HKUST

SurfaceNet: an End-to-end 3D Neural Network for Multiview Stereopsis (MVS)

Mengqi Ji*[1], Juergen Gall[3], Haitian Zheng[2], Yebin Liu[2] and Lu Fang†[2]

[1]Hong Kong University of Science and Technology, Hong Kong, China
[2]Tsinghua University, Beijing, China
[3]University of Bonn, Bonn, Germany

Presenter: Mengqi JI (HKUST)

HKUST

Contents

- Introduction to MVS
  - Existing works
- *SurfaceNet*
  - 2 views case
  - N views case
- Experiment
  - Prepare dataset
  - Comparison
- Conclusion

HKUST

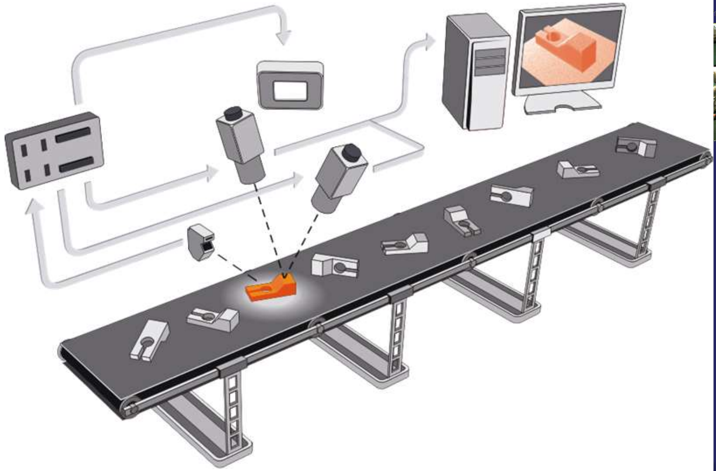
Contents

Introduction to MVS

- Multi-view Stereopsis (**MVS**) / 3D reconstruction
- **Task:**
  - **Inputs:** images with pose parameters
  - **Outputs**: reconstructed 3D representation, such as point cloud, mesh, volumetric …
- **Difficulties:**
  - A lot of information loss (Occlusions)
  - Non-Lambertian surface
  - Textureless region
  - …

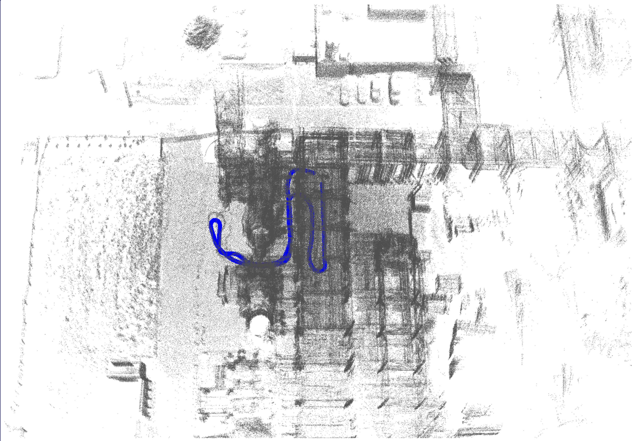
http://cs.bath.ac.uk/~nc537/images/projects/mvs_vase.png

