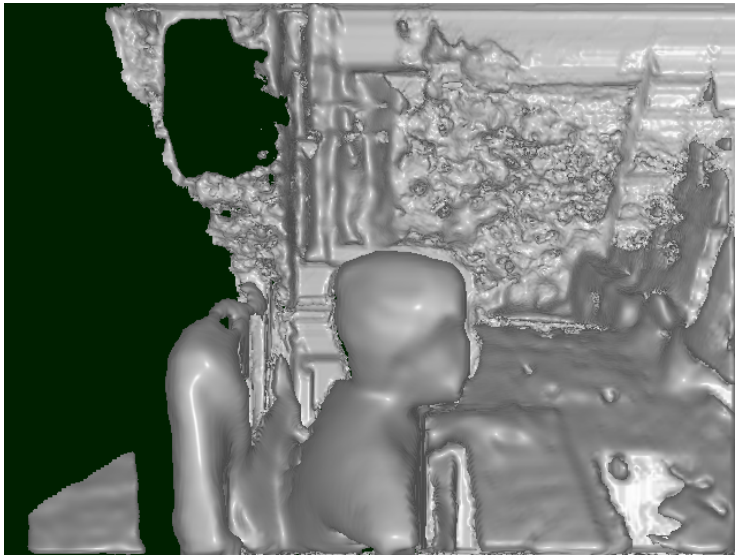
- StereoNet:



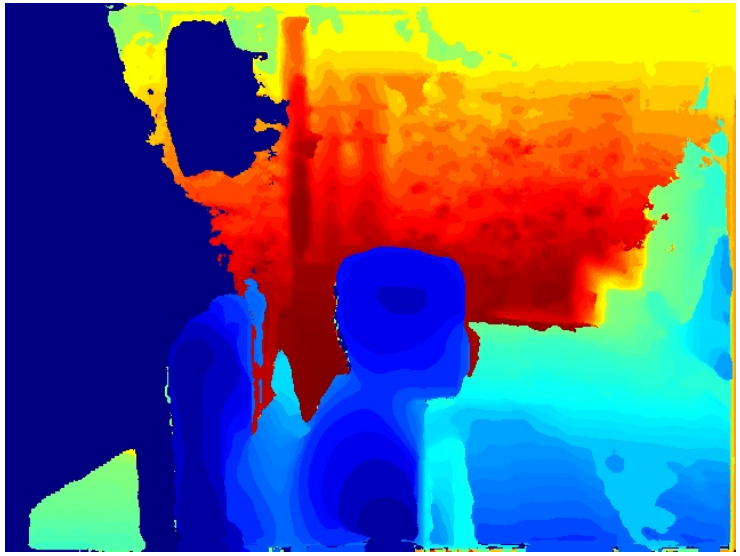
[6] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin and Shahram Izadi, StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware, ECCV 2018
[7] Yinda Zhang, Sameh Khamis, Christoph Rhemann,Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser and Sean Fanello, ActiveStereoNet: End-to-End Self-Supervised Learning for Active Stereo Systems, ECCV 2018

