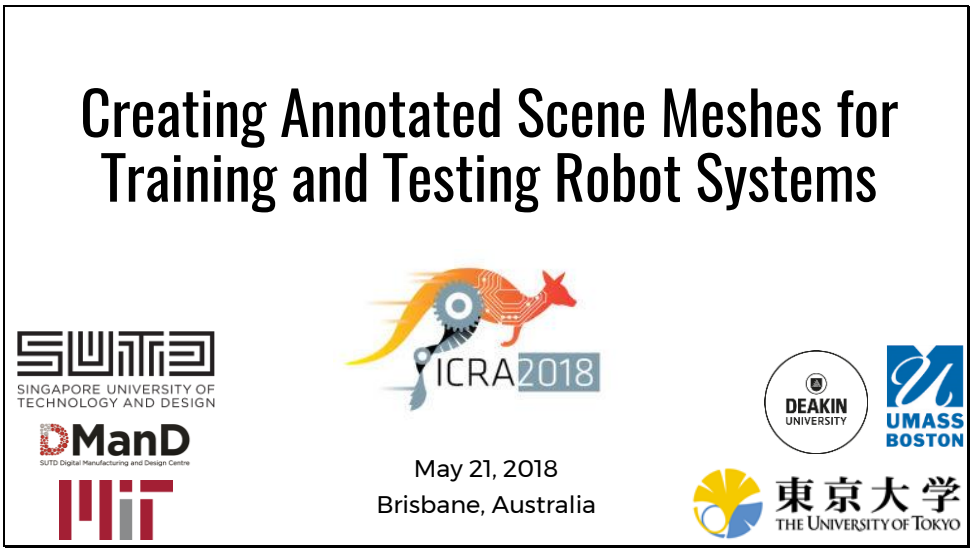
Slide 1



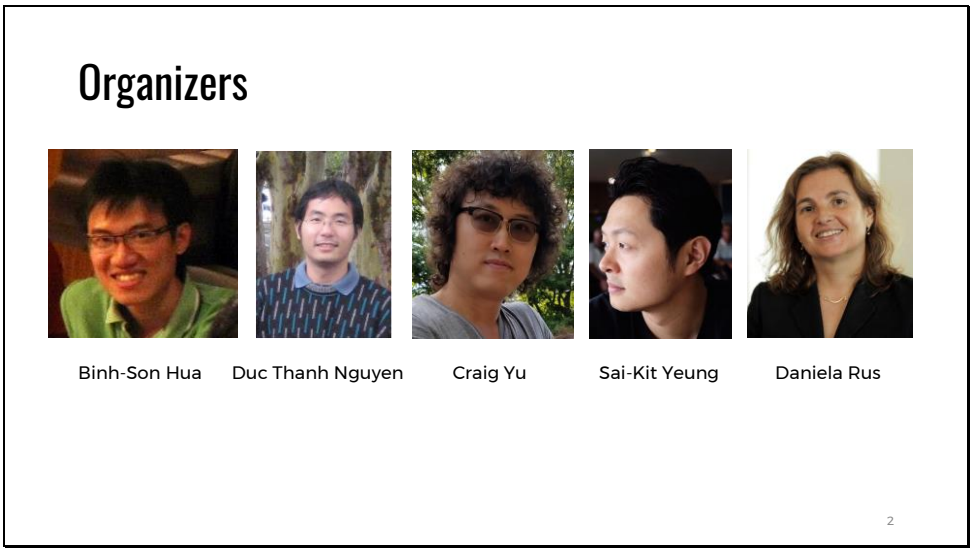
Welcome to our tutorial at ICRA 2018!

Today we are going to share our experience in a number of active research topics in computer vision and robotics, centralizing about creating datasets of 3D scene meshes and their applications to robotics using deep learning.

Our talk is going to be introductory but covers a board range of topics. At some points that we feel the technical details are important, we would discuss them in more depths. Our goal today is to provide the audience sufficient details to reimplement what we had been doing, which is important for industry practitioners, or adapt the ideas to their own research.

This is the first time we deliver such a tutorial, and we are looking to enhance our presentations better and better. We encourage discussions along the talks, so feel free to interrupt and ask questions.

Slide 2



Binh-Son Hua is currently a postdoctoral researcher in The University of Tokyo. Before that, he worked as a postdoctoral researcher for two years at Singapore University of Technology and Design. He received his PhD degree in Computer Science from National University of Singapore in 2015. His research interests are 3D reconstruction, 3D scene understanding, and physically based rendering. His recent works are published in both computer graphics and vision venues, including SIGGRAPH, Eurographics, TVCG, 3DV, and CVPR.