3D Reconstruction Applications

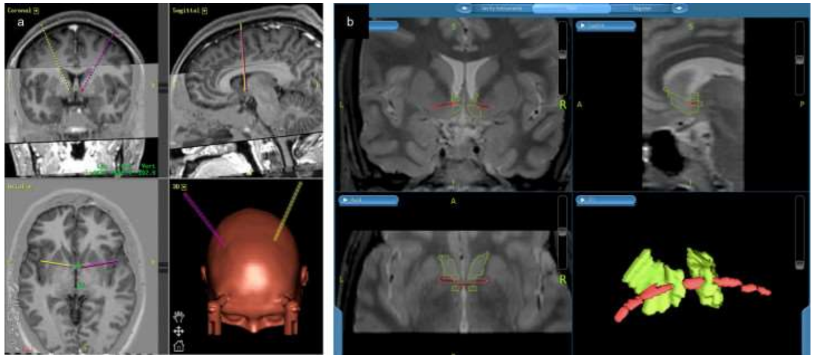


Inspection

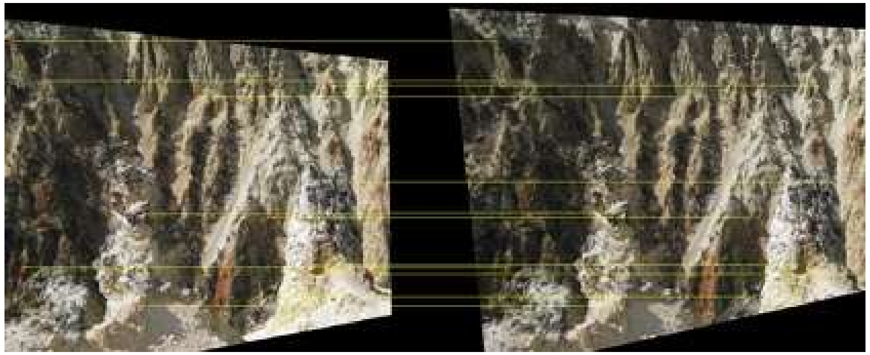
Motion Capture



Localization & Navigation



Medical Imaging



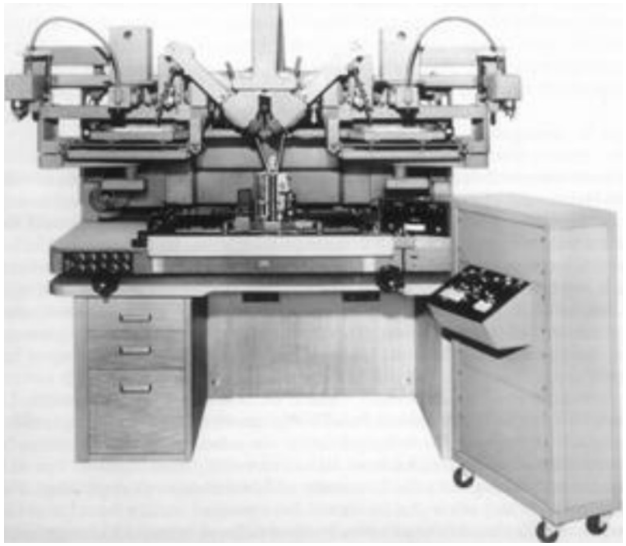
Accurate Measurement

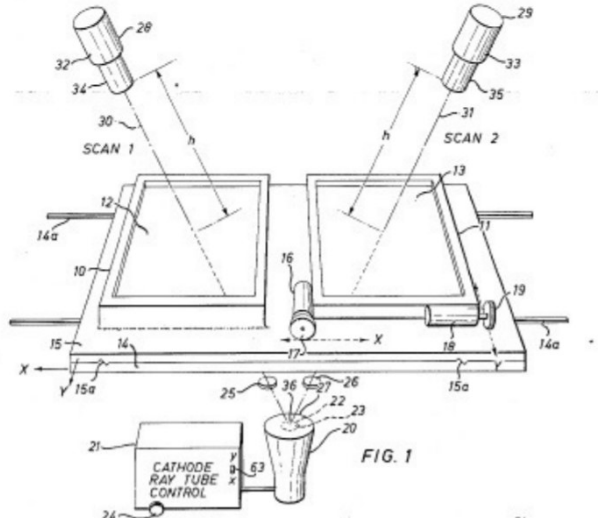
…

HKUST

3D Reconstruction History

- Before 1957, operators manually find correspondences
- In 1957, Gilbert Hobrough demonstrated an analog implementation of **stereo image correlation** (**patent** shown right).
  - 2 transparient images
  - 1 illuminator below
  - 2 sensors above → compare intensity difference

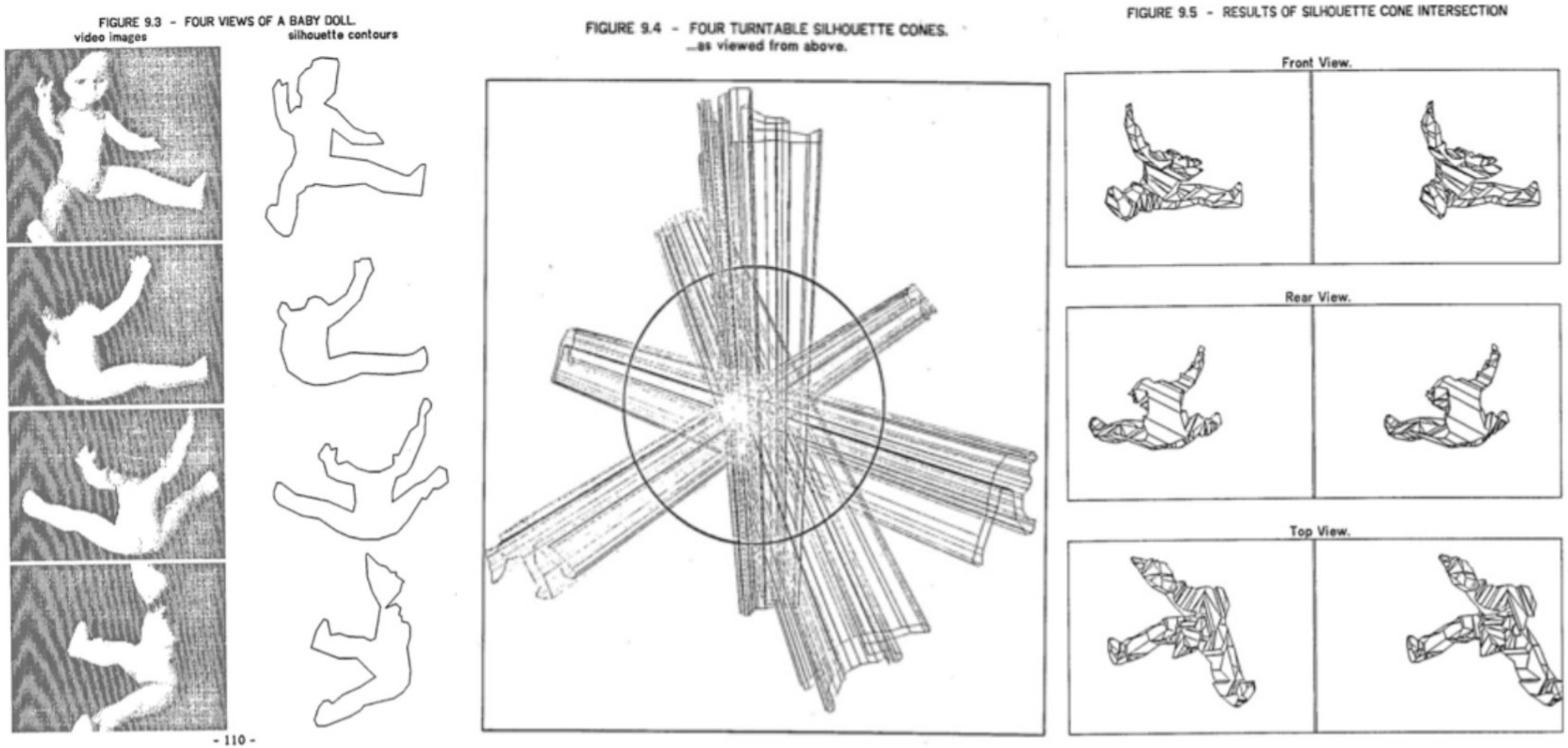




http://www.freepatentsonline.com/2964642.html

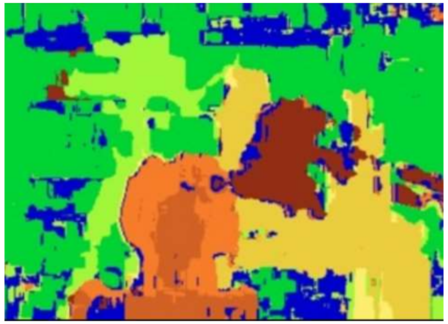
3D Reconstruction History

- 1974: shape from silhouettes [Bruce G. Baumgart, Ph.D Thesis]
    - But requires images to segmented.