Depth Estimation & Camera Pose

- KinectFusion [8]
  - Depth Map
  - Camera pose



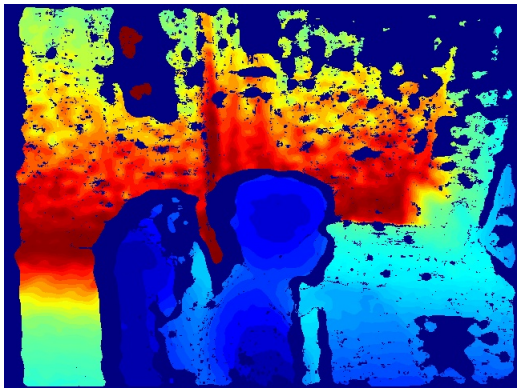
Rendered View of the TSDF



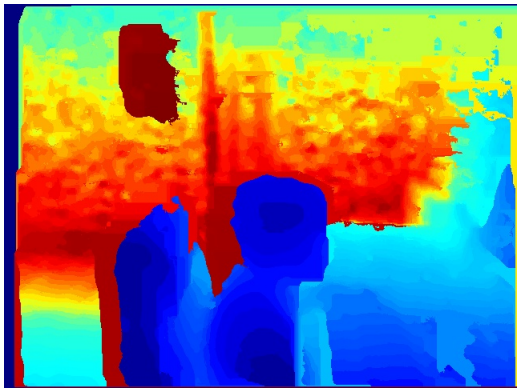
Depth Map

[8] R. A. Newcombe et al., "KinectFusion: Real-time dense surface mapping and tracking," 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, 2011, pp. 127-136, doi: 10.1109/ISMAR.2011.6092378.

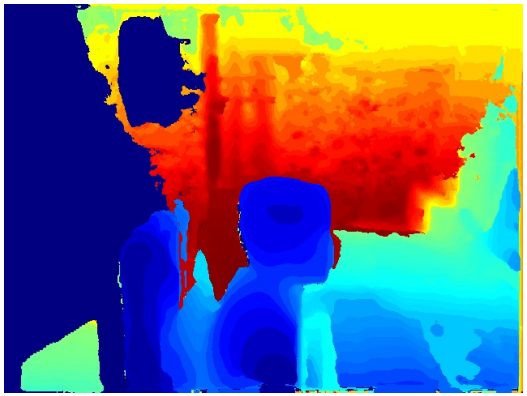
Comparison



Depth Map (Without filtering)



Depth Map (With filtering)



Depth Map (KinectFusion)

Depth-aware Mixed Reality: Capture the AR-Flag

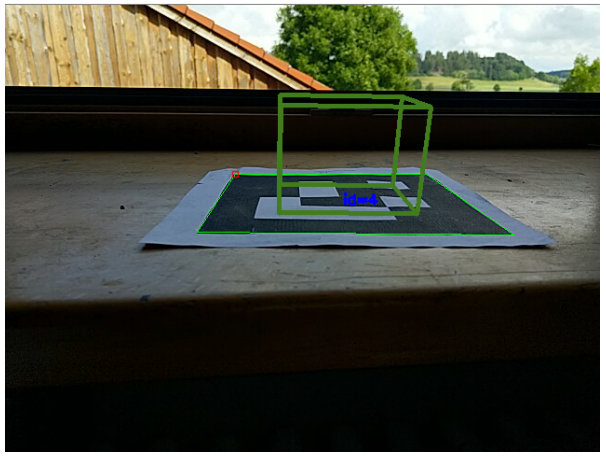
User Interaction and Scene Manipulation

Technische Universität München

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

User Interaction and Scene Manipulation

- ChArUco [9] marker tracking:
  - OpenCV provided implementation
  - Checkerboard pattern combined with ArUco markers
  - Pose estimation
- Improved results with calibrated camera
- 6 DoF interaction with the scene
  - Move the plane
- Estimated marker pose sent to the other scene



[9] Garrido-Jurado, Muñoz-Salinas, Madrid-Cuevas, Marín-Jiménez, Automatic generation and detection of highly reliable fiducial markers under occlusion, Pattern Recognition 2014.

User Interaction and Scene Manipulation

- Human pose estimation using Lightweight OpenPose [10]
- Optimized OpenPose [11]
- Detects keypoints (wrists, knees, ...) and then groups them, i.e. bottom-up approach
- Use estimated pose to interact with the scene
  - Grab the plane

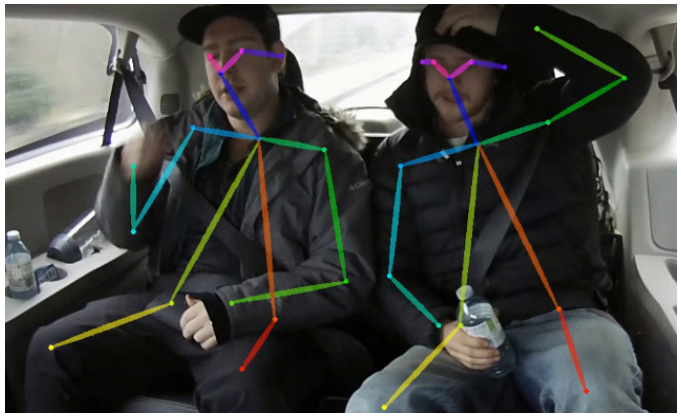


Image taken from [10]

[10] Osokin. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. ArXiv 2018.
[11] Cao, Hidalgo Martinez, Simon, Wei, Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. PAMI 2019.