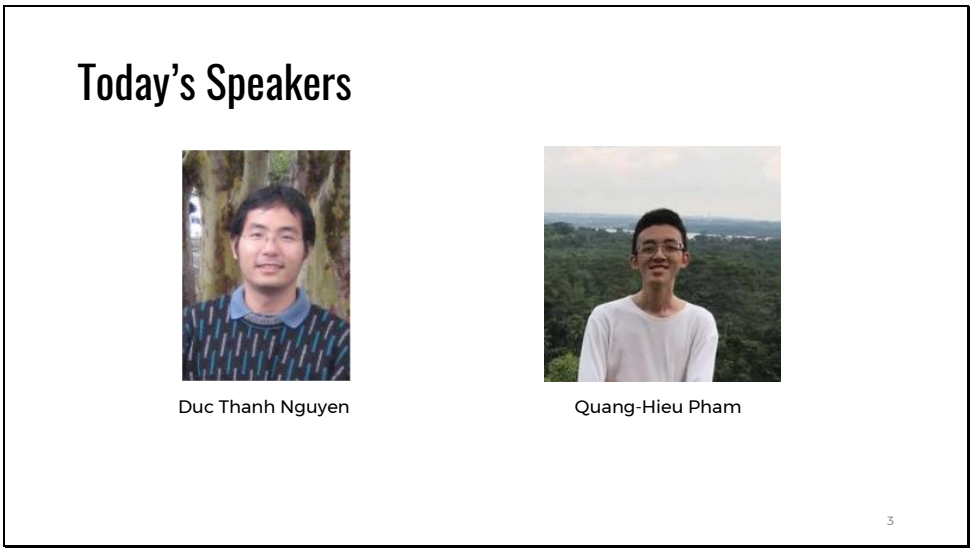
Duc Thanh Nguyen received his Ph.D. degree in Computer Science from the University of Wollongong, Australia, in 2012. Currently, he is a lecturer at the School of Information Technology, Deakin University, Australia. His research interests include Computer Vision and Pattern Recognition. Dr. Nguyen has published his work in highly-ranked publication venues in Computer Vision and Pattern Recognition such as the Journal of Pattern Recognition, CVPR, ICCV and ECCV. He also has served a technical program committee member of the IEEE Int. Conf. Image Process. (from 2012) and reviewers of the IEEE Trans. Intell. Transp. Syst., IEEE Trans. Image Process., IEEE Signal Processing Letters, Image and Vision Computing.

Lap-Fai (Craig) Yu is an assistant professor at the University of Massachusetts at Boston. He obtained his PhD degree in computer science from UCLA in 2013. His research interests are in computer graphics and vision, especially in the topics of synthesizing and analysing 3D models from the perspectives of functionality, physics, intentionality and causality. He is the recipient of the Cisco Outstanding Graduate Research Award, the UCLA Dissertation Year Fellowship, the Sir Edward Youde Memorial Fellowship and the Award of Excellence from Microsoft Research. His research has been featured in New Scientist, the UCLA Headlines and newspapers internationally. He regularly serves on the program committee of Eurographics, Pacific Graphics and IEEE Virtual Reality.

Sai-Kit Yeung is currently an Assistant Professor at the Singapore University of Technology and Design (SUTD), where he leads the Vision, Graphics and Computational Design (VGD) Group. He was also a Visiting Assistant Professor at Stanford University and MIT. Before joining SUTD, he had been a Postdoctoral Scholar in the Department of Mathematics, University of California, Los Angeles (UCLA). He was also a visiting student at the Image Processing Research Group at UCLA in 2008 and at the Image Sciences Institute, University Medical Center Utrecht, the Netherlands in 2007. He received his PhD in Electronic and Computer Engineering from the Hong Kong University of Science and Technology (HKUST) in 2009. He also received a BEng degree (First Class Honors) in Computer Engineering in 2003 and a MPhil degree in Bioengineering in 2005 from HKUST. His research interests include computer vision, computer graphics and computational fabrication.

Daniela Rus is the Andrew (1956) and Erna Viterbi Professor of Electrical Engineering and Computer Science and Director of the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT. Rus's research interests are in robotics, mobile computing, and data science. Rus is a Class of 2002 MacArthur Fellow, a fellow of ACM, AAAI, IEEE and RAS, and a member of the National Academy of Engineering, and the American Academy for Arts and Science. She earned her PhD in Computer Science from Cornell University. Prior to joining MIT, Rus was a professor in the Computer Science Department at Dartmouth College.