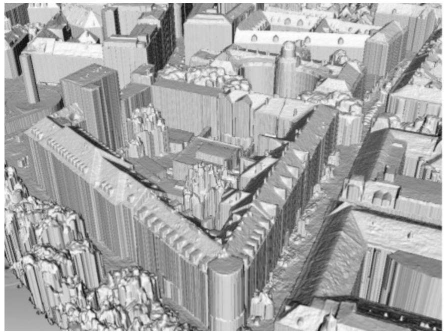


3D Reconstruction History

- 1998: more dense models
  - Graph cut era
  - Local priors: consider local smoothness assumption: nearby pixels are encouraged to have similar appearance and depth



1998 CVPR: Boykov, Veksler, Zabih, **Graph cut** Stereo



2006 PAMI: Hirschmueller

- 2010: large scale with fine geometry details

2010 PAMI: Furukawa et al.

HKUST

Contents

- **Introduction to MVS**
  - **Existing works**

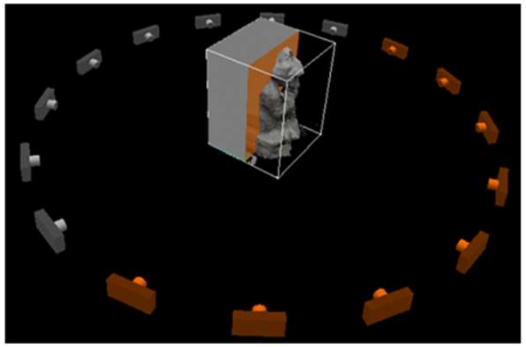
- *SurfaceNet*
  - 2 views case
  - N views case

- Experiment
  - Prepare dataset
  - Comparison

- Conclusion

Related Works

- **Standard pipelines**:
  1. Volumetric methods, such as:
     - space carving [Seitz & Dyer, CVPR 1997],
     - ray potential model [Ulusoy, Geiger, & Black, 3DV 2015].
  2. Depth map fusion methods.



http://www.ctralie.com/PrincetonUGRAD/Projects/SpaceCarving/

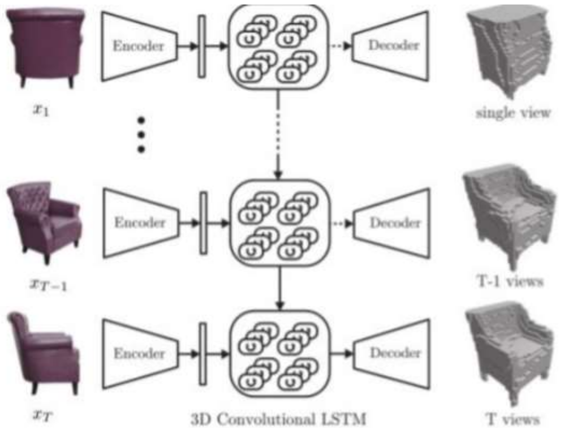
[Furukawa, et al. Multi-view stereo: A tutorial]

- **Problem**:
  1. Computationally expensive graph modelling.
     - Hard to model and solve
  2. Hand engineered pipeline.
     - Exist multiple potential sub-optimal choices.
- **Ours:**
  - Can we learn to reconstruct from data → easy to train & solve

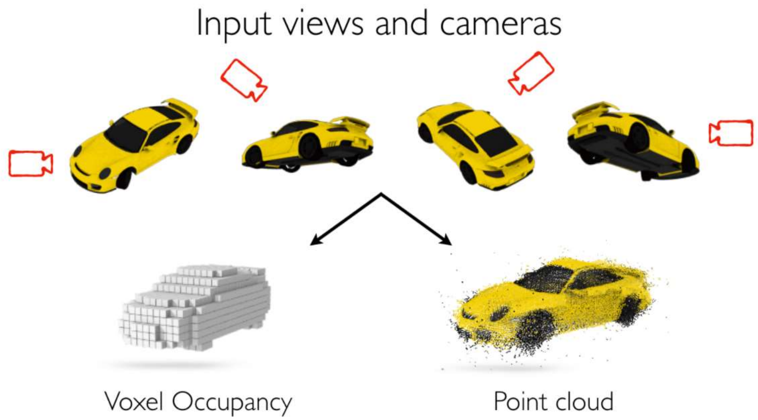
HKUST

Related Works

- **Learning based 3D Reconstruction**:
  - Idea: Learn a mapping from observations to their underlying 3D shape



[2016ECCV, Choy et al., 3D-R2N2]



[2017NIPS, Kar et al., Learning a Multi-View Stereo Machine]

- **Problem:**
  - Using Shape **Priors**: reconstruct specific type of models
  - Resolution limitation
- **Ours:**
  - More general 3D reconstruction **with** fine detail and **without** shape priors.

Contents

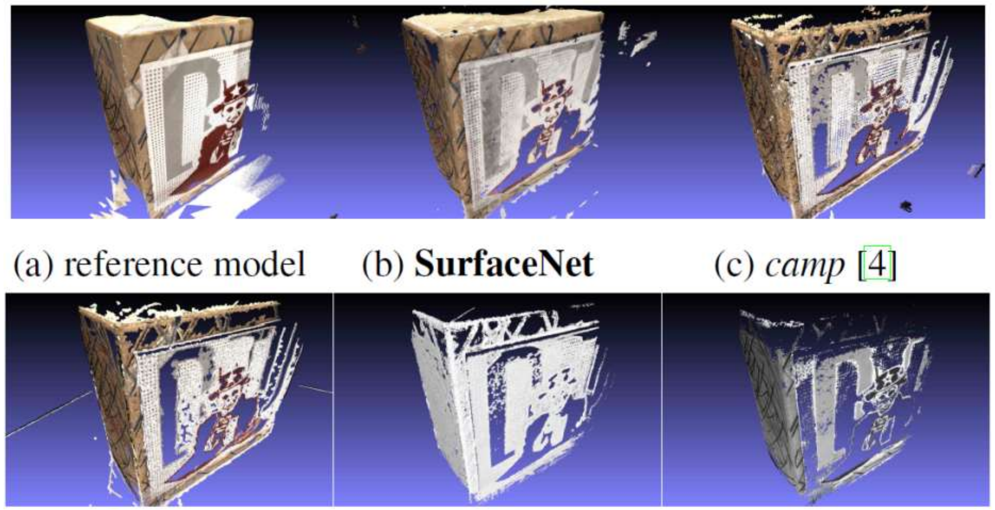
- Introduction to MVS
  - Existing works
- ***SurfaceNet***
  - 2 views case
  - N views case
- Experiment
  - Prepare dataset
  - Comparison
- Conclusion

