Skoltech

Skolkovo Institute of Science and Technology

Dense 2D Range-Images

Geometric Computer Vision

GCV v2021.1, Module 3

Alexey Artemov, Spring 2021

Lecture Outline

Skoltech
Skolkovo Institute of Science and Technology

§1. Range-images as a basic 3D modality [5 min]

1.1. Range-images as a result of scanning

1.2. Types and properties of range-images

§2. Representing a moving scene [external]

2.1. Rigid body motion

2.2. Camera motion, camera pose (extrinsic parameters)

Lecture Outline

Skoltech
Skolkovo Institute of Science and Technology

§3. Perspective projection [external]

3.1. The geometry of perspective projection

3.2. Camera intrinsic parameters

§4. Modeling dense range-images [5 minutes]

4.1. Raycasting

§5. Learning with dense range-images [10 minutes]

5.1. Single depth image: super-resolution, smoothing, …

5.2. Multiple depth images: geometric feature prediction

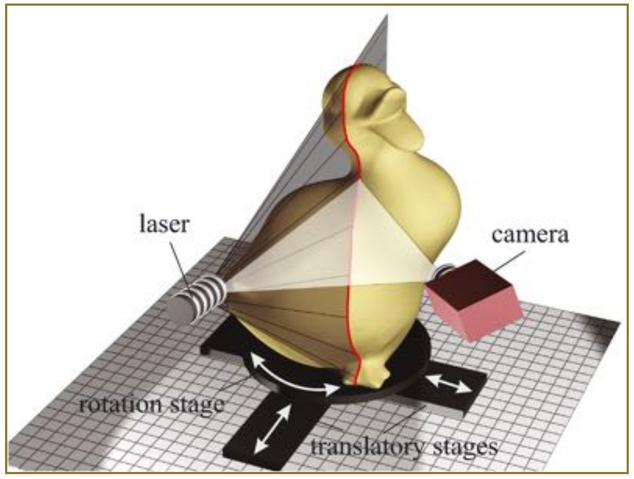
§1. Range-images as a basic 3D modality

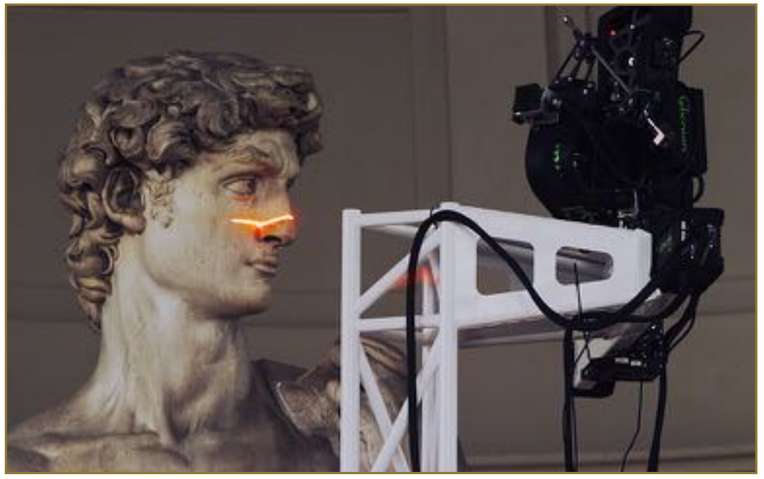
Range-images as a result of scanning

1.1. Range-images as a result of scanning

§1. Range-images as a basic 3D modality

- Captured directly by a large number of 3D sensors:
  - Laser stripes, structured light setups, time-of-flight cameras
- Reconstructed from images by stereo matching, etc.
  - Dense binocular stereo, active stereo, shape-from-X





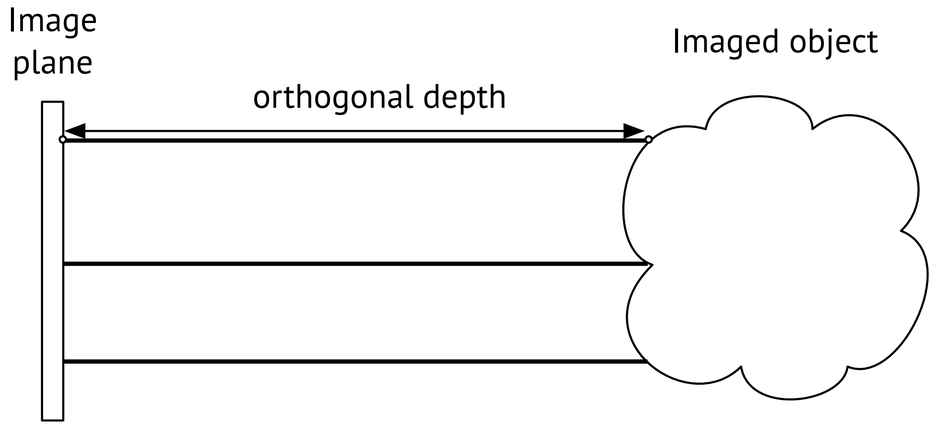
Types and properties of range-images

1.2. Types of range-images

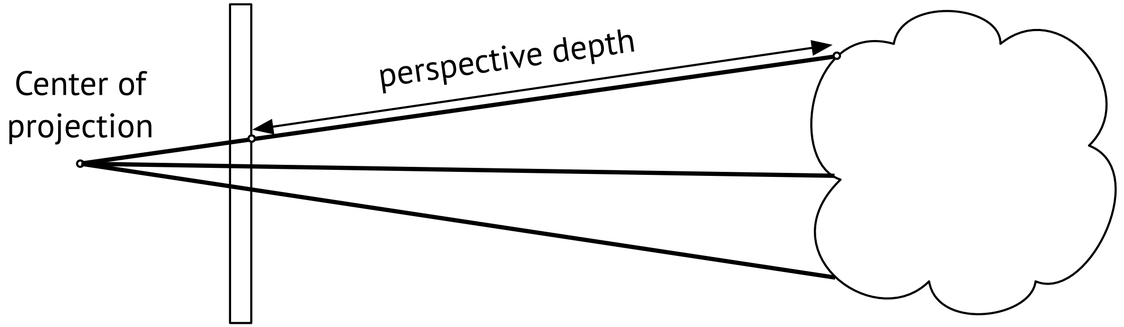
Skoltech
Skolkovo Institute of Science and Technology

§1. Range-images as a basic 3D modality

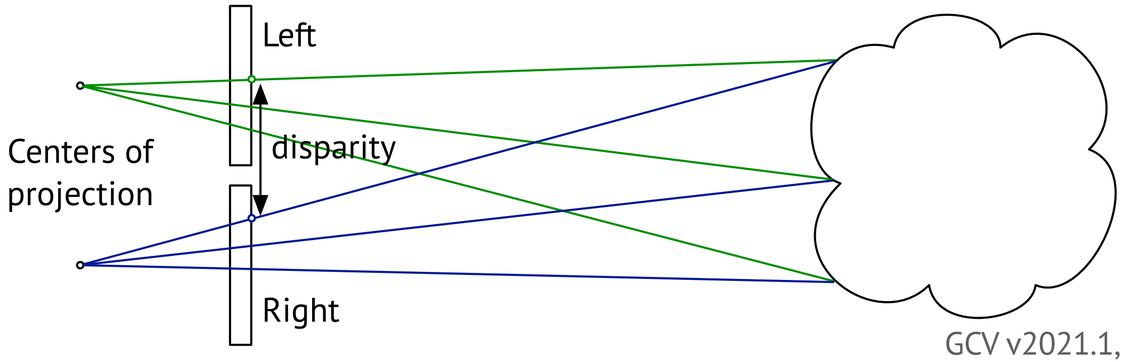
- Orthogonal depth



- Perspective depth



- Disparity: pixel space (left to right)
  - Inversely proportional to depth



1.2. Properties of range-images

Skoltech
Skolkovo Institute of Science and Technology

§1. Range-images as a basic 3D modality

- Comparatively low resolution compared to RGB photo-images
  - Intel RealSense D435 (real-time active stereo, 30 FPS):
    depth at 640x480 and RGB at 1920x1080
  - MS Kinect v2 (real-time structured light, 30 FPS):
    depth at 512x424 and RGB at 1920x1080
- **Can we obtain depth at the same spatial resolution as RGB images?**

1.2. Properties of range-images

Skoltech
Skolkovo Institute of Science and Technology

§1. Range-images as a basic 3D modality

- High amounts of noise and outliers, relatively low accuracy
  - MS Kinect v2 (real-time structured light, 30FPS): depth accuracy 2.5cm @2m
  - Commonly precision and accuracy proportional to depth value
- **Can we obtain cleaner depth images?**

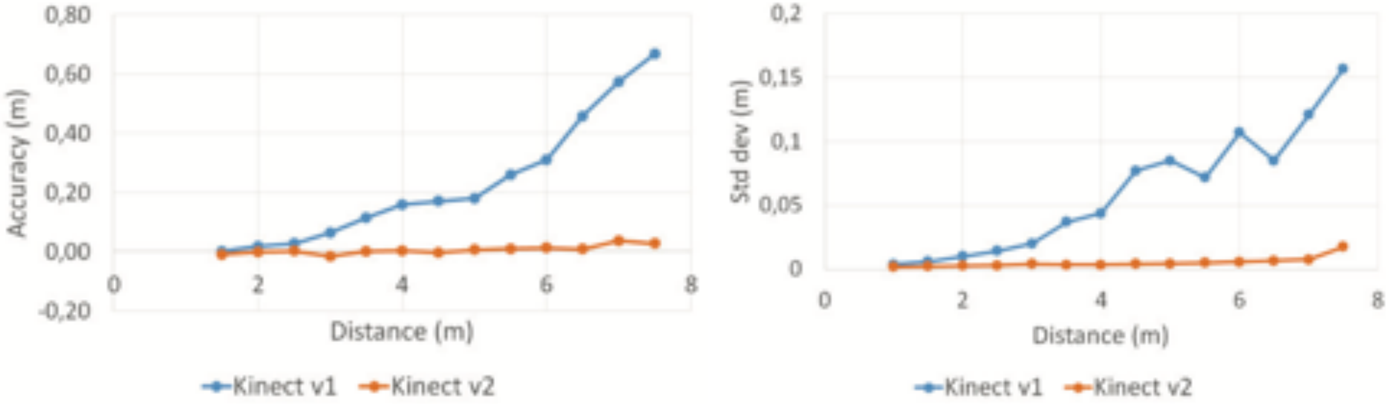
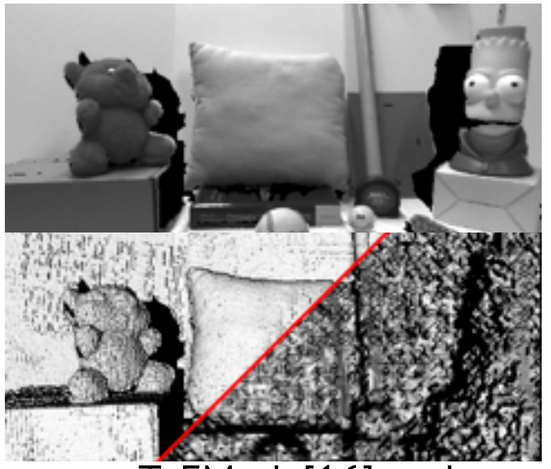


Figure Credit: (Zennaro et al., 2015)

Fig. 9. a) Depth accuracy and b) depth standard deviation of Kinect v1 and v2 at all ranges.

