Combining Machine Learning and Geometric Optimisation for Pose Estimation

Associate Professor Tat-Jun (“TJ”) Chin

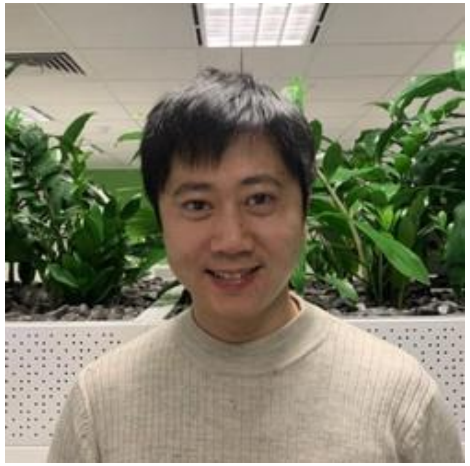
School of Computer Science, The University of Adelaide
and
Australian Institute for Machine Learning

THE UNIVERSITY of ADELAIDE

adelaide.edu.au/aiml

Introduction

The team



Dr Bo Chen

Dr Alvaro Parra Bustos

Associate Professor
Tat-Jun Chin

Adelaide, South Australia

Lot 14 Innovation Precinct

Australian Space Agency

- Established 1 July 2018.
- Annual budget: $9.8 Million (tiny!)
- Roles:
  - Providing **national policy** and strategic advice on the civil space sector.
  - Coordinating Australia's domestic **civil** space sector activities.
  - Supporting the **growth** of Australia's space industry and the use of space across the broader **economy**.
  - Leading **international** civil space **engagement**.
  - Administering space activities **legislation** and delivering on our international obligations.
  - **Inspiring** the Australian community and the next generation of space entrepreneurs.

SMARTSATCRC

BUILDING AUSTRALIA'S SPACE INDUSTRY

The SmartSat CRC is a consortium of industry and research organisations that will develop game changing technologies to bootstrap Australia's space industry and catapult it into the 1/2 trillion dollar global economy.

We invite you to join us in this nation-building journey.

ANNOUNCEMENT

$245M PARTICIPANT & GOVERNMENT FUNDING SECURED

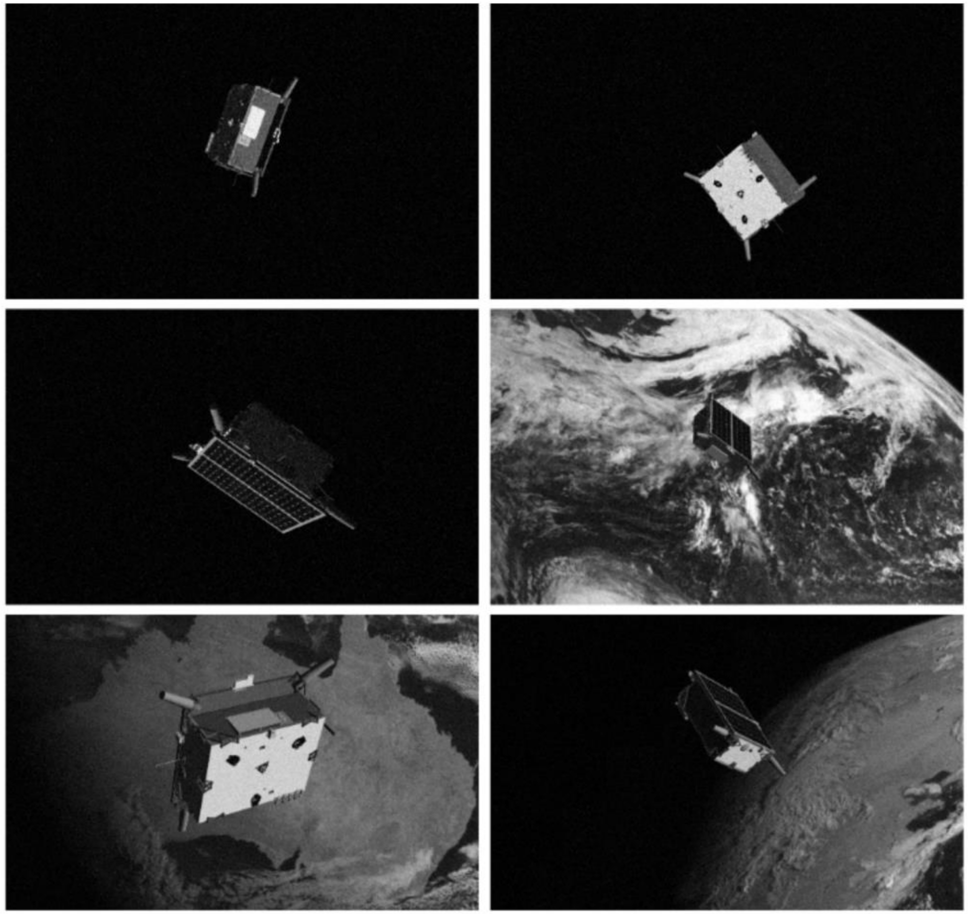
adelaide.edu.au/aiml

adelaide.edu.au/aiml

On to pose estimation…

Kelvins Pose Estimation Challenge

- Given an image of a satellite, estimate its 6DoF pose (position, orientation) w.r.t. to the observing camera.
- Available data:
  - Training images of satellite with ground truth pose.
- Challenges:
  - Varying lighting, scale, background.
- Assumptions:
  - It's always the same satellite…



More general pose estimation

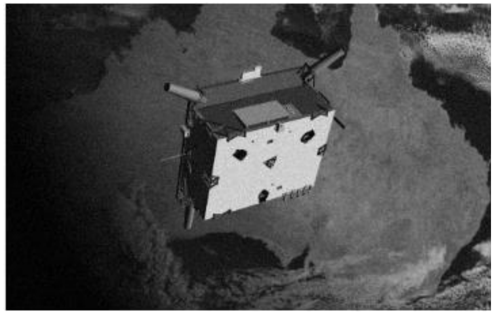
- Usually needs to work on an object class (e.g., faces, cars).
- Significant variations in the instances of the object class.
- Significant variations in object and environmental conditions.