Introduction to MVS

- **Question**: Can we design an end-to-end learning framework for MVS without shape priors?

- **Reinterpretation**: MVS predicts **2D** surface from a **3D** voxel space, analogous to boundary detection, which predicts a **1D** boundary from **2D** image input.

- ***SurfaceNet***: **first** end-to-end learning framework for MVS
  - takes the image + camera parameters and infers the 3D surface **directly**.
  - photo-consistency and geometric context for dense reconstruction
  - better completeness around the less textured regions compared with other methods.



(a) reference model (b) **SurfaceNet** (c) *camp* [4]

(d) *furu* [8] (e) *tola* [27] (f) *Gipuma* [9]

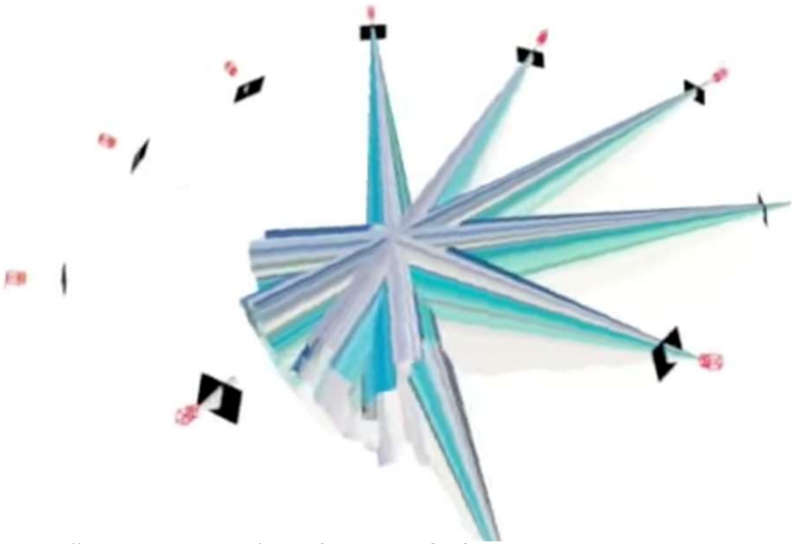
SurfaceNet ---- colored voxel cube (CVC)

HKUST

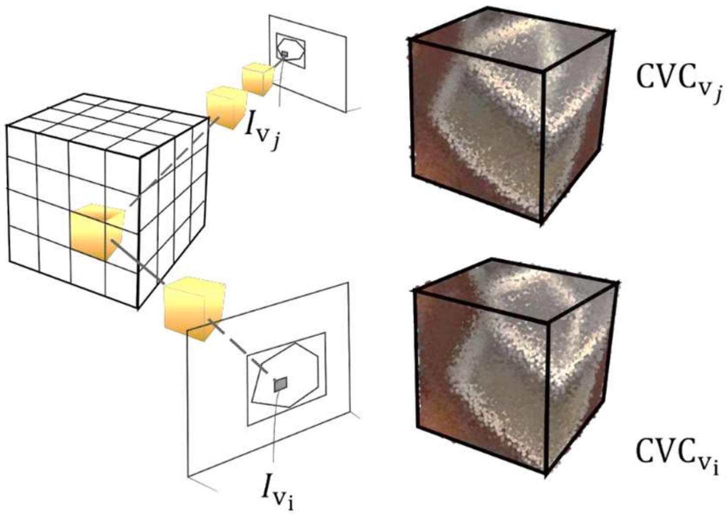
- **Problem**: how to embed the camera parameter into the network;
  perspective projection is straightforward and highly non-linear.

- **Solution**: 3D voxel representation for each view: *colored voxel cube (**CVC**)*
  - Scene → overlapping volumes → voxel grid
  - Each pixel corresponds to a voxel ray.
  - Colorize different voxels on the same voxel ray as the same color