Slide 3



Today we apologize that Son, our first organizer, cannot attend due to his job transition to the University of Tokyo in February 2018. However, he has been actively working with us to prepare the materials for this tutorial.

Hieu, a PhD candidate in Singapore University of Technology and Design, will present the tutorial on Son's behalf.
Hieu received his bachelor degree at Ho Chi Minh University of Science, Vietnam. He achieved third place in the Vietnam Computing Olympiad in 2010. His current research interests are 3D reconstruction, scene understanding and indoor scene navigation.

Hieu has been working very closely with our team since 2015. He co-authored our dataset paper and helps build the pipeline for scene annotation, and so he understands almost every detail in the pipeline.

Slide 4

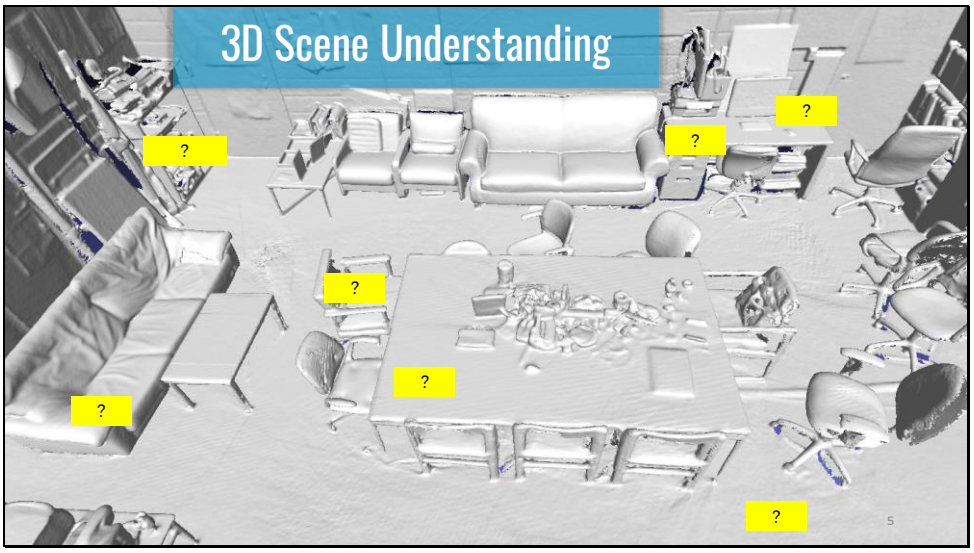
In general, our works (reconstruction and segmentation) are centralized around understanding the world, by inferring semantics of scene objects.

Here by semantics, we mean “the meaning” of objects in the scene. Think of yourself as a robot. When you interact with the environment, you have to know what you are interacting with. So, the “what” here is semantics. For example, cup, chair, table are some example objects that we often see in an indoor scene.

For human, scene understanding is often a natural task. What we look at an image of a room, we often understand whether it is a living room, a kitchen, a bedroom, a bathroom, or an office. We are also able to recognize rapidly what and where the objects are.

Understanding indoor scenes is an active research area in computer vision and robotics. Keywords are semantic segmentation, object recognition, object detection. You can find many new papers appearing almost every month, even every day if you follow arXiv. This is an active research direction that has been applied to robot navigation, object manipulation, self-driving car, and so on.