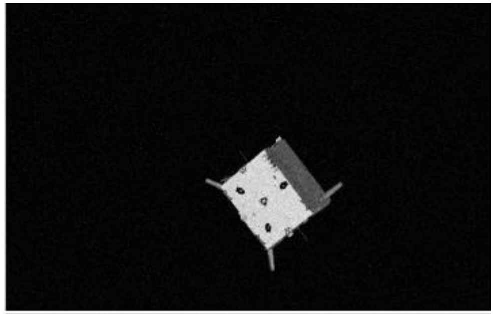
Yan Li, Leon Gu, Takeo Kanade. A Robust Shape Model for Multi-view Car Alignment. CVPR 2009.

Baseline 1: Geometric technique

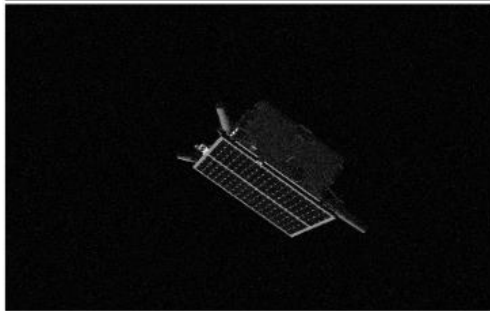
SIFT, SURF, etc.