- Implicitly encodes the camera parameters into a 3D *colored voxel cube*



https://www.youtube.com/watch?v=21YUA-SalO0



HKUST

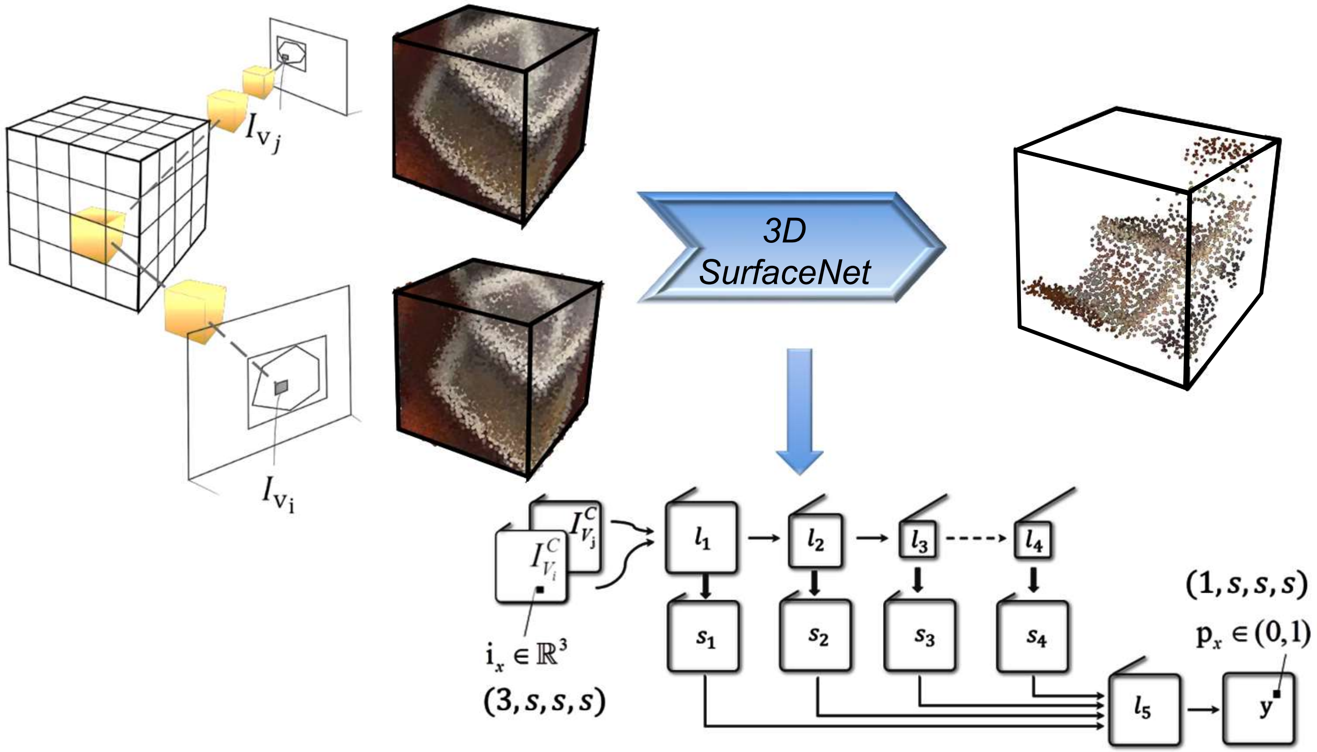
Contents

- Introduction to MVS
  - Existing works
- *SurfaceNet*
  - **2 views case**
  - N views case
- Experiment
  - Prepare dataset
  - Comparison
- Conclusion

HKUST

SurfaceNet ---- 2 views case

- pipeline
  - takes 2 colored voxel cubes from 2 different views as input
  - predicts for each voxel a binary occupancy attribute indicating if the voxel is on the surface or not.
- *SurfaceNet* predicts **2D** surface from a **3D** voxel space,
- analogous to boundary detection [2], which predicts a **1D** boundary from **2D** image input.



[2] Xie, Saining, and Zhuowen Tu. "Holistically-nested edge detection." Proceedings of the IEEE International Conference on Computer Vision. 2015.

HKUST

Contents

- Introduction to MVS
  - Existing works
- *SurfaceNet*
  - 2 views case
  - **N views case**
- Experiment
  - Prepare dataset
  - Comparison
- Conclusion

SurfaceNet ---- N view pairs

- **Fuse**: average N results from N view pairs.

- **Problem**: when there are multiple views, how to choose less views to get good 3D model.
    - 50 views → 1000+ view pairs

- **Solution**:
    - only use the valuable view pairs ranked by **relative importance w**
    - **w** is learned for each view pair based on baseline and the image appearance on both views

$$w_C^{(v_i,v_j)} = r\left(\theta_C^{(v_i,v_j)}, d_C^{(v_i,v_j)}, e(C, I_{v_i})^T, e(C, I_{v_i})^T\right)$$

$$d_C^{(v_i,v_j)} = \|e(C, I_{v_i}) - e(C, I_{v_i})\|_2$$

- Weighted average the results **p** from different view pairs

$$p_x = \frac{\sum_{(v_i,v_j)\in \mathbf{V}_C} w_C^{(v_i,v_j)} p_x^{(v_i,v_j)}}{\sum_{(v_i,v_j)\in \mathbf{V}_C} w_C^{(v_i,v_j)}}$$

SurfaceNet ---- N view pairs

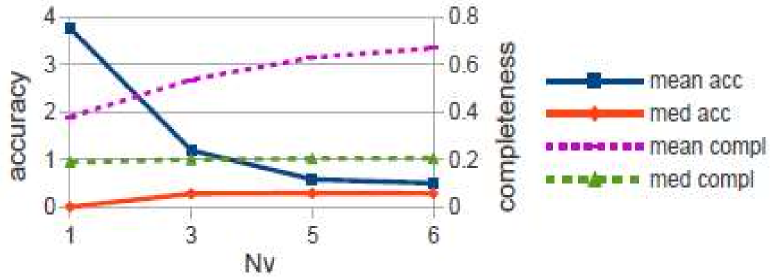
- Compare:
  - (**left**) Randomly select 5 view pairs out of 1000+.
  - (**Right**) Select 5 view pairs with top **w** value

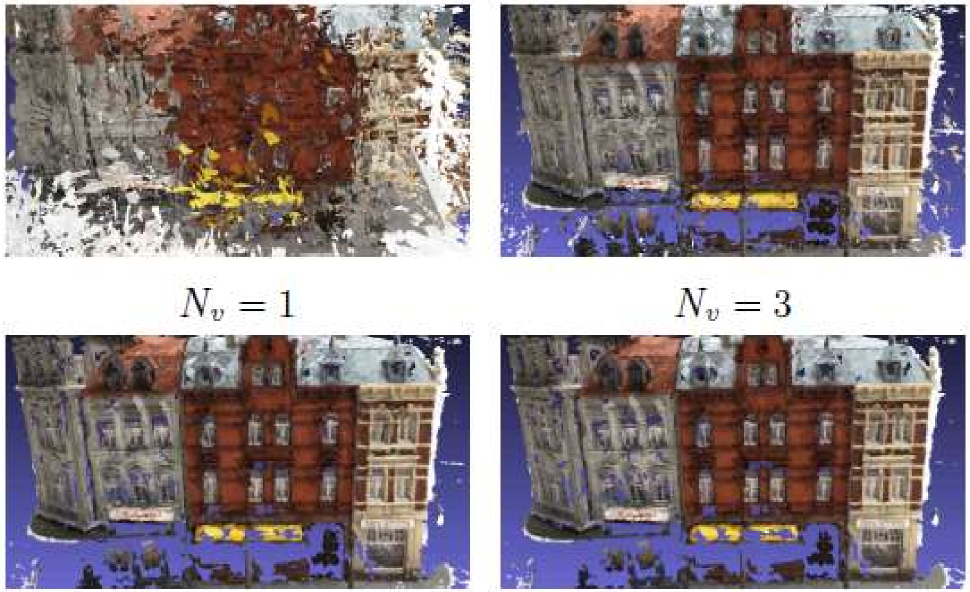
- (**Right**) is much complete with little accuracy drop than (**left**).