Depth-aware Mixed Reality: Capture the AR-Flag

Realism and Embedding

Technische Universität München

+

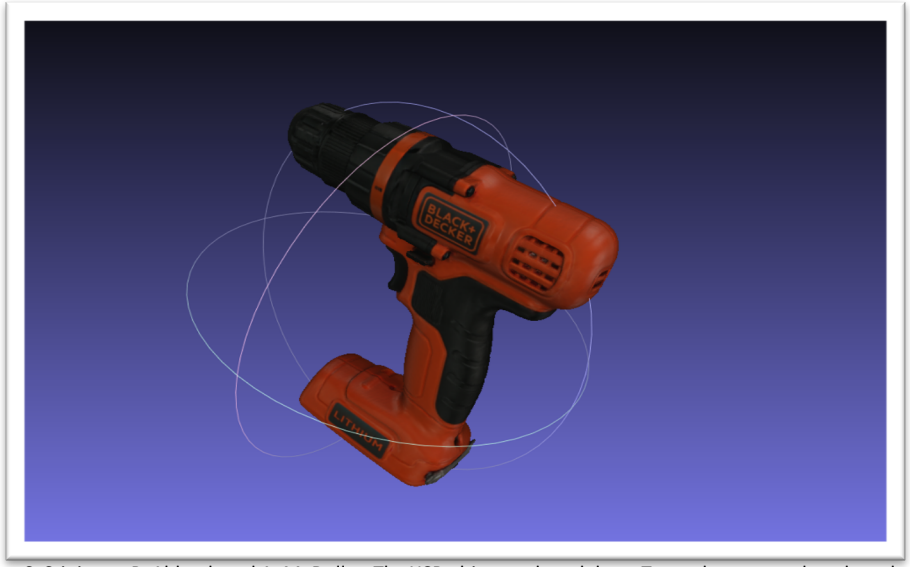
Johns Hopkins
Whiting School of Engineering

Blending Object with the Scene

- The depth map is used to blend the augmented object with occlusion awareness, given their pose.

- Triangle mesh of the object [12] is first rendered using Pyrender library in order to get the color and depth maps.

Triangle Mesh Represented in Meshlab



[12] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. ICAR 2015.

Blending Object with the Scene

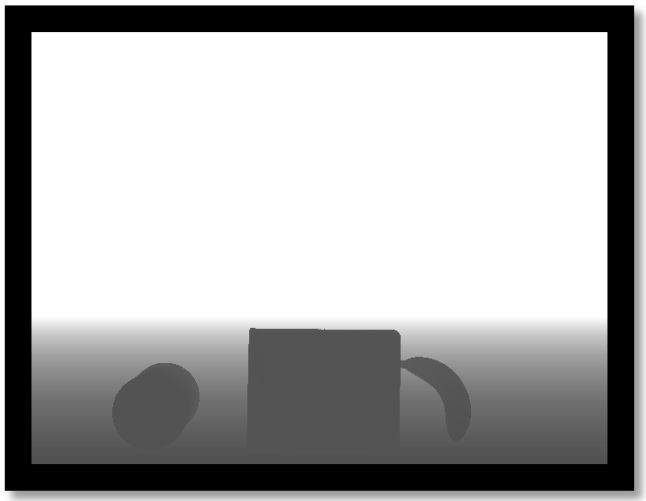
- (w,h,3) channel floating-point color image from Pyrender is on the left and (w,h) floating-point depth image from Pyrender is on the right. Camera parameters (intrinsic and pose) and triangle mesh as well as color and depth image of scene has been found online [12].