Testing image



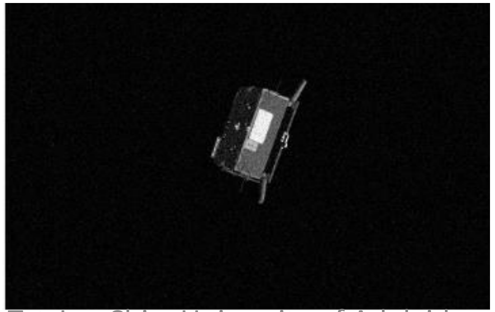


Pre-selected training image 1



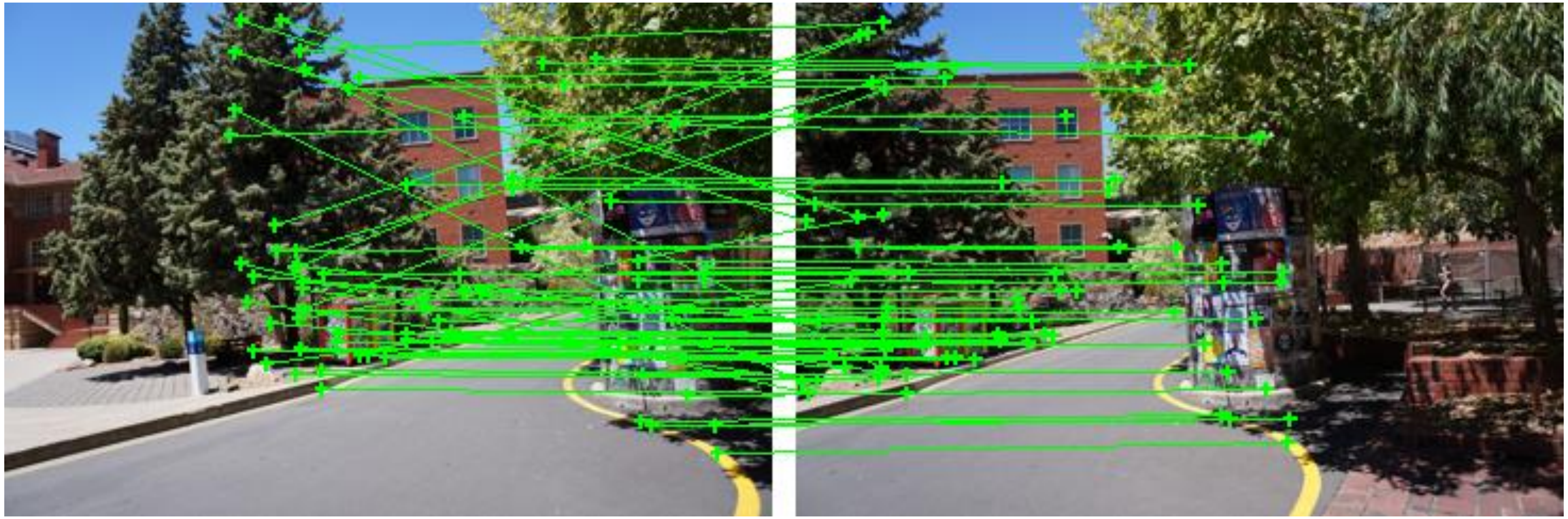
Pre-selected training image 2

⋮



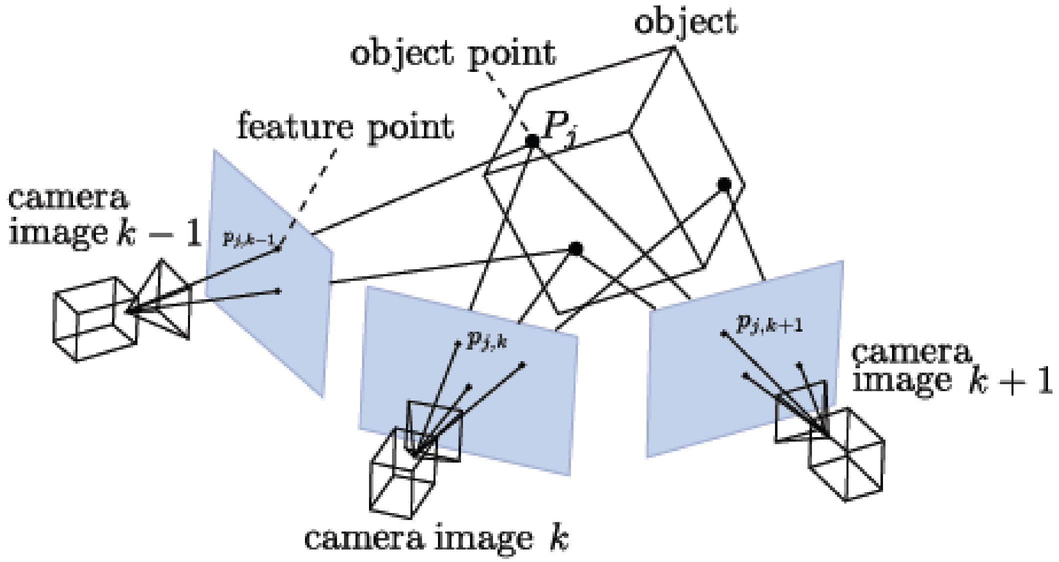
Pre-selected training image N

Keypoint matching and detection

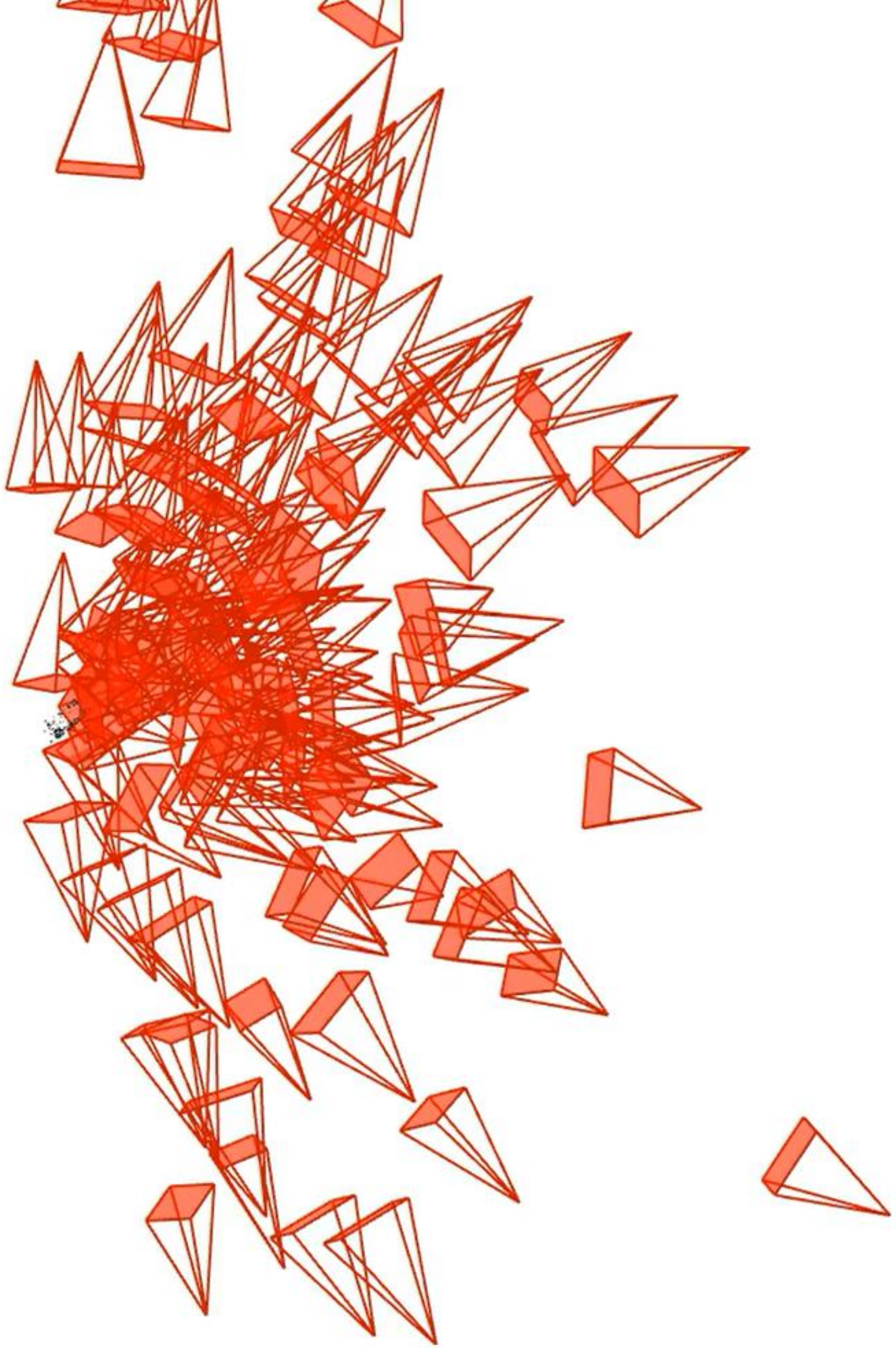


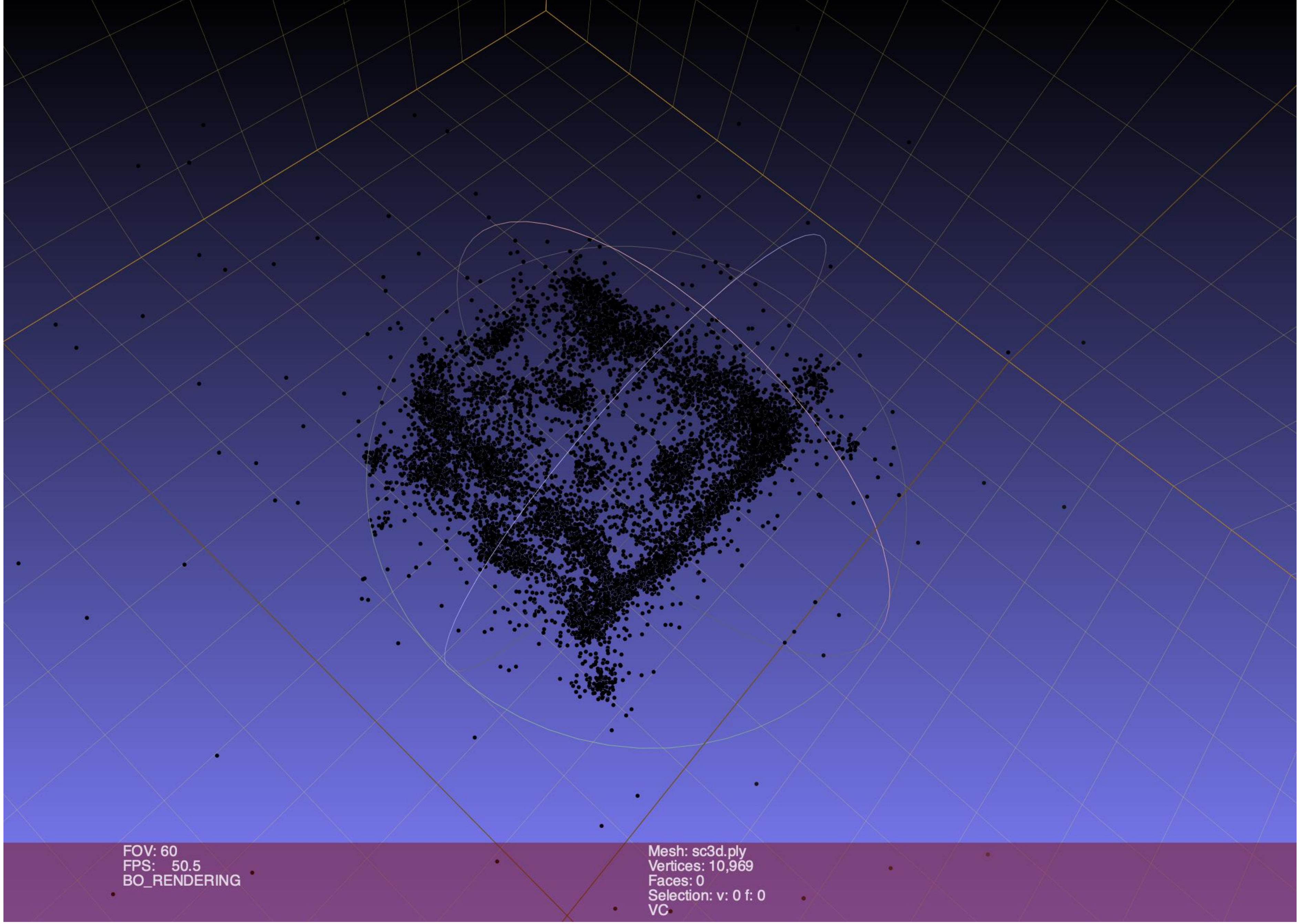
Structure-from-motion

- Given observations of a set of 3D points in a number of images, estimate the coordinates of the 3D points and relative poses of the cameras (that captured the images).