Slide 5



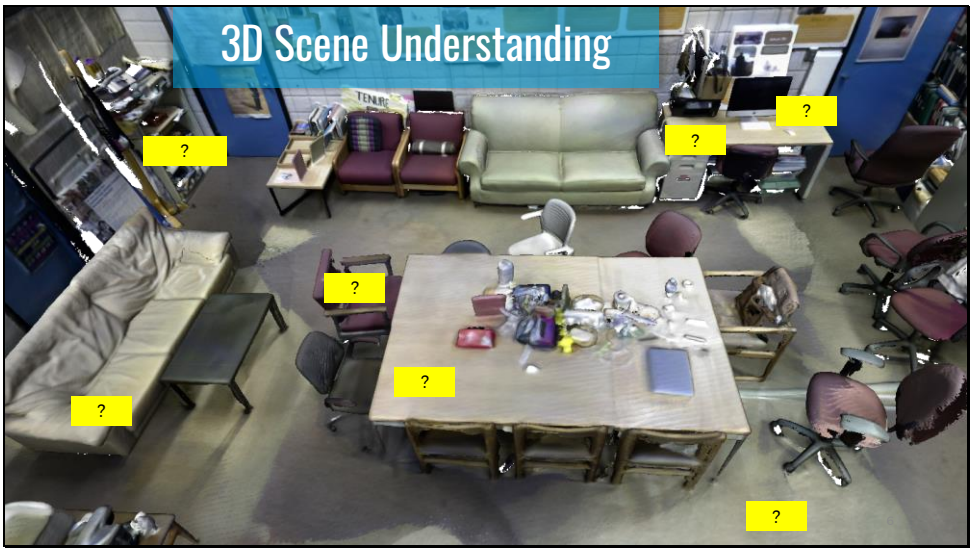
Today we are going to explore scene understanding in 3D. In general, 2D scene understanding works well from the massive availability of image data, however, it is challenging for a single image to capture full information of a scene. Also, it is also difficult to reconstruct depth from a single image in general.

Using multiple images could be a solution, as it helps capture more information and establish the correspondences in the images. For example, when a chair appears in two images, the correspondence helps reason whether they are the same or from two different scenes. In fact, scene understanding with multiple images is the implicit form of 3D scene understanding. It just does not deal directly with 3D data.

With the rise of depth sensors, 3D data will become more popular, and being able to work on 3D data could help improve the accuracy in several scene understanding tasks. We can ask the same question as with 2D scene understanding task, including object classification, semantic segmentation, and object detection.

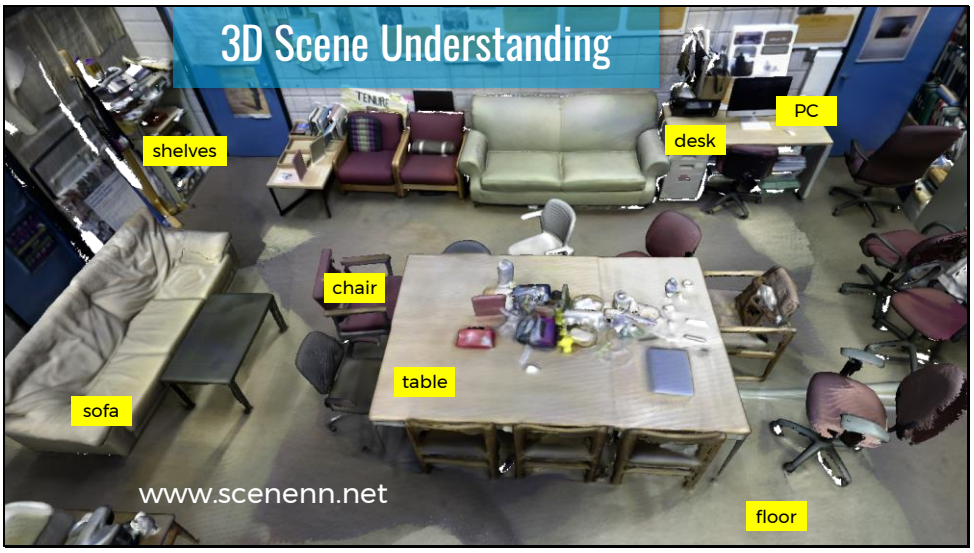
As you can see in this example, this is a 3D reconstructed scene from a consumer depth sensor. It is generally believed that reasoning with 3D data is more accurate.

Slide 6



This is the same scene but displayed in color. So in 3D, we can capture rich information along with depth information in the scene, from 3D coordinates, surface normal, vertex color, to thermal and spectral information. Such data will provide stronger cues for scene understanding, compared to 2D which we do not have depth and it is difficult to align color and other information such as thermal or spectral data into a single representation.