SUN RGBD [15] low-quality real



ToFMark [16] real HQ high-res + super LQ low-res

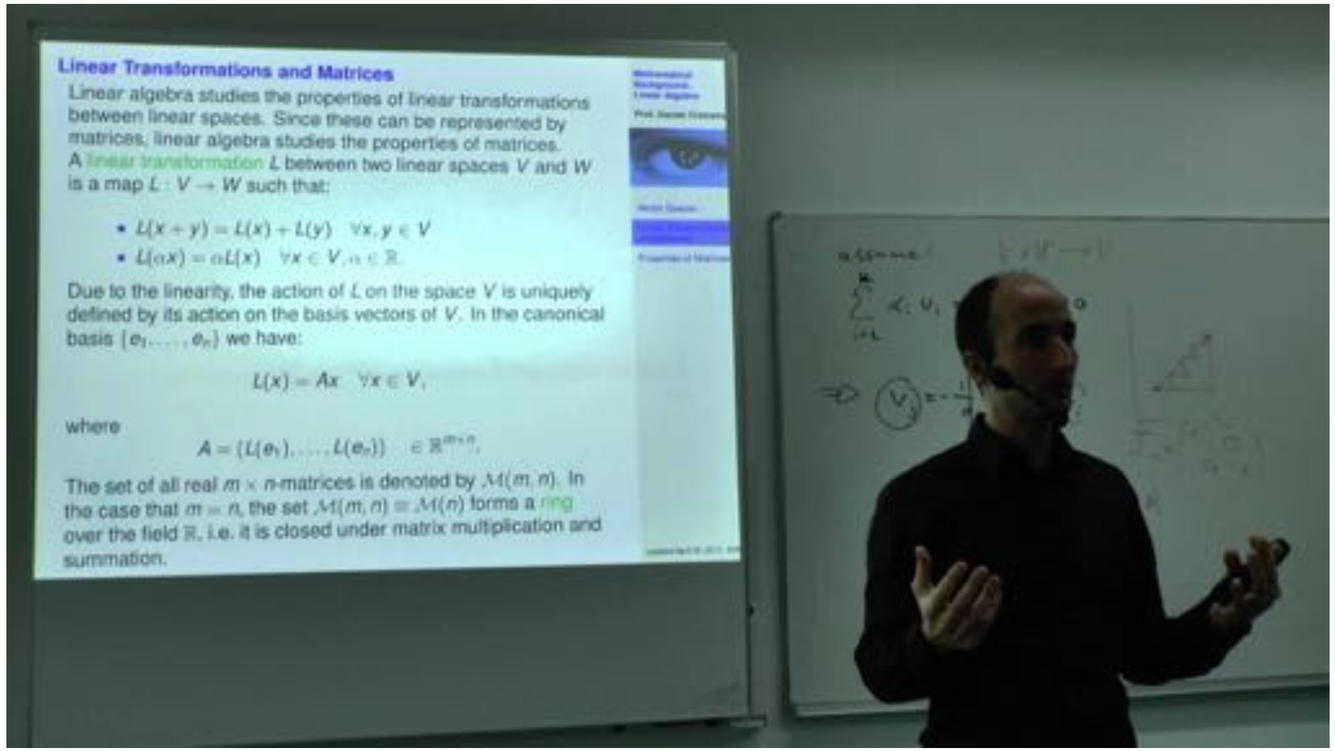
§2. Representing a moving scene

A lecture on representing a moving scene

Skoltech Skolkovo Institute of Science and Technology

§2. Representing a moving scene

- Prof. D. Cremers: Multiple View Geometry (YouTube playlist with 14 lectures on SfM)



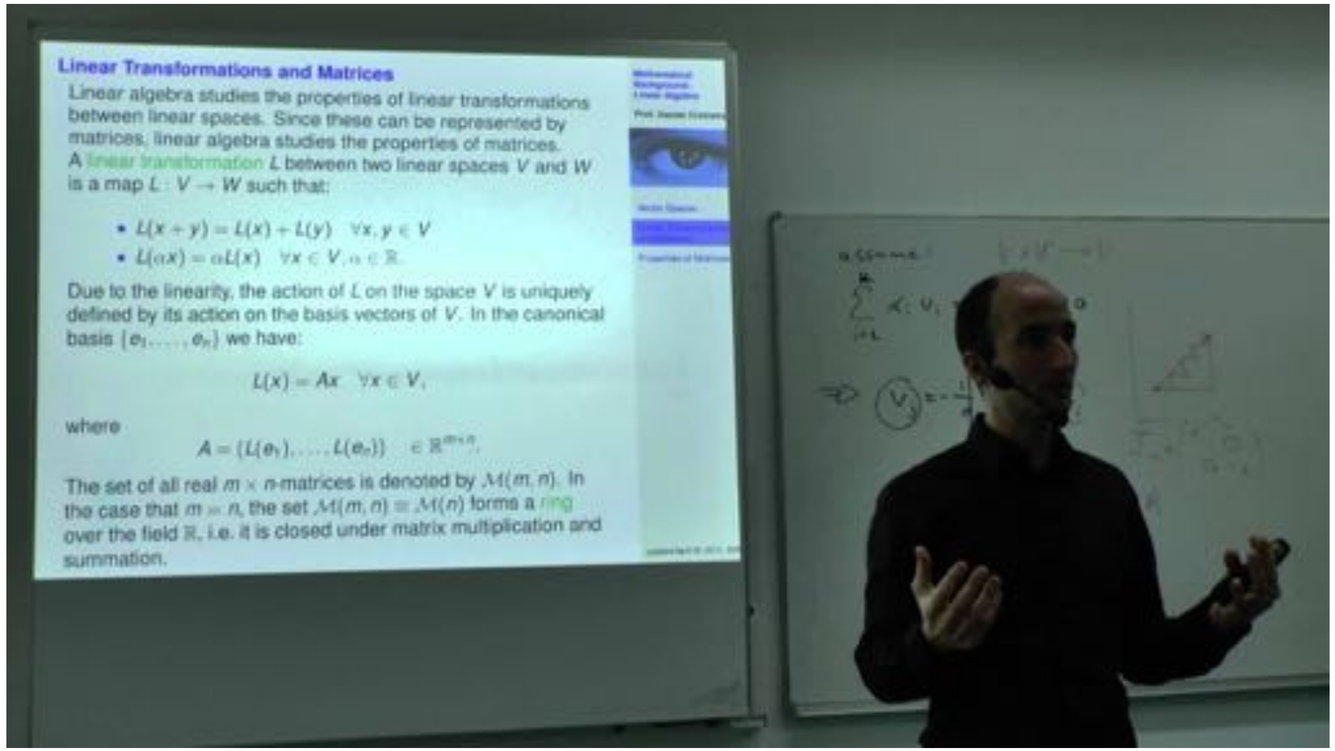
§3. Perspective projection

A lecture on perspective projection

Skoltech
Skolkovo Institute of Science and Technology

§3. Perspective projection

- Prof. D. Cremers: Multiple View Geometry (YouTube playlist with 14 lectures on SfM)



§4. Modeling dense range-images

Ray-casting

4.1. Ray-casting

Skoltech
Skolkovo Institute of Science and Technology

§4. Modeling dense range-images

- Given a 3D scene, compute *(render)* a 2D image of the scene
- This course: model 2D range-images (learning with 3D shapes)
- Basic rendering technique: *ray-casting*
- Created as early as during the late 1970s to replace wireframe renderings of solids