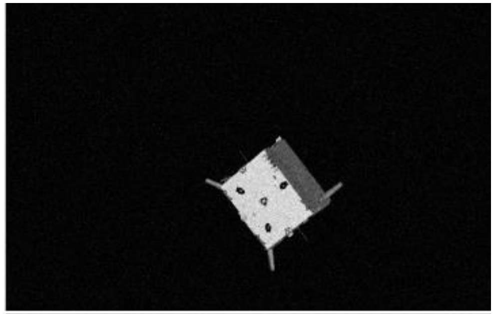
Source: [openmvg.readthedocs.org/en/latest/_images/structureFromMotion.png]




FOV: 60
FPS: 50.5
BO_RENDERING
Mesh: sc3d.ply
Vertices: 10,969
Faces: 0
Selection: v: 0 f: 0
VC

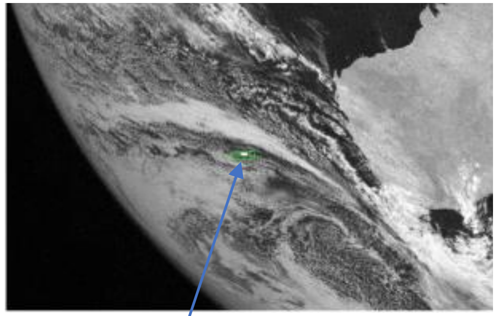
Baseline 1: Geometric technique

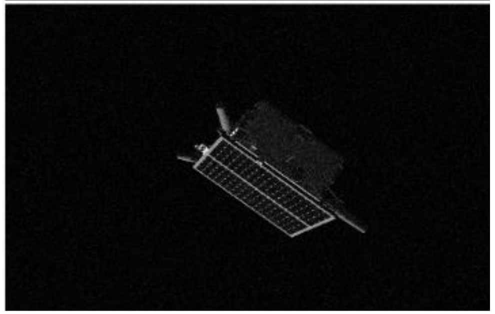
SIFT, SURF, etc.



Pre-selected training image 1

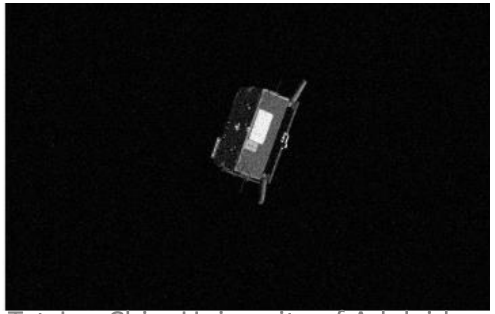
Testing image





Pre-selected training image 2

⋮



Pre-selected training image N

Target satellite
(about 30x30 pixels)