[12] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. ICAR 2015.

Blending Object with the Scene

- The object is blended into the scene using its color and depth image. For this, the depth and color images of the scene are used.

Depth image of the scene



RGB Color image of the scene

Blending Object with the Scene

Ground Truth Blending

Our blending output

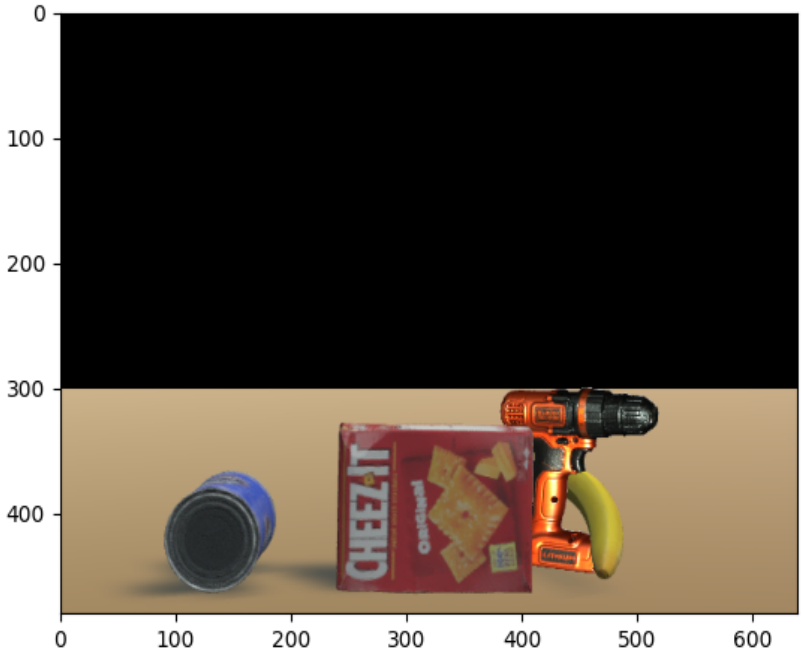
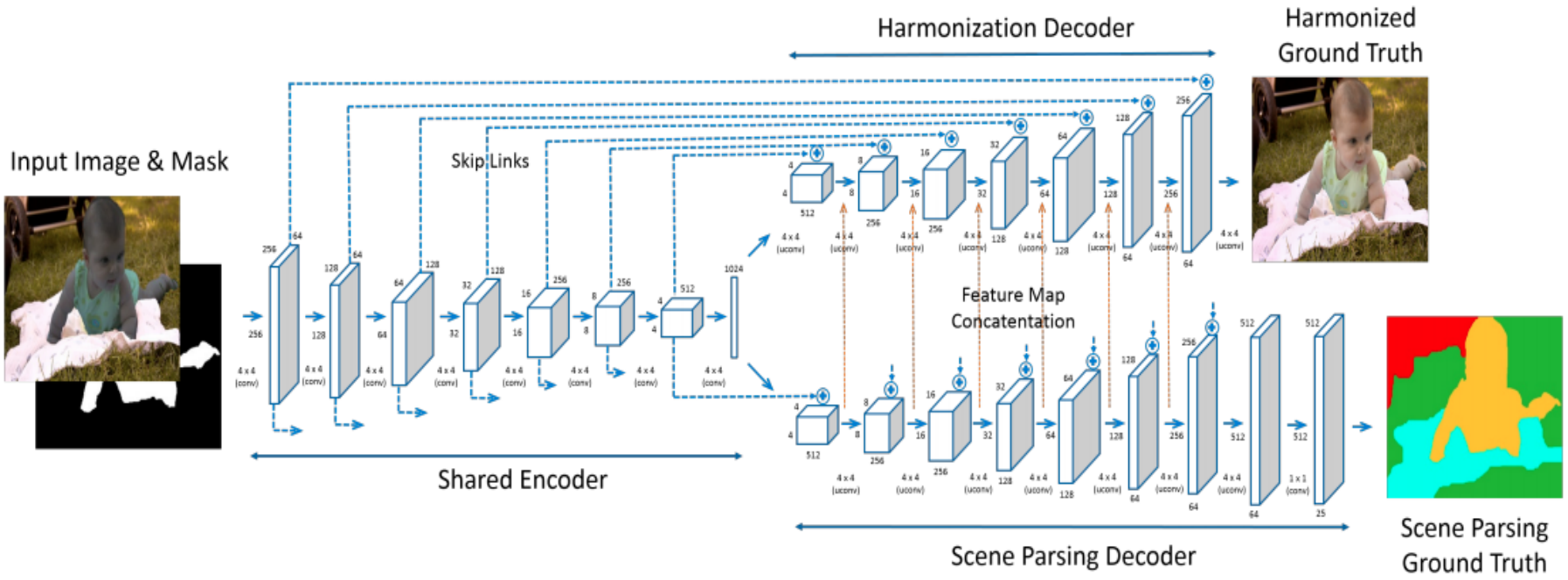