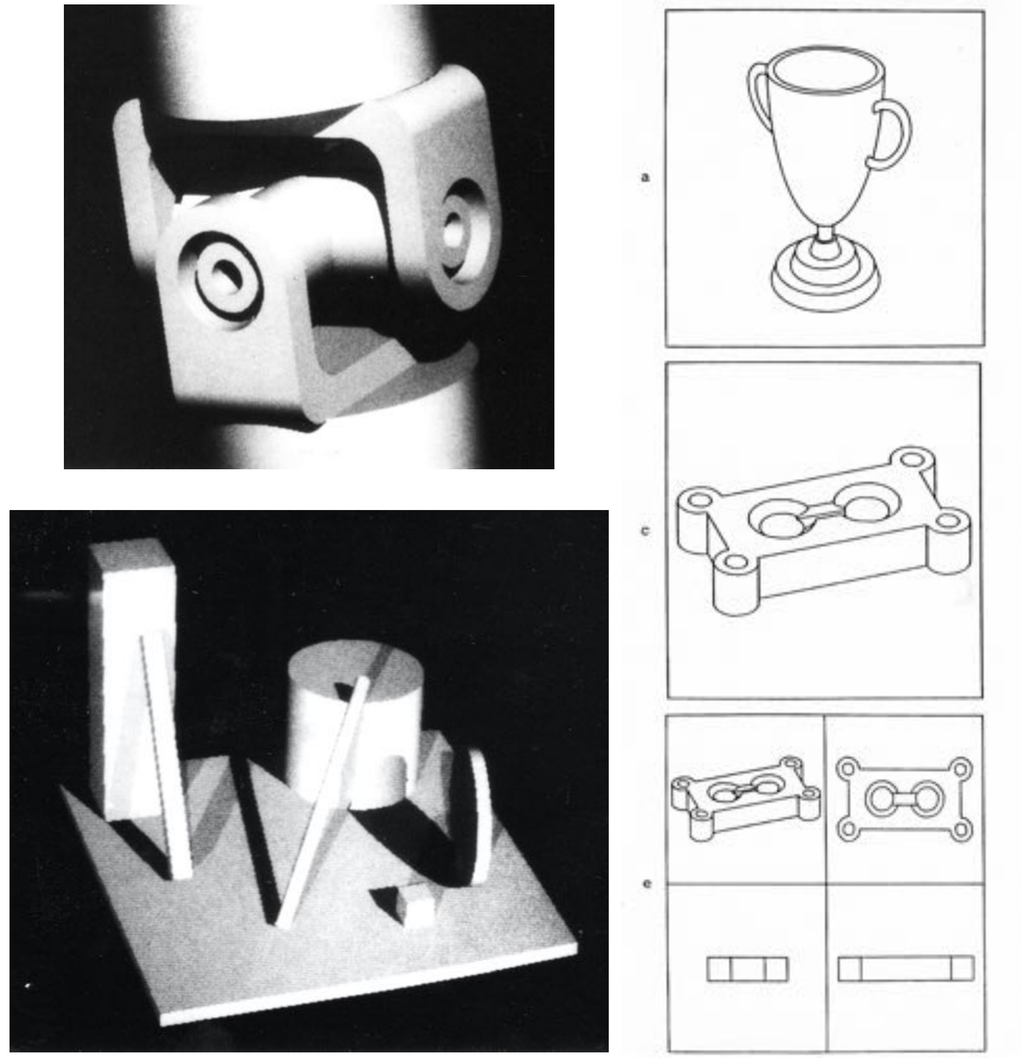


Figure Credit: Wikipedia

4.1. Ray-casting

Skoltech
Skolkovo Institute of Science and Technology

§4. Modeling dense range-images

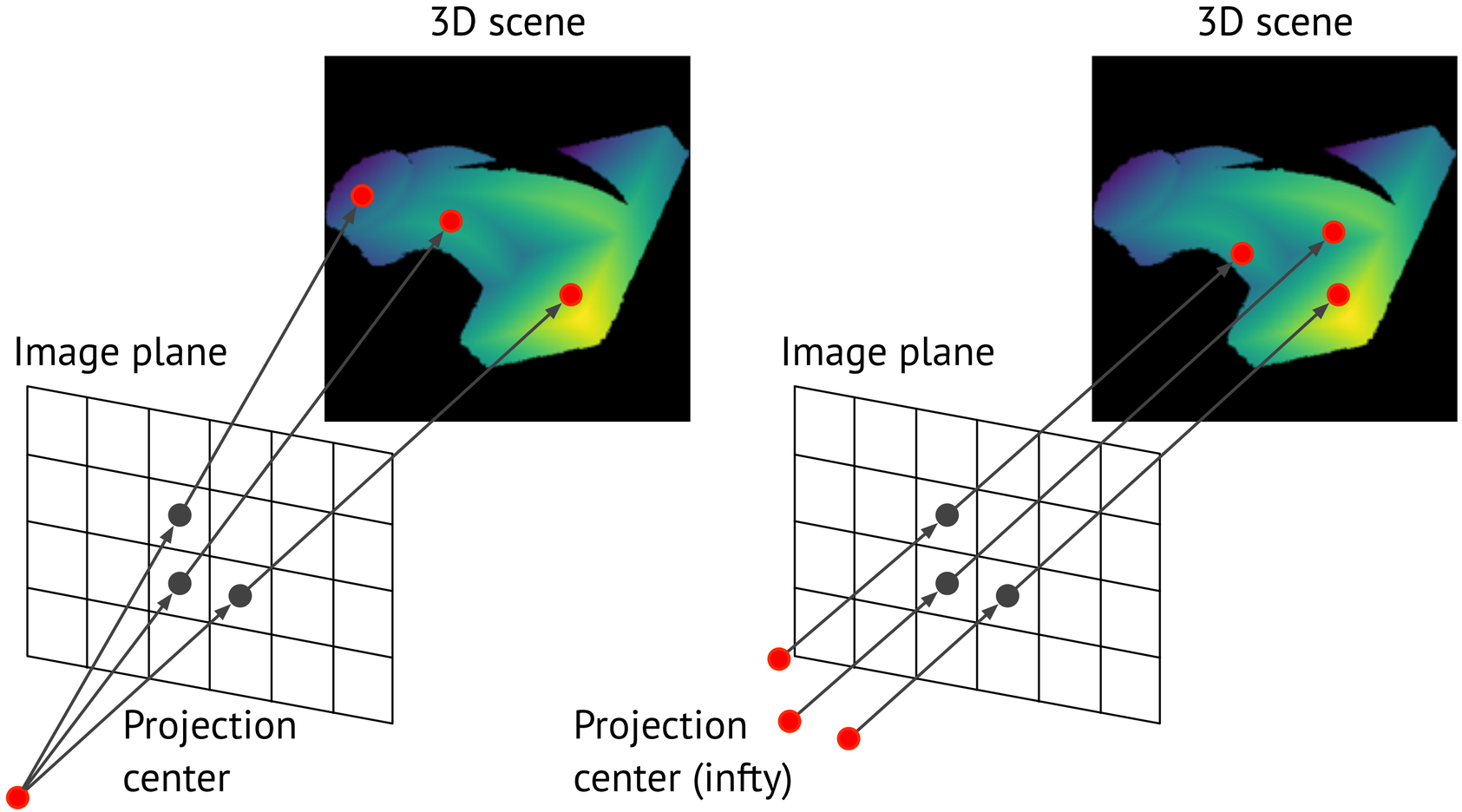
- *Cast* geometric *rays* from the eye though each pixel in the image
- Compute locations of intersection between the ray and each surface point along its path
- Find the closest object blocking the path of each ray
  - Other applications: wireframe drawing, compute volumes, …

4.1. Ray-casting: overview

Skoltech
Skolkovo Institute of Science and Technology

§4. Modeling dense range-images

1. Generation of rays (one per pixel)
2. Intersection with objects in the scene
3. *(For graphics applications)* Shading (aka compute the color of the pixel)



4.1. Ray-casting: intersections

Skoltech
Skolkovo Institute of Science and Technology

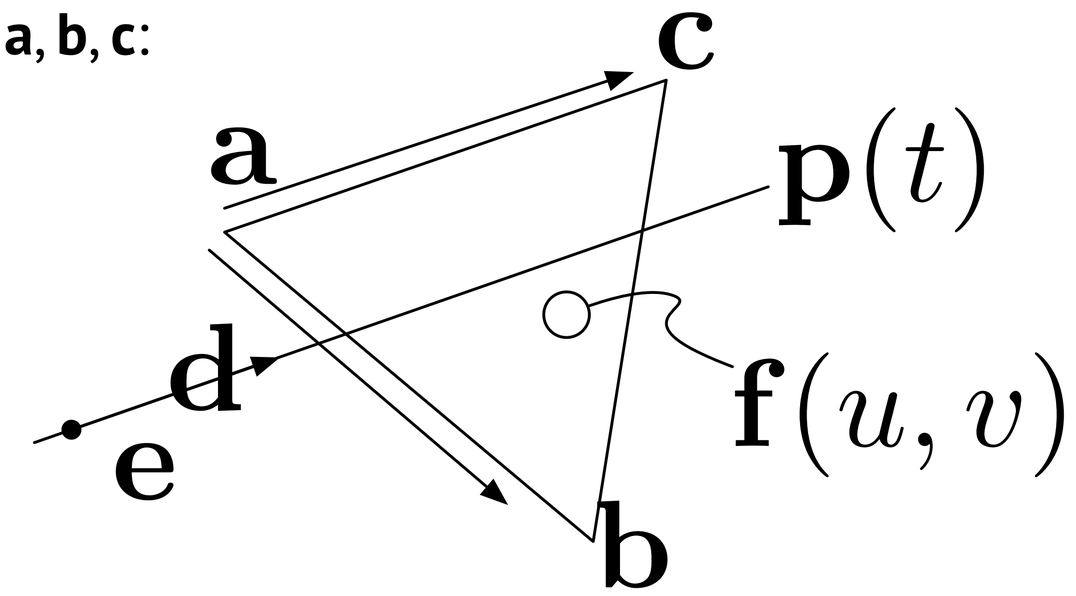
§4. Modeling dense range-images

- This is an expensive operation
- There is a large literature, a good overview is here: http://www.realtimerendering.com/intersections.html
- We will study the one useful case:
  - Triangles (by combining many triangles you can approximate complex surfaces)

4.1. Ray-casting: ray-triangle intersection

§4. Modeling dense range-images

- Explicit parametrization of a triangle with vertices **a**, **b**, **c**:
  $\mathbf{f}(u, v) = \mathbf{a} + u(\mathbf{b} - \mathbf{a}) + v(\mathbf{c} - \mathbf{a})$
- Explicit ray:
  $\mathbf{p}(t) = \mathbf{e} + t\mathbf{d}$
- The ray intersects the triangle if a **t**, **u**, **v** exist s.t.:
  $$\mathbf{f}(u, v) = \mathbf{p}(t),$$
  $$t > 0, \quad 0 \leqslant u, v, \quad u + v \leqslant 1$$
- Solve by yourself to identify appropriate **t**, **u**, **v**



4.1. Ray-casting: intersections

Skoltech
Skolkovo Institute of Science and Technology

§4. Modeling dense range-images

- It is simple, intersect it with all of them and only keep the closest intersection
- To speed up computation, you can use a spatial data structure to prune the number of collisions that you need to check

- Bounding volume hierarchy:
  - compute bounding 3D boxes (e.g. AABBs) on sets of objects, group hierarchically
  - quickly discard objects from ray-box intersection checks
- Used by Intel Embree CPU impl.