Baseline 1: Geometric technique

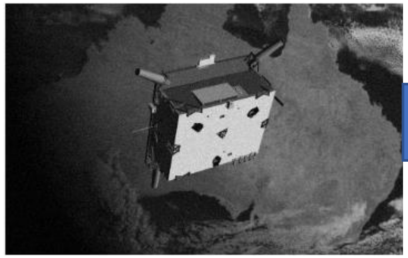
Pros

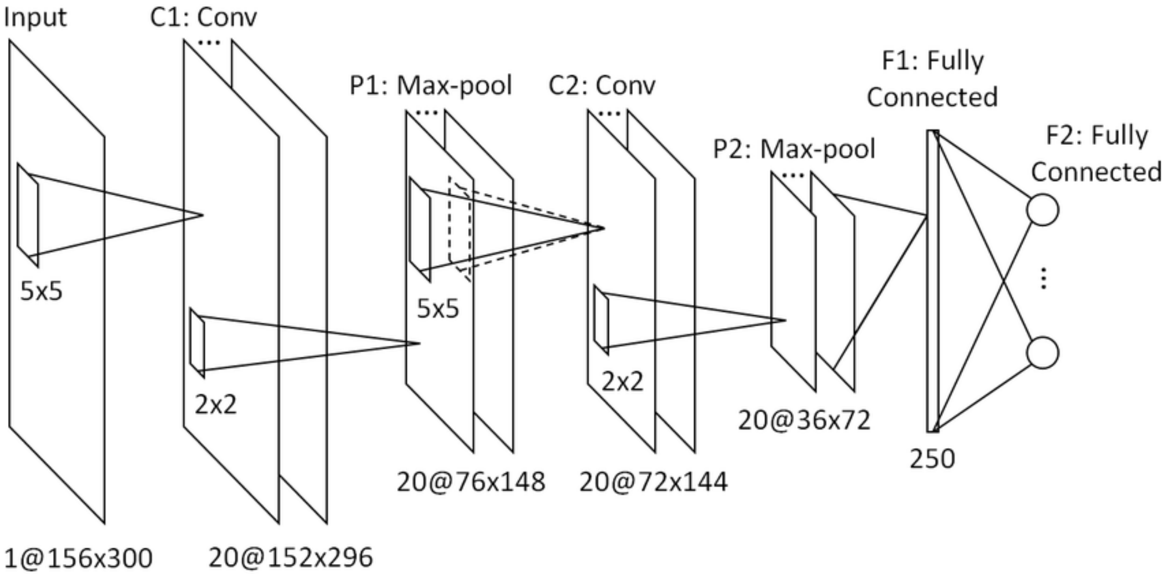
- Gives very accurate pose estimates when it works – when sufficient good keypoint matches can be established.

Cons

- Fails catastrophically on images that don't work – target too small, background too cluttered.

Baseline 2: End-to-end learning

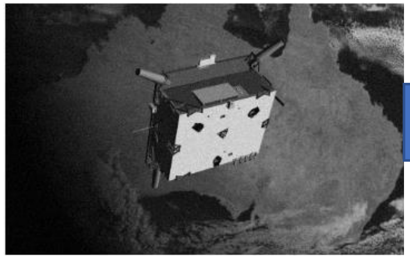


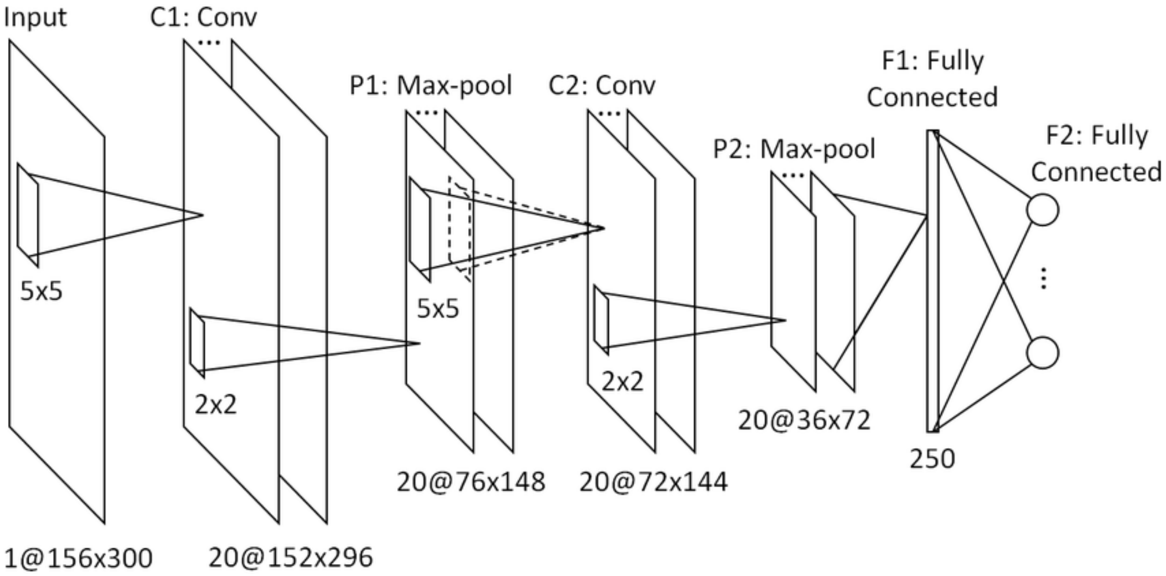


Geometric loss function
(e.g., Kendall et al., ICCV 2015)

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

Baseline 2: End-to-end learning



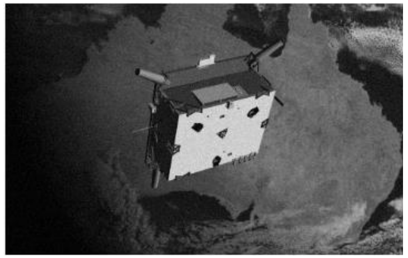


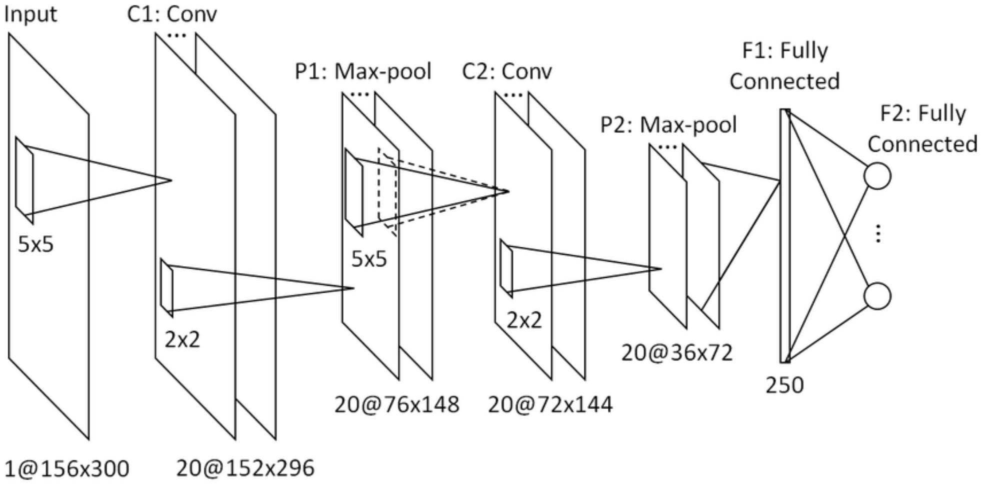
Geometric loss function
(e.g., Kendall et al., ICCV 2015)