Image Harmonization

- For image harmonization, a pretrained caffe model was used [13].
- The model takes an RGB composite image and a mask of the object to be harmonized. The model consist of shared encoder, scene parsing decoder and harmonization decoder.

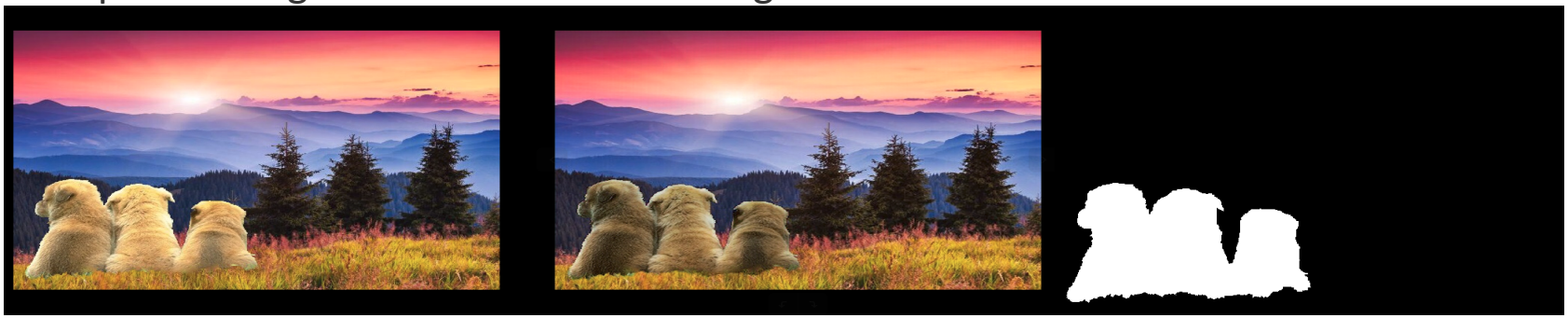


[13] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, Ming-Hsuan Yang. Deep Image Harmonization. CVPR 2017.