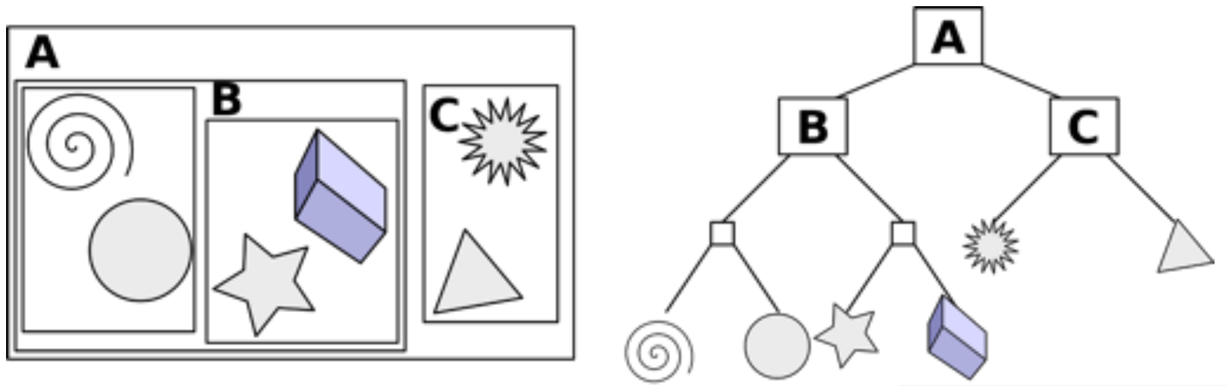


Figure Credit: Wikipedia

Bounding volume hierarchy

Slide Credit: Daniele Panozzo GCV v2021.1, Module 3

4.1. Ray-casting: intersections

Skoltech
Skolkovo Institute of Science and Technology

§4. Modeling dense range-images

- Ray-casting: trace rays from the “eye” of the observer into the scene
- Compute intersection locations with scene geometry
- Prune parts of geometry for efficiency

Slide Credit: Daniele Panozzo

§5. Learning with dense range-images

Single depth image

5.1. Single depth image

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- Range-images: low spatial resolution, sparsity / holes, high noise
- RGB images: high resolution, fully dense, can be made nearly noiseless

- **Can we obtain depth at the same spatial resolution as RGB images?**
- **Can we achieve high degree of noise-suppression in depth images?**
- Common tasks: scene understanding from RGBD (classification, segmentation, retrieval…)

5.1. Single depth image: tasks

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- **Upsampling/super-resolution**
  - Goal: obtaining depth at a higher spatial resolution
  - Input: low-resolution depth (+possibly high-resolution RGB)
  - Output: high-resolution depth

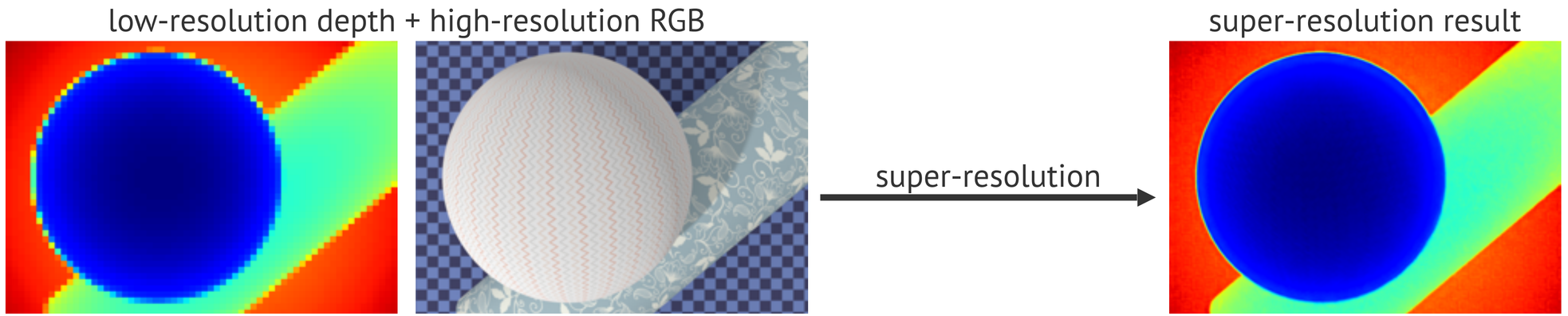


Figure Credit: Voynov et al., 2019

5.1. Single depth image: tasks

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- **Densification / completion**
  - Goal: obtaining dense depth from sparse samples
  - Input: sparse depth (+possibly high-resolution RGB)
  - Output: dense depth (desirably: high-resolution)

low-resolution depth + high-resolution RGB

densification

densification result

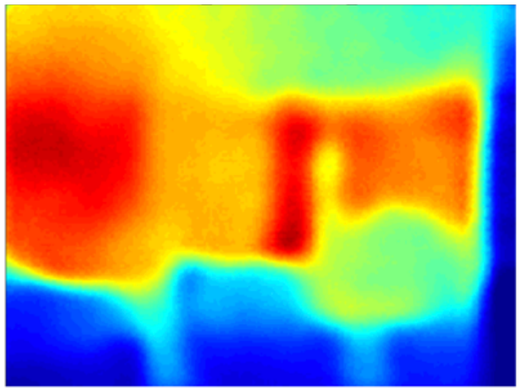


Figure Credit: Ma et al., 2018

5.1. Single depth image: tasks

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- **Denoising / enhancement**
  - Goal: reducing noise in the depth measurements, depth interpolation
  - Input: noisy/coarse depth (+possibly high-resolution RGB)
  - Output: denoised depth

high-resolution RGB + noisy / quantized depth

denoising

denoising result

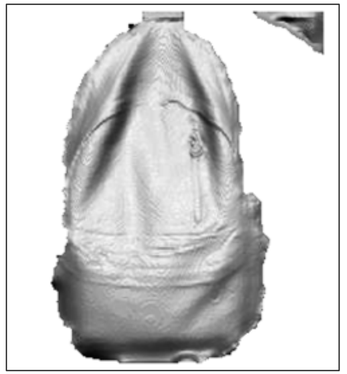


Figure Credit: Yan et al., 2018

5.1. Single depth image: overview

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- Challenges:
  - Fine structure ambiguity, texture transfer
- Visual cues leveraged in depth processing tasks:
  - RGB guidance, edge-awareness, multi-scale structure/guidance, visual loss functions
- Strategies:
  - Upsample + refine,

5.1. Single depth image: challenges

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

RGB | GT | 8x low-res | SRCNN [1] | MSG-Net [2]

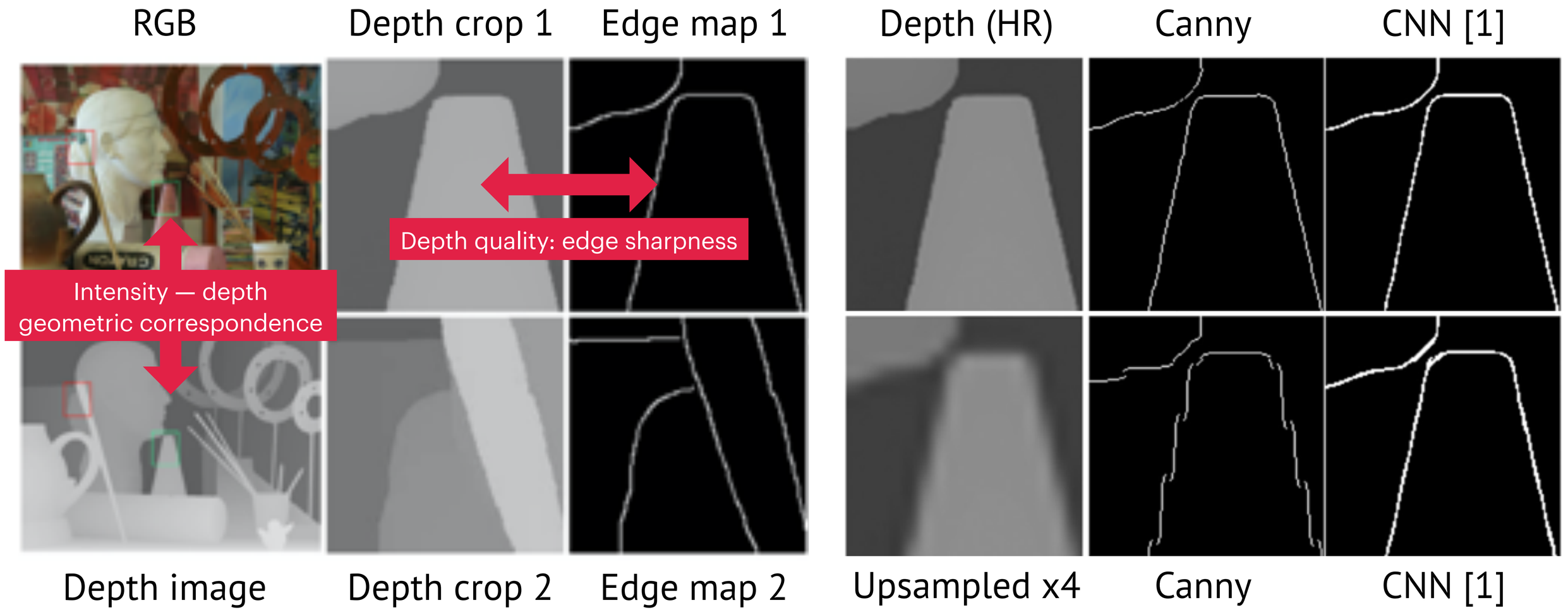
- High-res -> low-res: distortion of fine structures
- Ambiguity in super-resolving fine structures (blurring of the estimated depth map)