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

“Magic” parameter

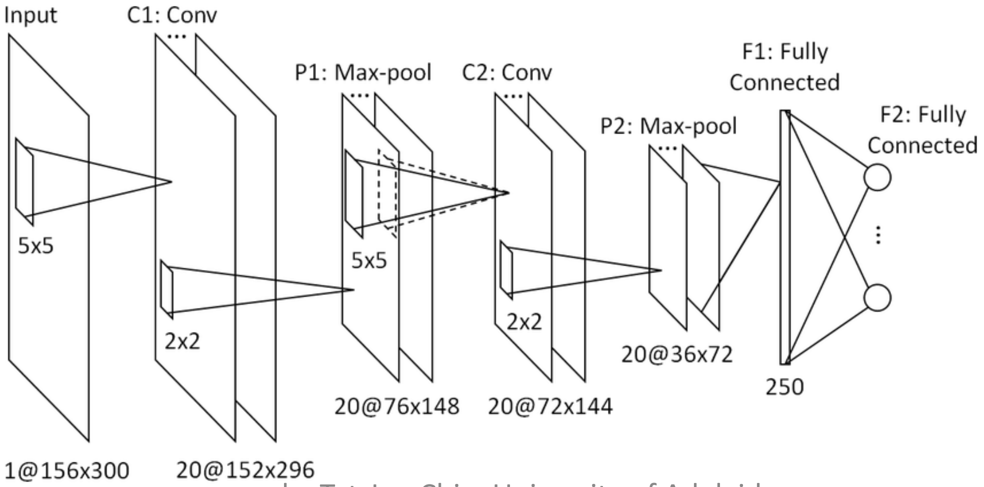
Baseline 2: End-to-end learning





Minimise position error:

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2$$



Minimise orientation error:

$$loss(I) = \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

Baseline 2: End-to-end learning

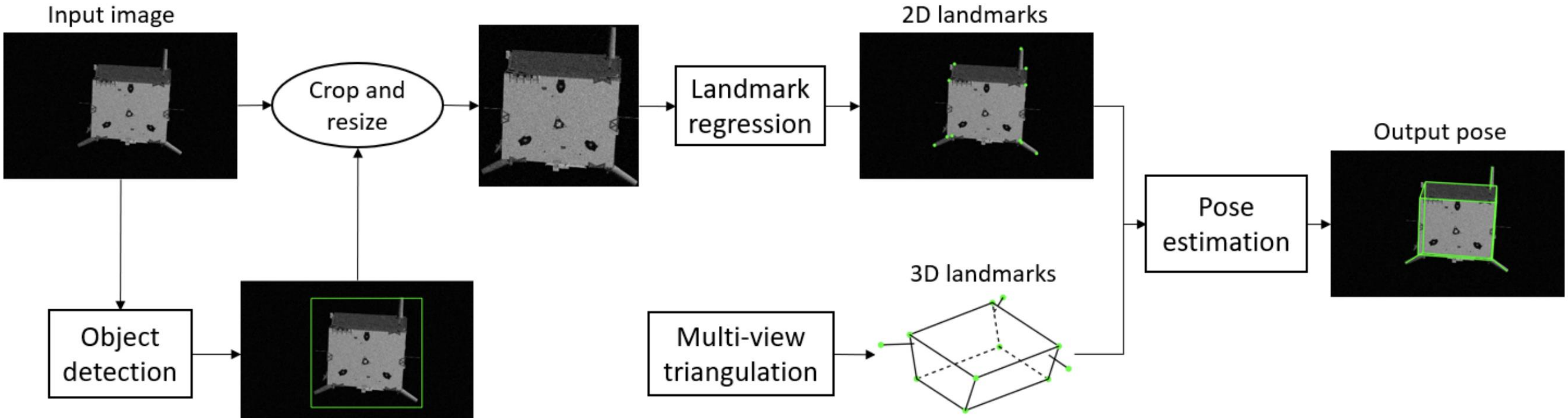
Pros

- Easy to implement (PyTorch, TensorFlow, etc.).
- Tends to not fail catastrophically.

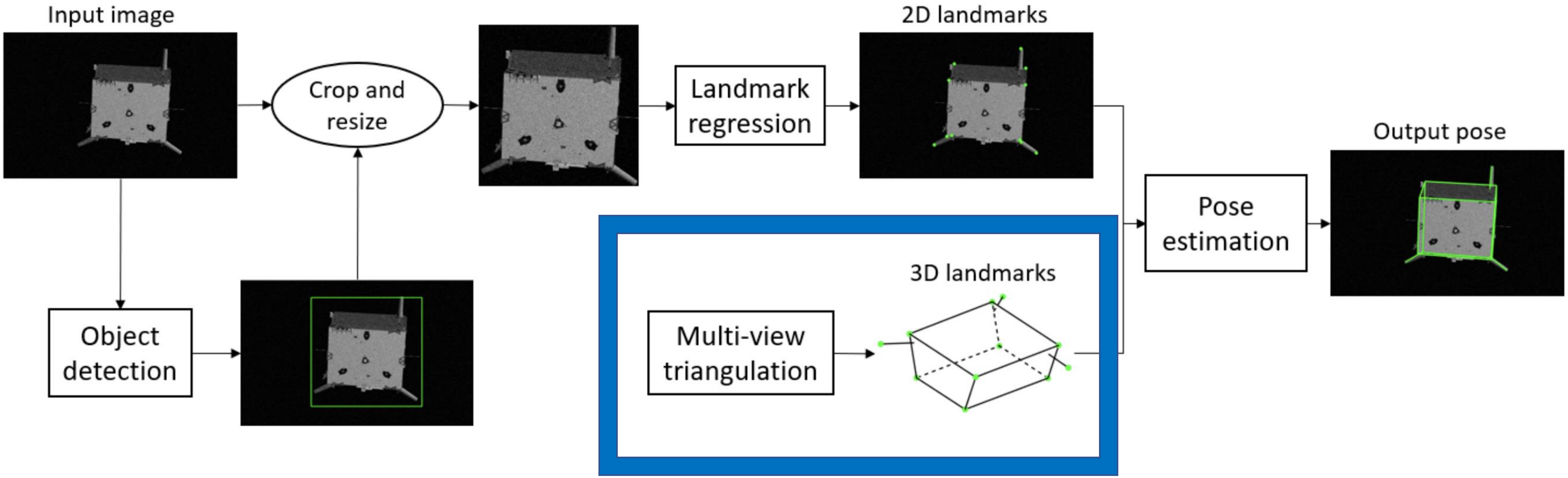
Cons

- Rather imprecise/inaccurate.
- Not explainable.