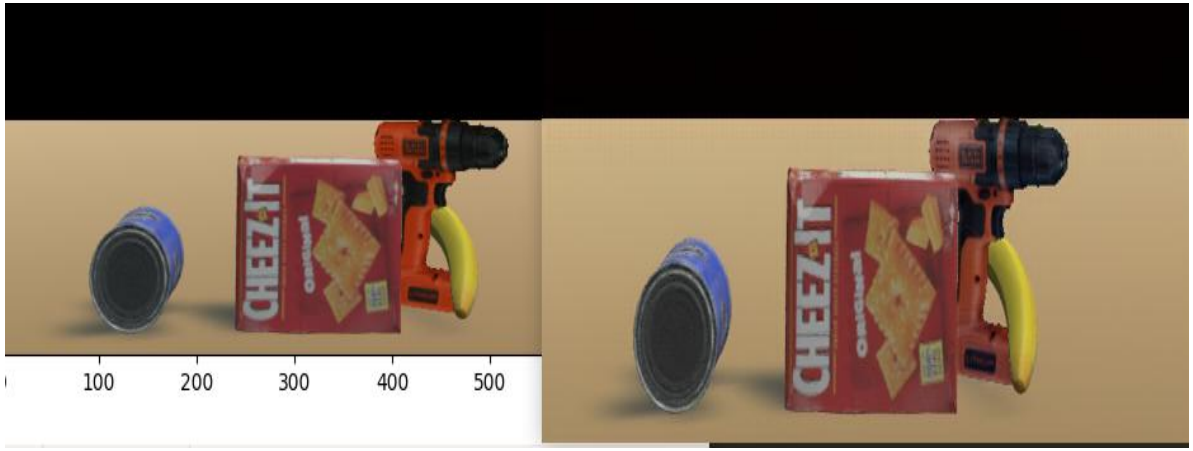
Image Harmonization

- Example results from the git repository [13]:

- Composite image Harmonized image Used Mask



- Results on sample data [14]:
- Composite image Harmonized image



[13] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, Ming-Hsuan Yang. Deep Image Harmonization. CVPR 2017.

[14] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols. RAM 2015.

Depth-aware Mixed Reality: Capture the AR-Flag

Project Demo…

Technische Universität München

+

Johns Hopkins
Whiting School of Engineering

Depth-aware Mixed Reality: Capture the AR-Flag

Students:
Muhammad Faizan
Ihsan Berkan Balaban
Jeremias Neth

Tutors:
Patrick Ruhkamp
Benjamin Busam
Hyunjun Junq

Limitations

- Slow FPS because of remote connections.
- KinectFusion having difficulty in reconstructing a texture-less white wall.
- Human pose estimation model is limited by inference time.
- Image harmonization model is limited to images of certain sizes.

Conclusion

- We managed to properly occlude an object in the scene given their depth and color maps as well as their pose while harmonizing the object at the same time.
- Best results are observed when the filtered depth map is used and image harmonization is enabled.
- The scene can be interacted with in 3D through human pose estimation and marker tracking.

Thank you very much ☺

Any question?

Technische Universität München

+

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

REFERENCES

[1] Intel RealSense D435: https://www.intelrealsense.com/depth-camera-d435/, Accessed: 2020-06-24

[2] Eduardo S.L. Gastal, Manuel M. Oliveira. Domain Transform for Edge-Aware Image & Video Processing (https://www.inf.ufrgs.br/~eslgastal/DomainTransform/)

[3] Intel RealSense SDK Post-Filters:https://github.com/IntelRealSense/librealsense/blob/master/doc/post-processing-filters.md

[4] Disparity map post-filtering: https://docs.opencv.org/trunk/d3/d14/tutorial_ximgproc_disparity_filtering.html

[5] Clément Godard, Oisin Mac Aodha, Michael Firman and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. ICCV 2019

[6] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin and Shahram Izadi, StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware, ECCV 2018

[7] Yinda Zhang, Sameh Khamis, Christoph Rhemann,Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser and Sean Fanello. ActiveStereoNet: End-to-End Self-Supervised Learning for Active Stereo Systems. ECCV 2018

[8] R. A. Newcombe et al. KinectFusion: Real-time dense surface mapping and tracking. ISMAR 2011.

[9] Garrido-Jurado, Muñoz-Salinas, Madrid-Cuevas, Marín-Jiménez, Automatic generation and detection of highly reliable fiducial markers under occlusion, Pattern Recognition 2014.

[10] Osokin. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. ArXiv 2018.

[11] Cao, Hidalgo Martinez, Simon, Wei, Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. PAMI 2019.

[12] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The YCB object and model set: Towards common benchmarks for manipulation research. ICAR 2015.

[13] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, Ming-Hsuan Yang. Deep Image Harmonization. CVPR 2017.

[14] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols. RAM 2015.