Figure Credit: MSG-Net

5.1. Single depth image: visual cues

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images



5.1. Single depth image: challenges

Skoltech
Skolkovo Institute of Science and Technology

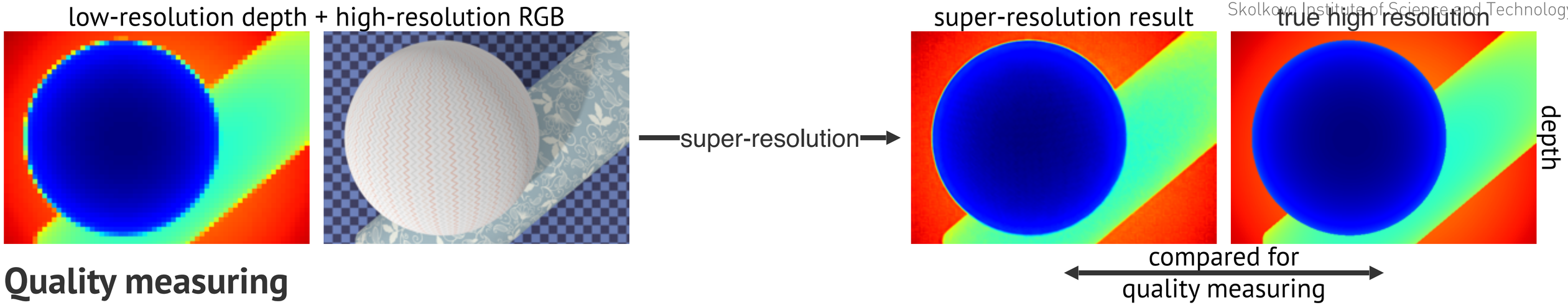
§5. Learning with dense range-images

- Use intensity (RGB) image to guide upsampling
- Texture artefacts (RGB features transferred into depth images)

Figure Credit: MSG-Net

5.1. Single depth image: visual cues

Skoltech
Skolkovo Institute of Science and Technology

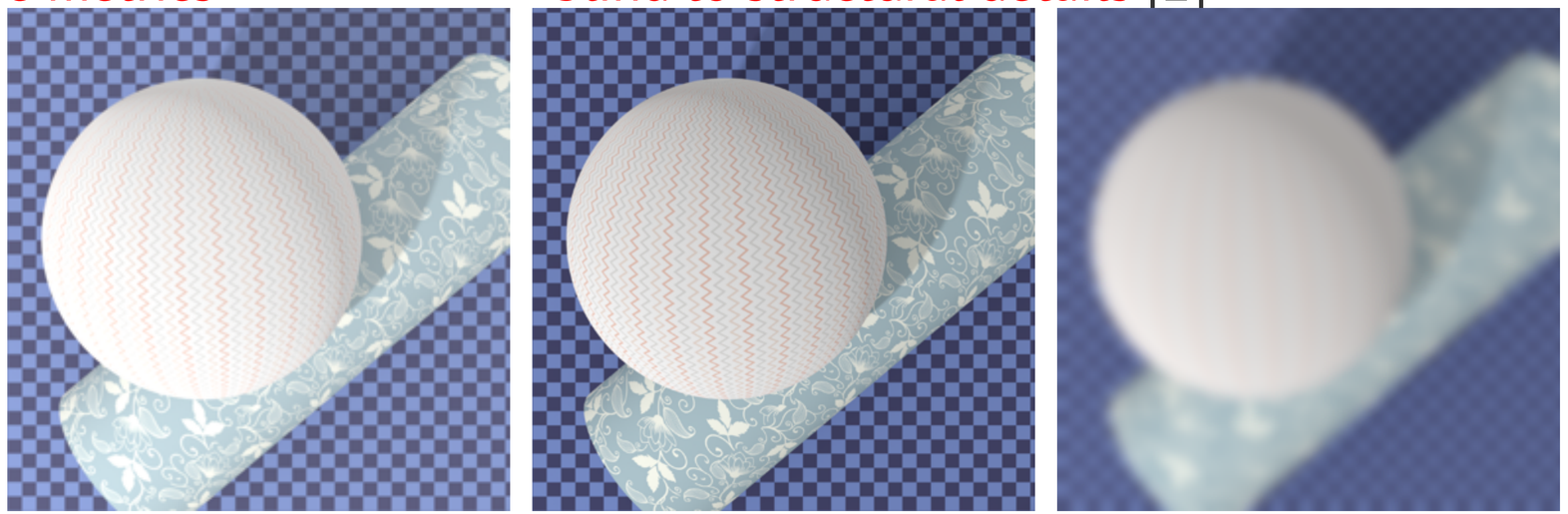


Quality measuring

Common approach: compare depth maps directly with pixel-wise metrics — blind to structural details [2]

Example for photos:

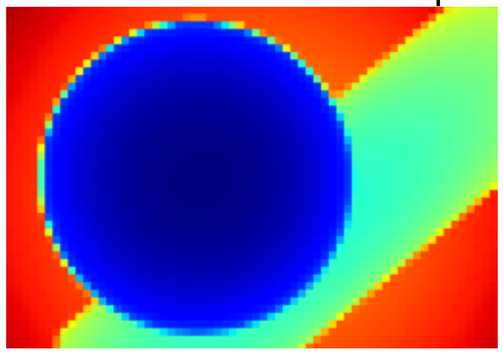
the brighter version is farther from the reference in a pixel-wise metric than the blurred version

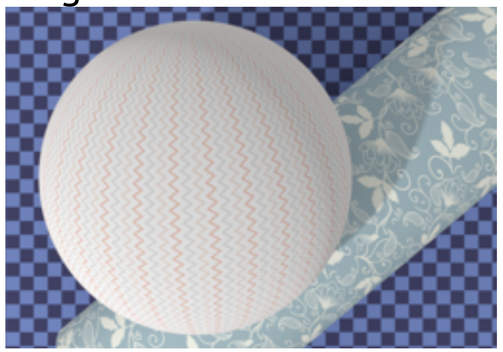


[2] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. IEEE Conference on Computer Vision and Pattern Recognition

5.1. Single depth image: visual cues

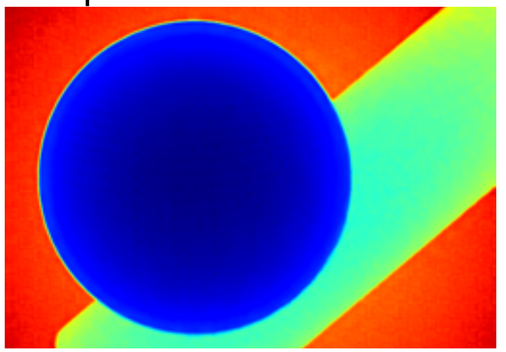
low-resolution depth + high-resolution RGB



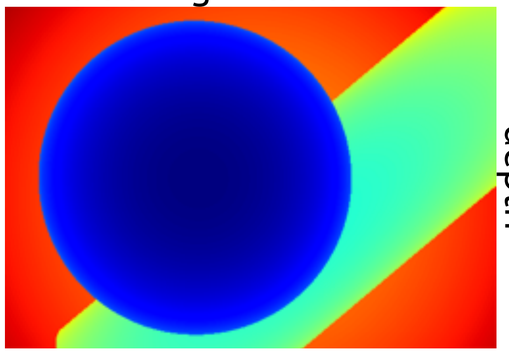


super-resolution

super-resolution result



true high resolution



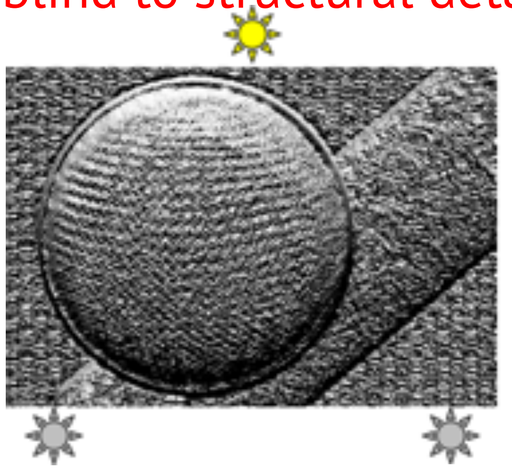
depth

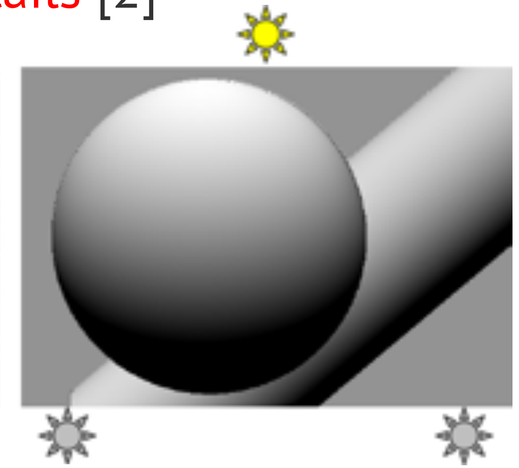
almost identical

Quality measuring

Common approach: compare depth maps directly — blind to surface artefacts [3, 4]
with pixel-wise metrics — blind to structural details [2]

Proposed approach: compare surface visualizations
with perceptual metrics [2]





surface visualization

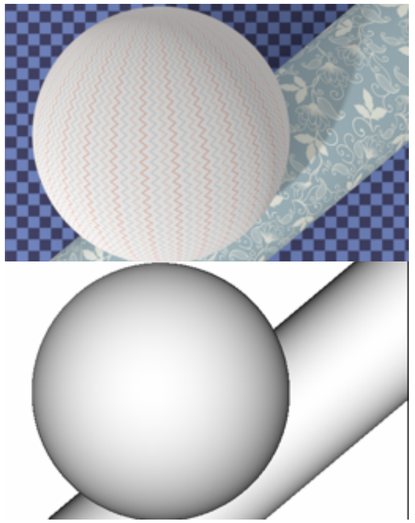
completely different

[2] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. IEEE Conference on Computer Vision and Pattern Recognition
[3] Mechrez, R., Talmi, I., Shama, F., & Zelnik-Manor, L. (2019). Maintaining Natural Image Statistics with the Contextual Loss. Asian Conference on Computer Vision
[4] Xie, J., Feris, R. S., & Sun, M. (2016). Edge-Guided Single Depth Image Super Resolution. IEEE Transactions on Image Processing