Combine deep learning and geometry



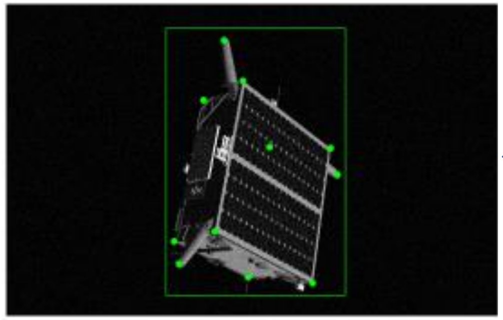
Combine deep learning and geometry

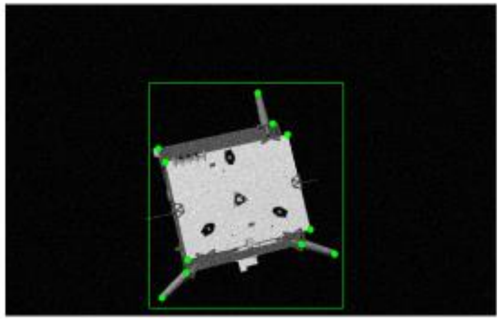


Offline processing

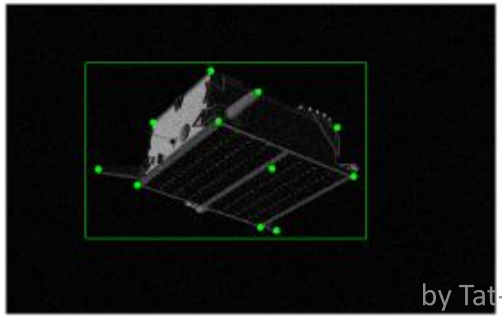
Build 3D model (selected landmarks only)

Selected training images with manually ”clicked” landmarks

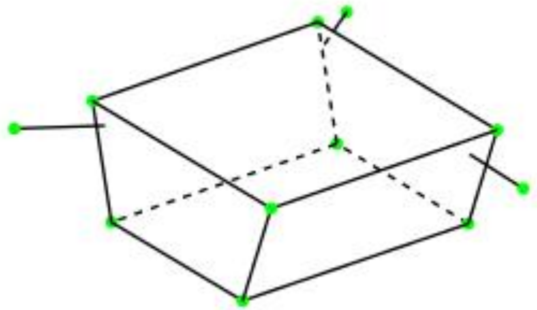




⋮



Multi-view triangulation



Multi-view triangulation

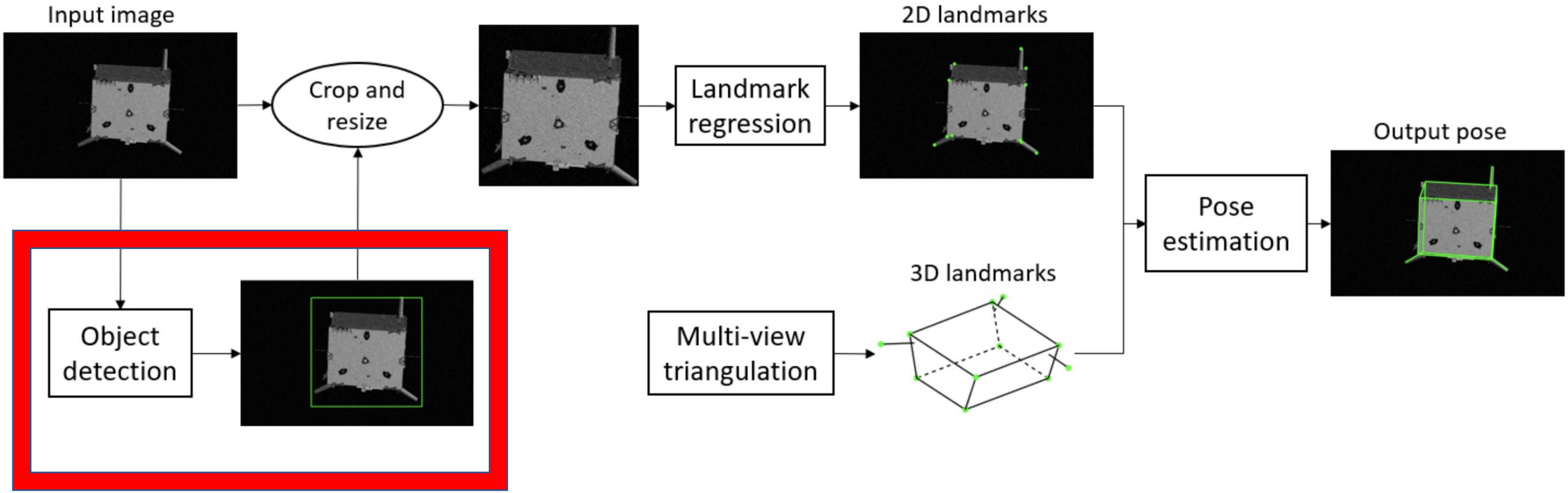
- Given 2D observations $\{(x_i, y_i)\}_{i=1}^{N}$ of a landmark and camera pose $\mathbf{p}_i$ behind each observation (x_i, y_i), solve

$$\min_{\mathbf{X}} \sum_{i}^{N} \|(x_i, y_i) - f(\mathbf{X} \mid \mathbf{p}_i)\|_2^2$$

where $f(\mathbf{X} \mid \mathbf{p}_i)$ computes the pinhole projection of 3D point $\mathbf{X}$ onto image with camera pose $\mathbf{p}_i$.

- Typical algorithm: get initial solution using algebraic least squares, then refine using nonlinear least squares (Levenberg-Marquardt).