Random model 9	mean accuracy	median accuracy	mean completeness	median completeness
Randomly select view pairs (**Left**)	**0.421**	**0.268**	16.611	1.219
Select top view pairs based on relative importance rank (**Right**)	2.777	0.364	**4.669**	**0.281**

SurfaceNet ---- N view pairs

- Quantitative and qualitative evaluation of N
  - the lower, the better
  - Only take the best view pair, N = 1:
    - Very noisy inaccurate results
  - N = 3:
    - The accuracy is substantially improved.
  - N = 5 + :
    - The accuracy slightly improves.
    - Time consumption linear increases.
  - Trade off choice: N = 5

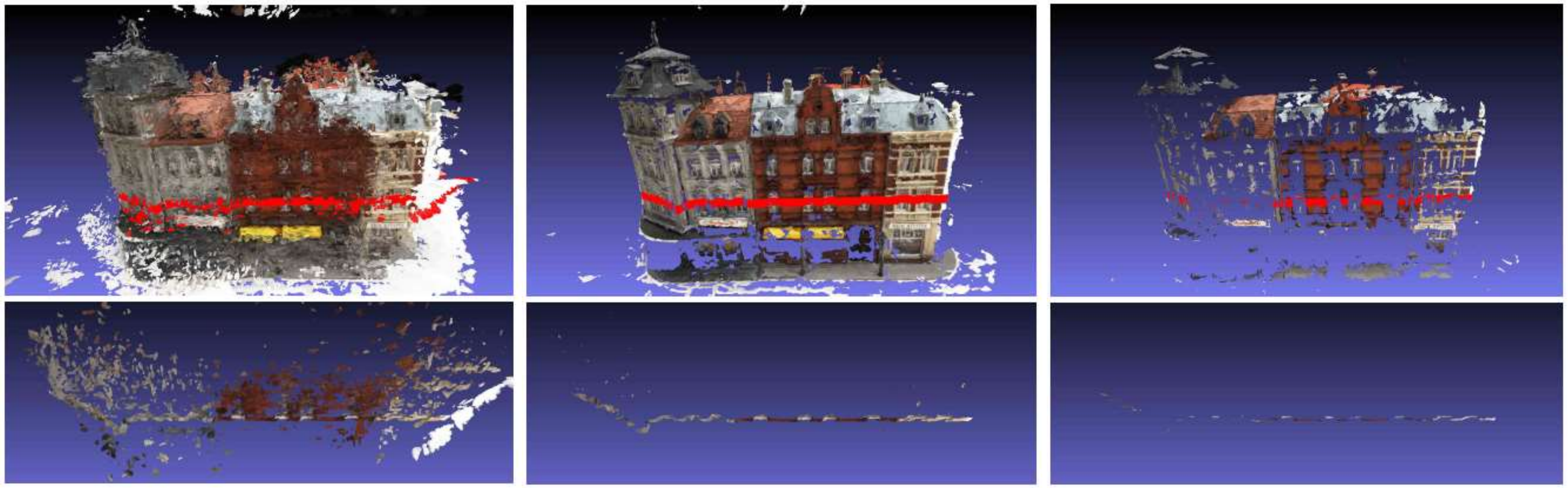




$N_v = 1$ $N_v = 3$ $N_v = 5$ $N_v = 6$

SurfaceNet ---- N view pairs

- **Binarization**: converts the probability map
  - **Uniform** threshold:



$\tau = 0.5$ $\tau = 0.7$ $\tau = 0.9$

  - **Adaptive** threshold: Since the neighboring cubes are helpful for the binarization.

$$E(\tau_C) = \sum_{C' \in \mathcal{N}(C)} \psi\left(S^C(\tau_C), S^{C'}(\tau_{C'})\right) \qquad \tau_C \in [0.5, 1)$$

$$\psi(S^C, S^{C'}) = \sum_{x \in C \cap C'} (1 - s_x)s'_x + s_x(1 - s'_x) - \beta s_x s'_x$$

HKUST

Contents

- Introduction to MVS
    - Existing works
- *SurfaceNet*
    - 2 views case
    - N views case
- **Experiment**
    - **Prepare dataset**
    - Comparison
- Conclusion

Experiments: Prepare Dataset

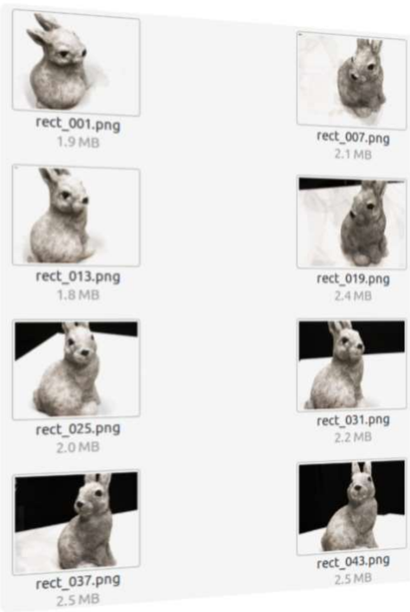
- Use the DTU dataset [3]
  - To our knowledge, [3] is the only large scale MVS benchmark.
  - Contain 80 different scenes seen from 49 camera positions.