Slide 7

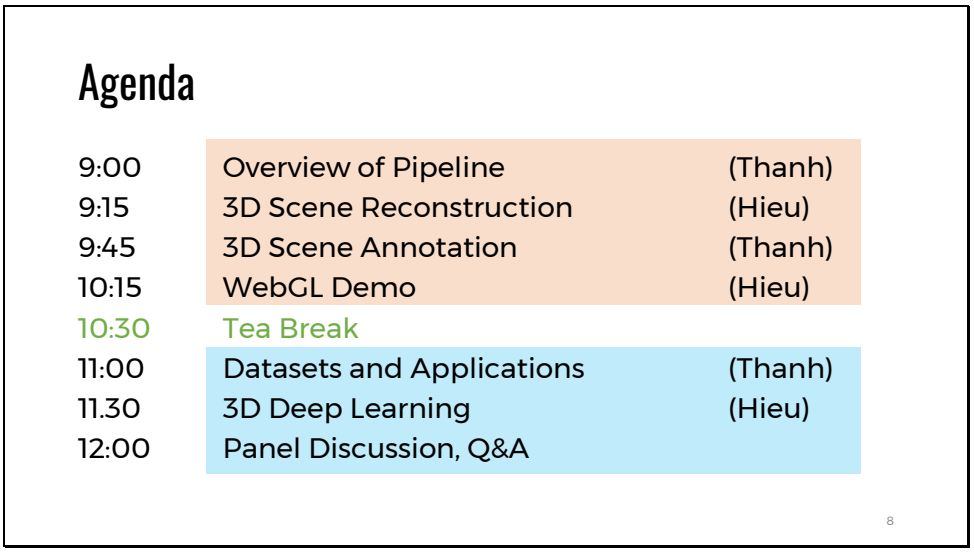


Today we will share our experience in acquisition, annotation of 3D scenes, and how we built a dataset of 3D annotated scene meshes with more than 100 scenes.

The image you see here is one of the scenes in the dataset. If you are interested, you can check our homepage at www.scenenn.net for more details.

The agenda is as follows.

Slide 8



Here is the agenda. In general, there are two main parts: the pipeline, and its applications.

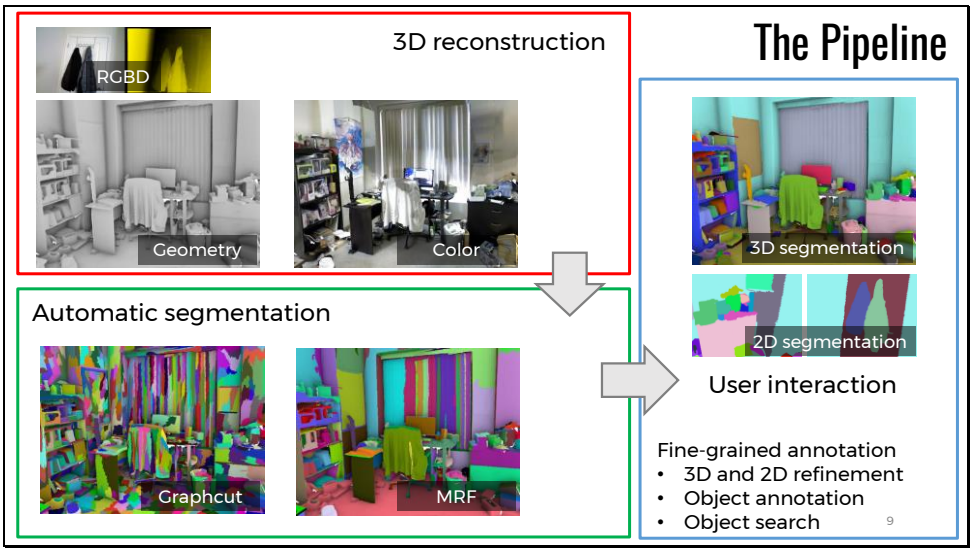
In the first part, we will discuss in details various stages to build a 3D scene dataset. We will also evaluate the advantage and disadvantages of the current pipeline, and suggest how to improve for a more cost-effective pipeline.

In the second part, we will briefly summarize the dataset we created with this pipeline, and discuss potential applications by using this dataset such as object classification and functionality reasoning. A powerful tool to solve such problems nowadays is deep learning.

Therefore, we will spend about half an hour to discuss recent advances of deep learning with 3D data and highlight the state-of-the-art techniques. We hope that this will give the audience the big picture of the entire process, from creating datasets to training the machine intelligence.

After today's panel discussion at the end of the tutorial, you are welcome to email us for any comment and feedback.

Slide 9



Here are the main stages in our pipeline to create annotated scene meshes. This pipeline has been utilized for building the SceneNN dataset with more than 100 indoor scenes.

Our proposed system has three main components: 3D reconstruction, automatic segmentation, user interaction.

We will quickly demonstrate 3D reconstruction before diving into more details in automatic segmentation and user interaction. We will also present advanced features such as 2D refinement and object search to support user to annotate more efficiently.