5.1. Single depth image: datasets

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

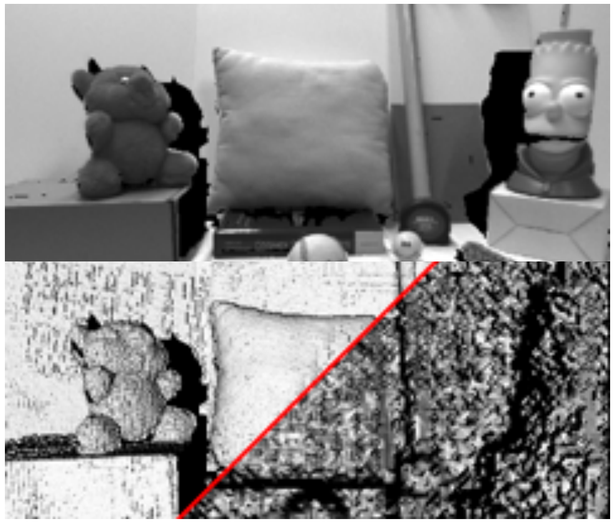


SimGeo
simple synthetic

ICL-NUIM [13]
complex synthetic

Middlebury [14]
high-quality real

SUN RGBD [15]
low-quality real



ToFMark [16] real
HQ high-res + super LQ low-res

- A variety of datasets with range-images:
  - Differ by acquisition method/device, ground-truth, type of scene

[13] Handa, A., Whelan, T., McDonald, J., & Davison, A. J. (2014). A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. IEEE International Conference on Robotics and Automation
[14] Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., & Westling, P. (2014). High-resolution stereo datasets with subpixel-accurate ground truth. German Conference on Pattern Recognition
[15] Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. IEEE Conference on Computer Vision and Pattern Recognition
[16] Ferstl, D., Reinbacher, C., Ranftl, R., Rüther, M., & Bischof, H. (2013). Image guided depth upsampling using anisotropic total generalized variation. IEEE International Conference on Computer Vision

5.1. Single depth image: recap

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- Deal with range-image acquisition artefacts (low resolution, holes, sparsity…)
- Commonly use RGB intensity image as guidance
- Visual depth image similarity assessment

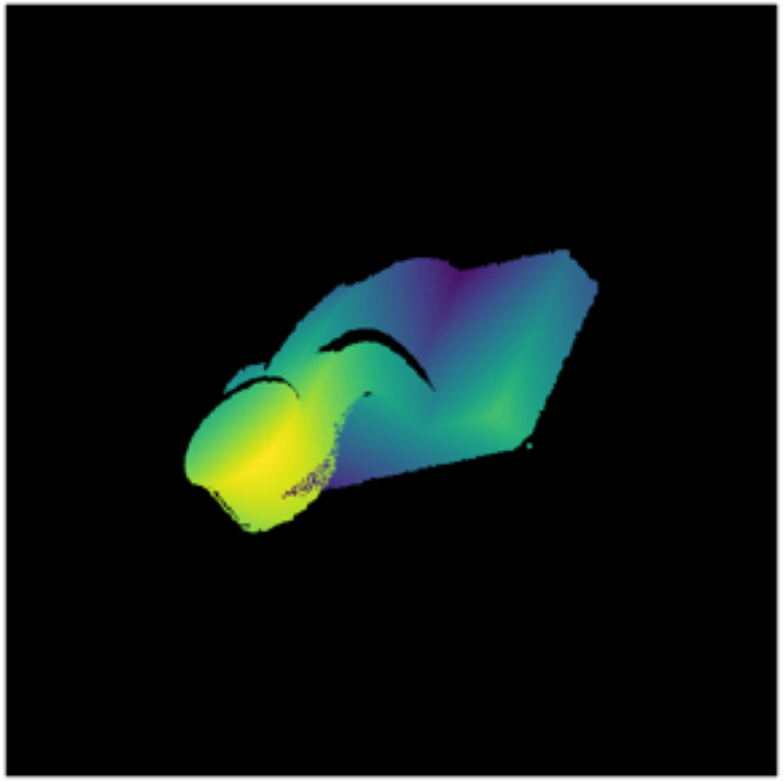
Multiple depth images

5.2. Why multiple depth images?

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

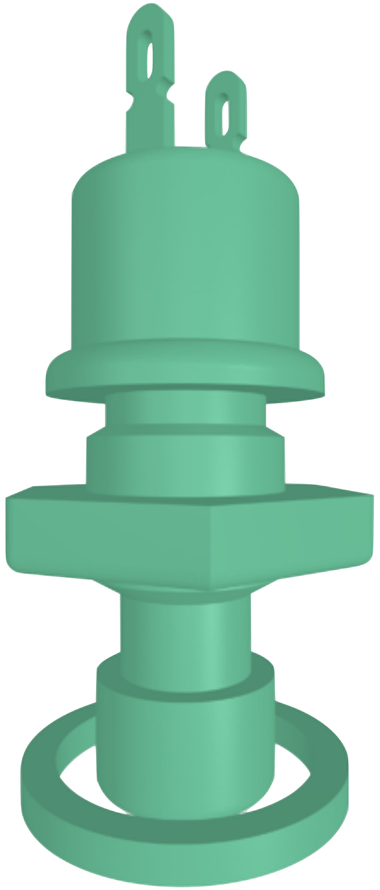
- Range-images: acquired from different (possibly intersecting) views
  - Multiple views provide complementary contexts
  - “Freely” available with standard scanning sequences (e.g. RGBD video, structured light)
- Many applications as range-images are common
  - Depth fusion, feature line estimation, differential quantities (normals / curvature / …), shape segmentation…