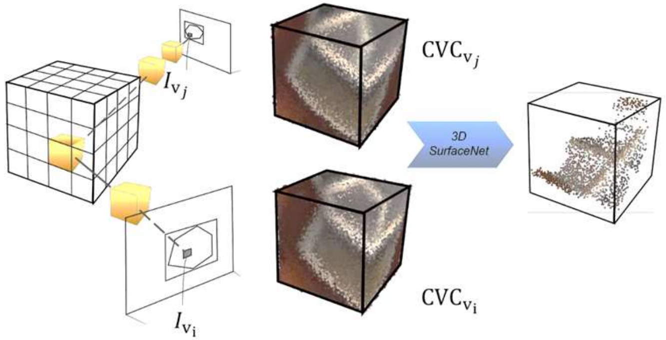
- Limited by the GPU memory, the cube size is set to (32, 32, 32)
  - The cubes are randomly cropped on the training model surface.
  - Data augmentation: rotation and translation

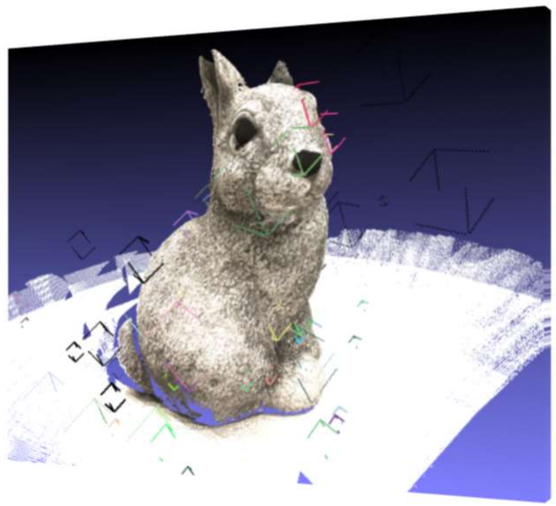
[3] Aanæs, Henrik, et al. "Large-scale data for multiple-view stereopsis." International Journal of Computer Vision 120.2 (2016): 153-168.

Experiments: Prepare Dataset

- {Net_inputs, Net_gt} pairs for training:
  - Posed images → CVCs
  - Laser scanned 3D model → gt (surface points in cube)





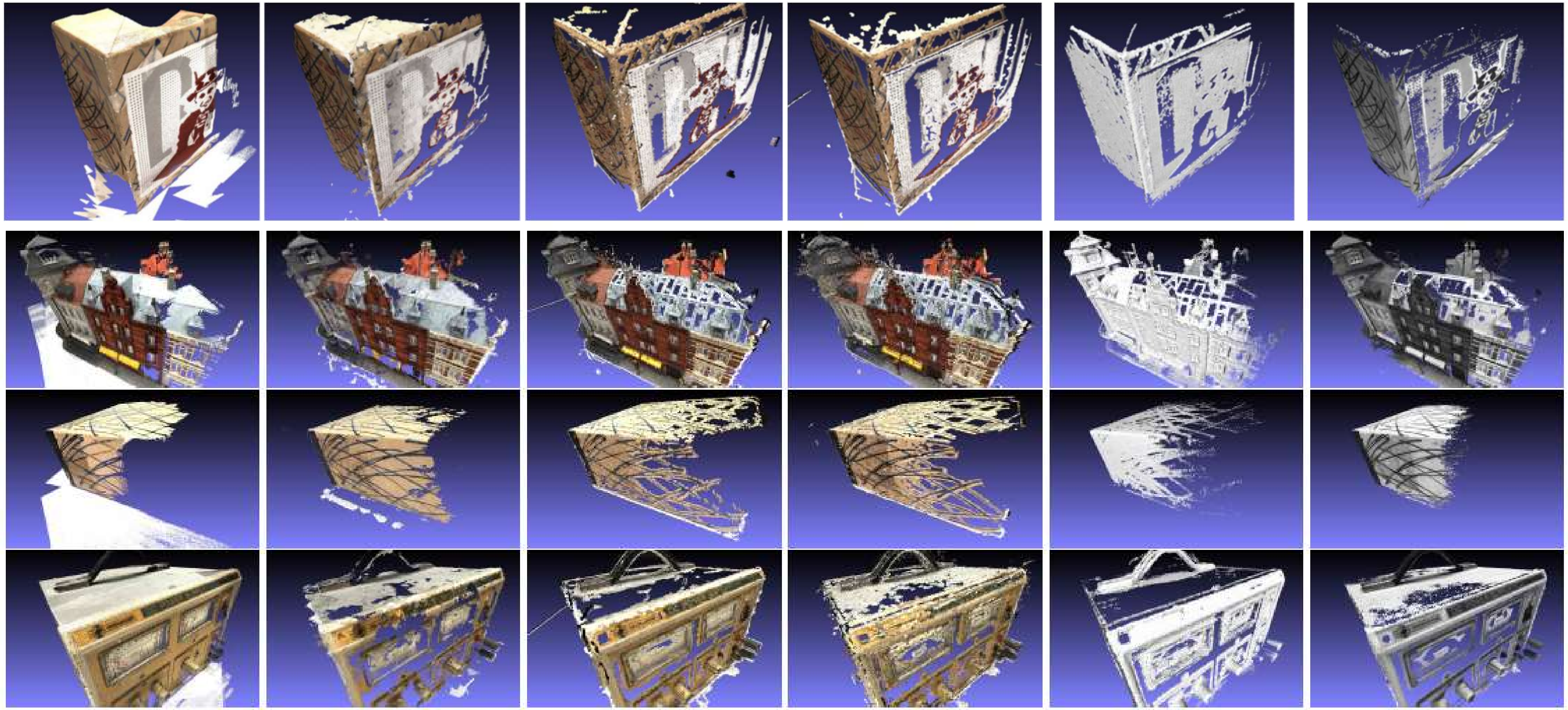


HKUST

Contents

HKUST

Experiments: Compare with others



a: reference model b: **SurfaceNet** c: *camp* [4] d: *furu* [8] e: *tola* [27] f: *Gipuma* [9]

[3] N. D. Campbell, G. Vogiatzis, C. Hern´andez, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In European Conference on Computer Vision, pages 766–779. Springer, 2008.