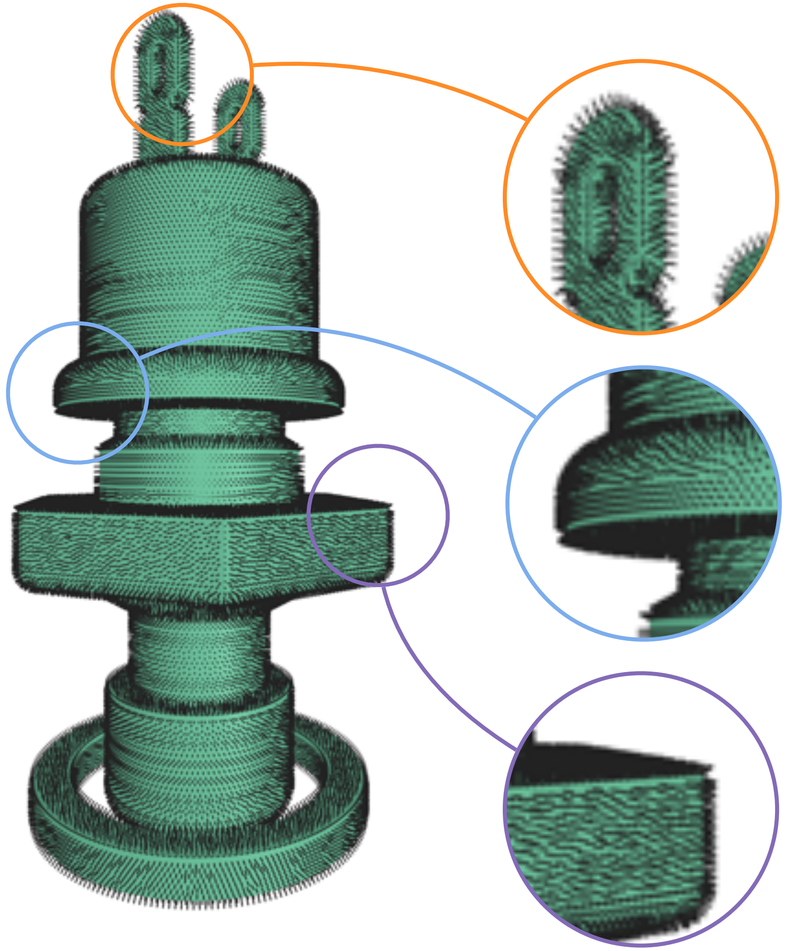
Input: multiple-view range-images

5.2. Multiple depth images: datasets

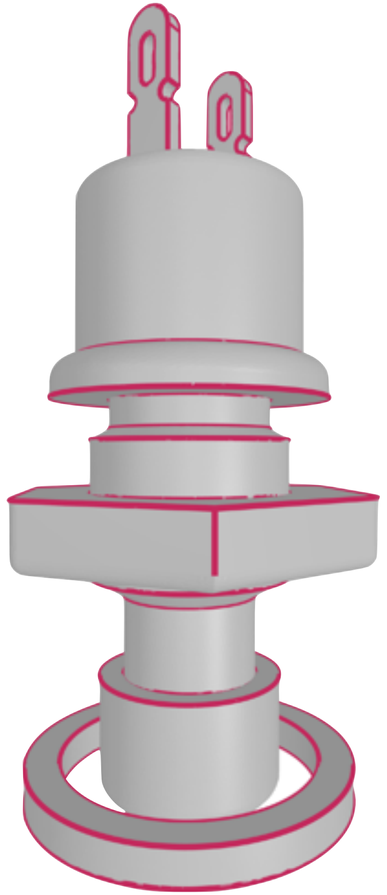
§5. Learning with dense range·



CAD model



Rapid change in normals

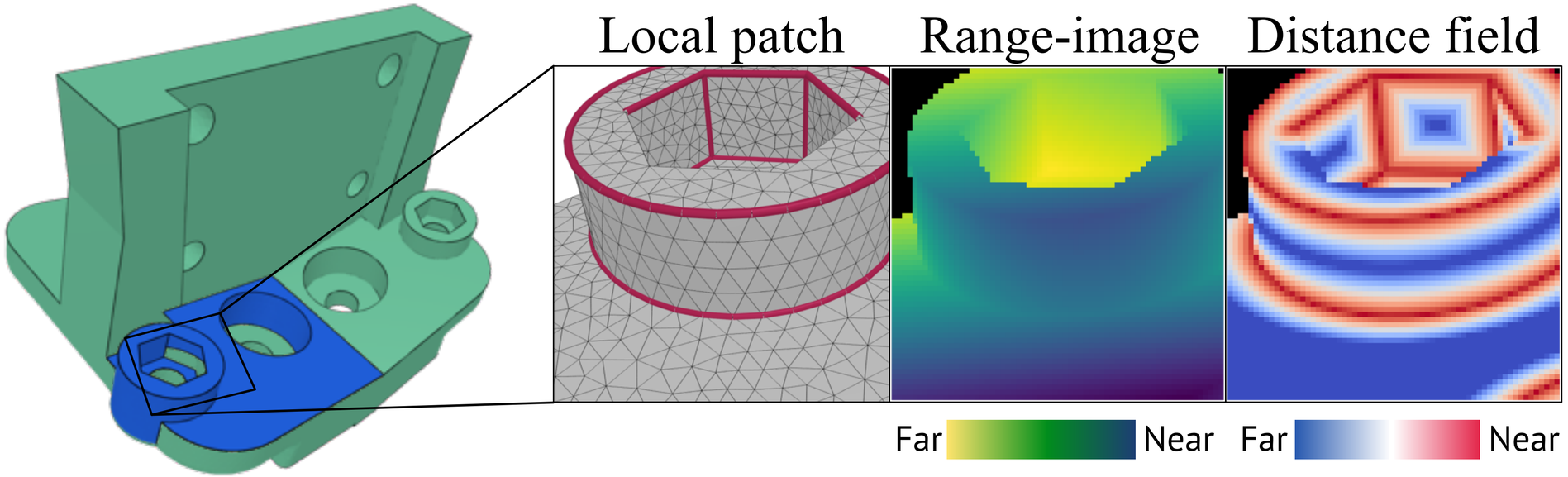


Sharp feature lines

5.2. Multiple depth images: datasets

§5. Learning with dense range-images

3D CAD shapes
+ feature annotations



- Extract local patches → ray-cast and compute range-image → compute distance-to-feature (use closest sharp feature line only)

Figure credit: Matveev, A., Artemov, A., Rakhimov, R., Bobrovskikh, G., Panozzo, D., Zorin, D., & Burnaev, E. (2020). DEF: Deep Estimation of Sharp Geometric Features in 3D Shapes. *arXiv preprint arXiv:2011.15081*.

5.2. Multiple depth images

Skoltech
Skolkovo Institute of Science and Technology

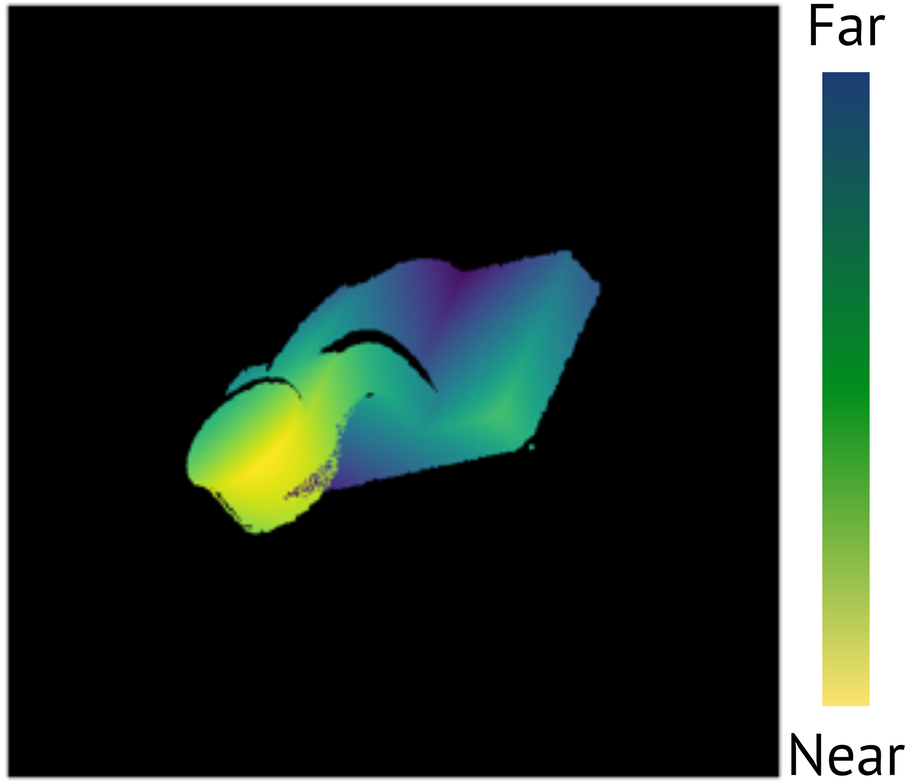
§5. Learning with dense range-images

- Common learning scenarios with range-images:
  - **Input:** a dense range-image (+a camera pose and intrinsic parameters)
  - **Output:** per-pixel quantity of interest (e.g., semantic label, distance to geometric feature curve, curvature, normals, …)
- Commonly approached using CNN machinery
- **Challenge:** occlusions would not generally allow accurately predicting for ambiguous regions (e.g. little context, shape boundaries, …)

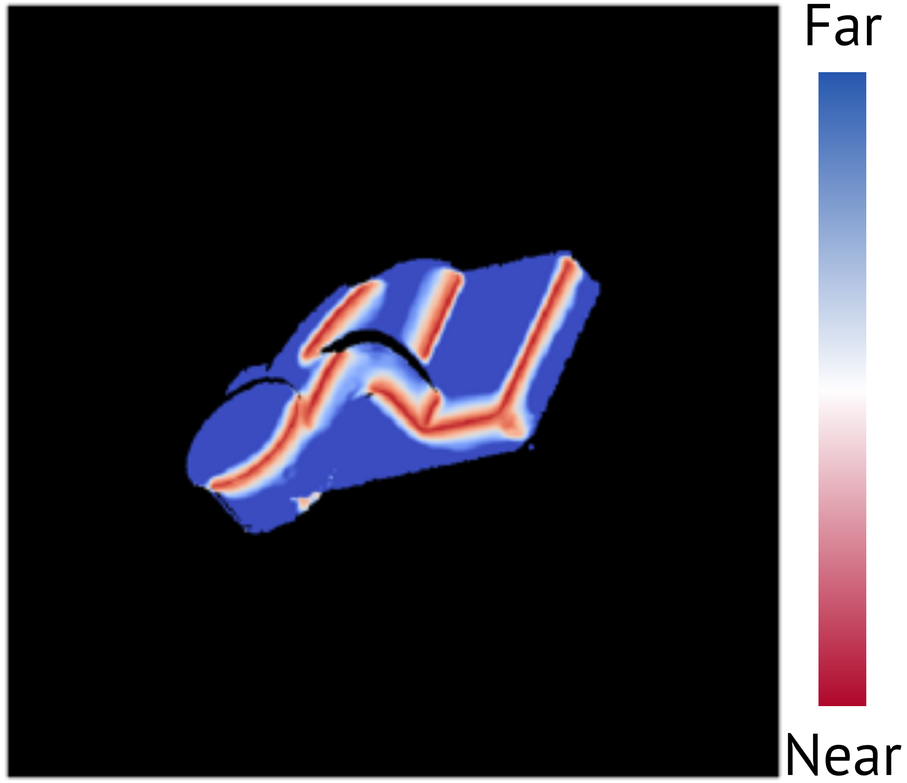
5.2. Multiple depth images

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images



Input: range-image



Output: distance-to-feature

5.2. Multiple depth images

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- Leveraging multiple views is possible with:
  - Naive implementation: **“fuse” predictions** from individual views (practice, homework)
  - More advanced: **multiple view consistency** in the scene during training
- Same e.g. for RGB images (e.g., see “Learning Unsupervised Multi-View Stereopsis via Robust Photometric Consistency”)