[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust mul-tiview stereopsis. IEEE transactions on pattern analysis and machine intelligence, 32(8):1362–1376, 2010.

[8] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, pages 873–873–881, 2015.

[24] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications, pages 1–18, 2012.

Experiments: Compare with others

- The structured output surface leads better completeness around the less textured regions compared with other methods.
- *SurfaceNet* outperforms *camp* [3] and *furu* [7]
- It's comparable to *tola* [24] and *Gipuma* [8].

Table 3: Comparison with other methods. The results are reported for the test set consisting of 22 models.

methods (mm)	mean acc	med acc	mean compl	med compl
camp [4]	0.834	0.335	**0.987**	**0.108**
furu [8]	0.504	0.215	1.288	0.246
tola [27]	0.318	0.190	1.533	0.268
Gipuma [9]	**0.268**	0.184	1.402	0.165
$s = 32, \tau = 0.7, \gamma = 0\%$	1.327	0.259	1.346	0.145
$s = 32, \tau = 0.7, \gamma = 80\%$	0.779	0.204	1.407	0.172
$s = 32$, adapt $\beta = 6, \gamma = 80\%$	0.546	0.209	1.944	0.167
$s = 64, \tau = 0.7, \gamma = 0\%$	0.625	0.219	1.293	0.141
$s = 64, \tau = 0.7, \gamma = 80\%$	0.454	0.191	1.354	0.164
$s = 64$, adapt $\beta = 6, \gamma = 80\%$	0.307	**0.183**	2.650	0.342

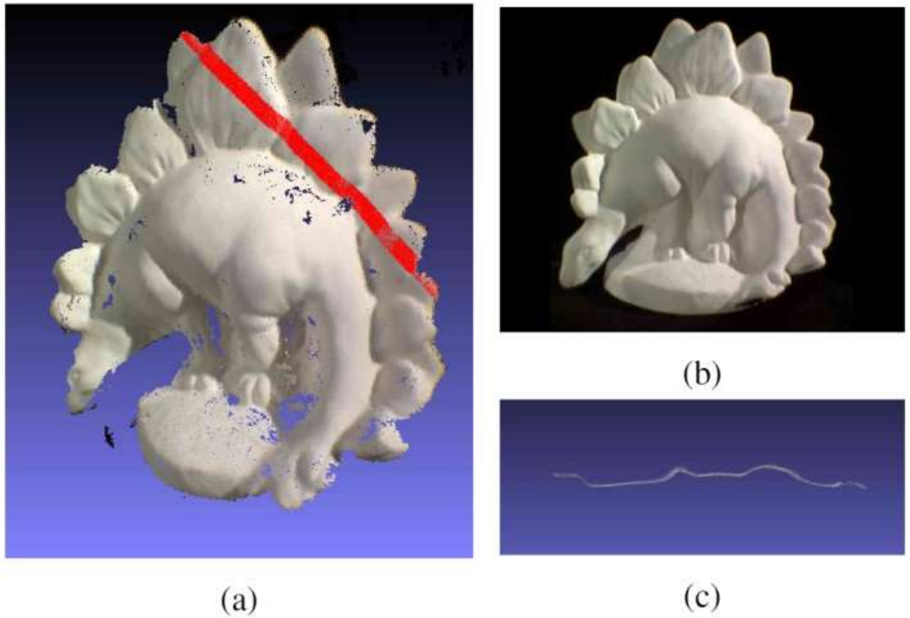


Figure 9: (a) Reconstruction using only 6 images of the dinoSparseRing model in the Middlebury dataset [21]. (b) One of the 6 images. (c) Top view of the reconstructed surface along the red line in (a).

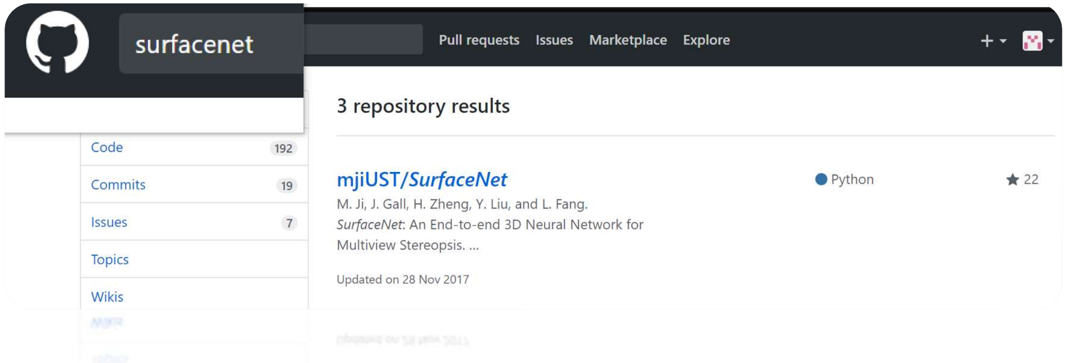
[3] N. D. Campbell, G. Vogiatzis, C. Hern´andez, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In European Conference on Computer Vision, pages 766–779. Springer, 2008.
[7] Y. Furukawa and J. Ponce. Accurate, dense, and robust mul-tiview stereopsis. IEEE transactions on pattern analysis and machine intelligence, 32(8):1362–1376, 2010.
[8] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the IEEE International Conference on Computer Vision, pages 873–881, 2015.
[24] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications, pages 1–18, 2012.

HKUST

Contents

Conclusion

- Presented the first end-to-end learning framework for MVS.
- To effectively encode the camera parameters, we introduced a novel representation: *colored voxel cubes* for each viewpoint.
- Our qualitative and quantitative evaluation on the DTU dataset demonstrated that our network can accurately reconstruct the surface of 3D objects. While our method is currently comparable to the state-of-the-art.
- 3D reconstruction stages
  - Manual labor → analog implementation → silhouettes projection method → graph cut era → depth fusion → deep learning era



Q&A