5.2. Multiple depth images

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

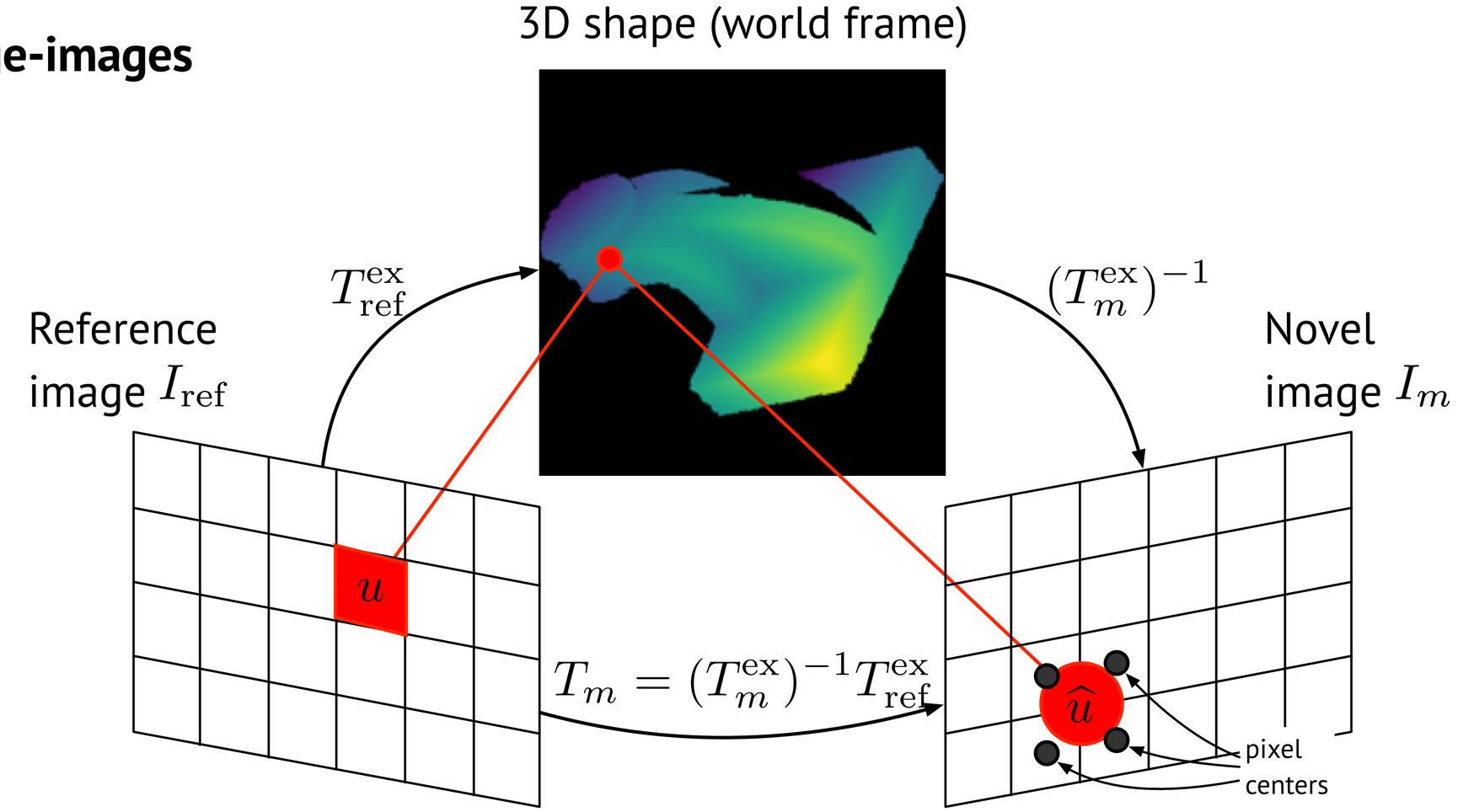
- I_{ref}: a particular (reference) depth image
- $\{I_m\}_{m=1}^M$: additional depth images (views of the same scene)
- $D_{\text{ref}}, \{D_m\}_{m=1}^M$: the quantity of interest (e.g. distance-to-feature annotations)
- u: pixel in I_{ref}, i.e. $I_{\text{ref}}(u)$ represents depth value at u
- K: camera intrinsics (shared across views)
- T_m: relative extrinsic parameters between I_{ref} and I_m

5.2. Multiple depth images

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- 3D coordinates of pixel u in reference image frame:
  $[xyz]_{\text{ref}} = [K^{-1}u; I_{\text{ref}}(u)]$
- Relation between pixel coordinates in two views:
  $\widehat{u} = KT_m[K^{-1}u; I_{\text{ref}}(u)]$
- Note: warped coordinates $\widehat{u}$ need re-sampling on the image grid!



Re-project to 3D + transform from reference to world frame:
$X = T_{\text{ref}}^{\text{ex}}[K^{-1}u; I_{\text{ref}}(u)]$

Transform from world to novel frame + project:
$\widehat{u} = K(T_m^{\text{ex}})^{-1}X$

5.2. Multiple depth images: fusion

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- **Input:** depth images $I_1, \ldots, I_M$;
  predictions $D_1, \ldots, D_M$ obtained for each view independently (e.g. $D_i = \mathrm{CNN}(I_i)$)
- **Goal:** obtaining predictions that are consistent (approximately the same) across views
- For each pair of views (I_{ref}, I_m) (commonly fix I_{ref}, iterate over I_m):
  - Compute novel view pixel coordinates $\widehat{u} = KT_m[K^{-1}u; I_{\mathrm{ref}}(u)]$ for each pixel u in I_{ref}
  - Computed an interpolated prediction in the pixel space: $\widehat{D}^m_{\mathrm{ref}}(u) = \mathrm{Interp}(D_m(\widehat{u}))$
  - Append $\widehat{D}^m_{\mathrm{ref}}(u)$ and a validity mask $\widehat{V}^m_{\mathrm{ref}}(u)$ to the list of predictions in for each pixel u
- Aggregate by minimizing the pixelwise difference in valid pixels

$$\widehat{D}_{\mathrm{ref}} = \arg\min_{\widehat{D}_{\mathrm{ref}}} \sum_{m=1}^{M} ||(\widehat{D}_{\mathrm{ref}} - \widehat{D}^m_{\mathrm{ref}}) \odot \widehat{V}^m_{\mathrm{ref}}||$$

5.2. Multiple depth images: fusion

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

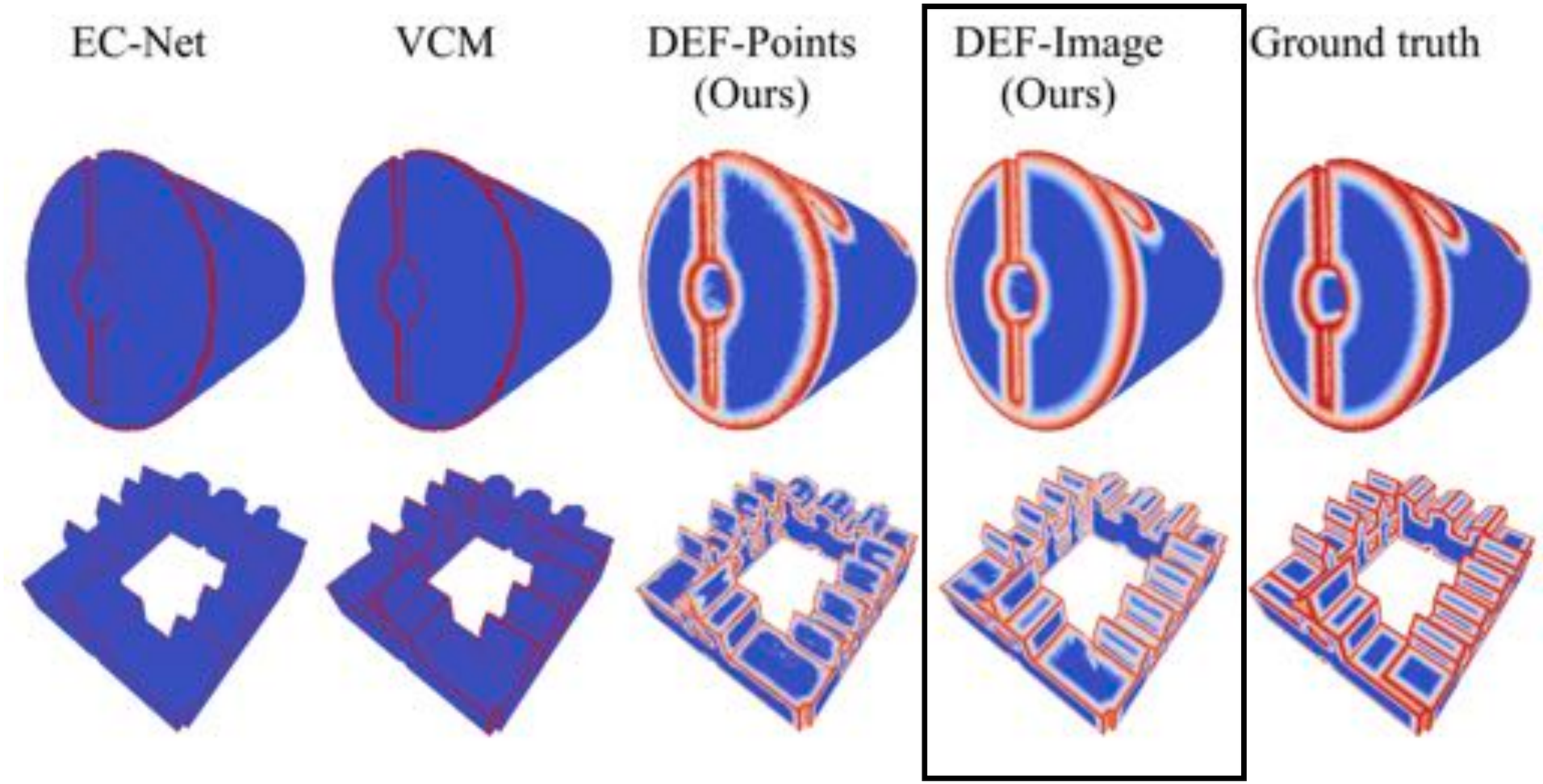


Figure credit: Matveev, A., Artemov, A., Rakhimov, R., Bobrovskikh, G., Panozzo, D., Zorin, D., & Burnaev, E. (2020). DEF: Deep Estimation of Sharp Geometric Features in 3D Shapes. *arXiv preprint arXiv:2011.15081*.

5.2. Multiple depth images: view consistency

§5. Learning with dense range-images

- Virtually the same as the previous slide, but done **during training**
- Predict for each view independently (e.g. $D_i = \mathrm{CNN}(I_i)$)
- Reproject, interpolate, concatenate (all ops must be differentiable!)
- Compute the pixelwise difference in valid pixels (multiple view consistency assumption)

$$L_{\mathrm{reproj}}(\widehat{D}_{\mathrm{ref}}) = \sum_{m=1}^{M} ||(\widehat{D}_{\mathrm{ref}} - \widehat{D}^m_{\mathrm{ref}}) \odot \widehat{V}^m_{\mathrm{ref}}||$$

- Minimize $L_{\mathrm{reproj}}(\widehat{D}_{\mathrm{ref}}) + \text{other losses}$

5.2. Multiple depth images: recap

Skoltech
Skolkovo Institute of Science and Technology

§5. Learning with dense range-images

- Multiple views provide complementary contexts
- Boost **method** performance either during **post-processing** by **fusing predictions** in the same spatial positions, visible from different viewpoints
- Boost **network** performance by **training** with **multiple view consistency** assumption

References

1. Botsch, M., Kobbelt, L., Pauly, M., Alliez, P., & Lévy, B. (2010). *Polygon mesh processing*. CRC press.
2. Ma, Y., Soatto, S., Kosecka, J. and Sastry, S.S., 2012. *An invitation to 3-d vision: from images to geometric models* (Vol. 26). Springer Science & Business Media.
3. Hartley, R. and Zisserman, A., 2003. *Multiple view geometry in computer vision*.