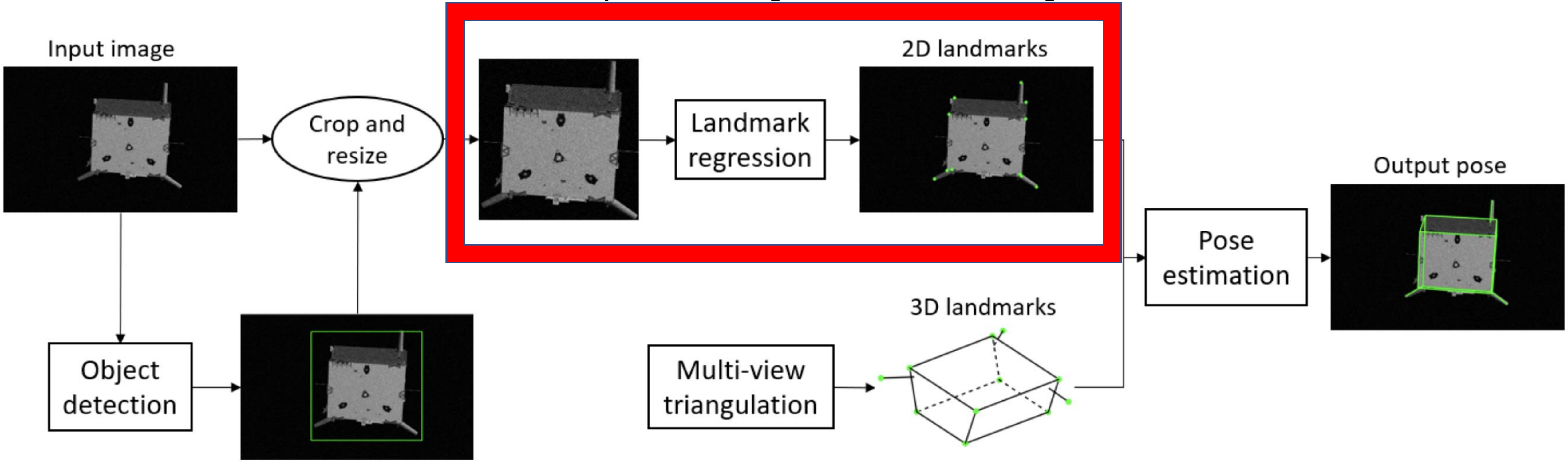
Combine deep learning and geometry



Online processing, offline training, standard object detection method (Faster RCNN)

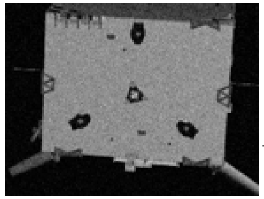
Combine deep learning and geometry

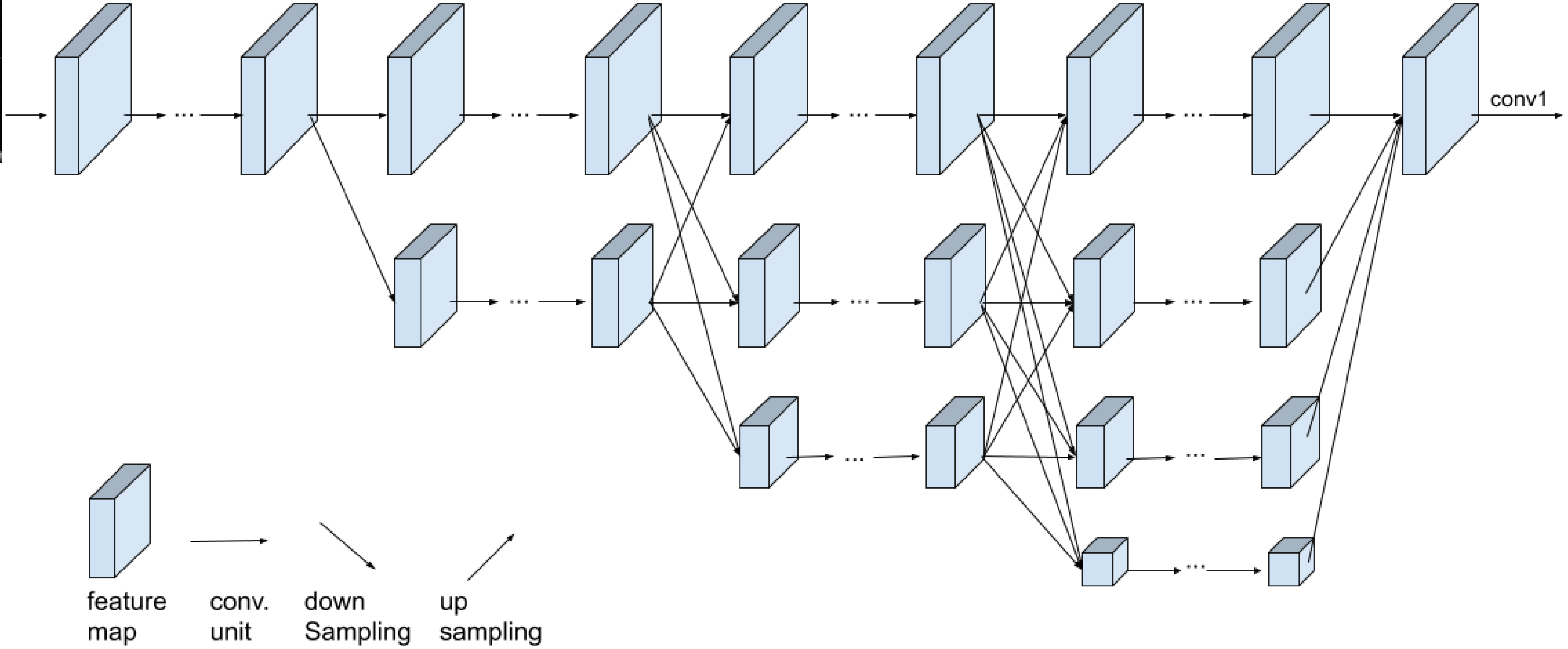
Online processing, offline training



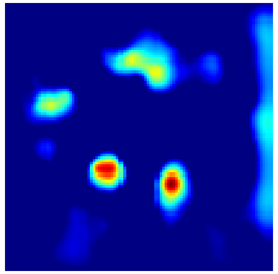
Landmark regression using HRNet

HRNet by Sun et al., Deep high-resolution representation learning for human pose estimation, CVPR 2019.

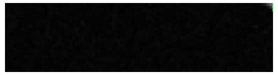




Landmark heatmap



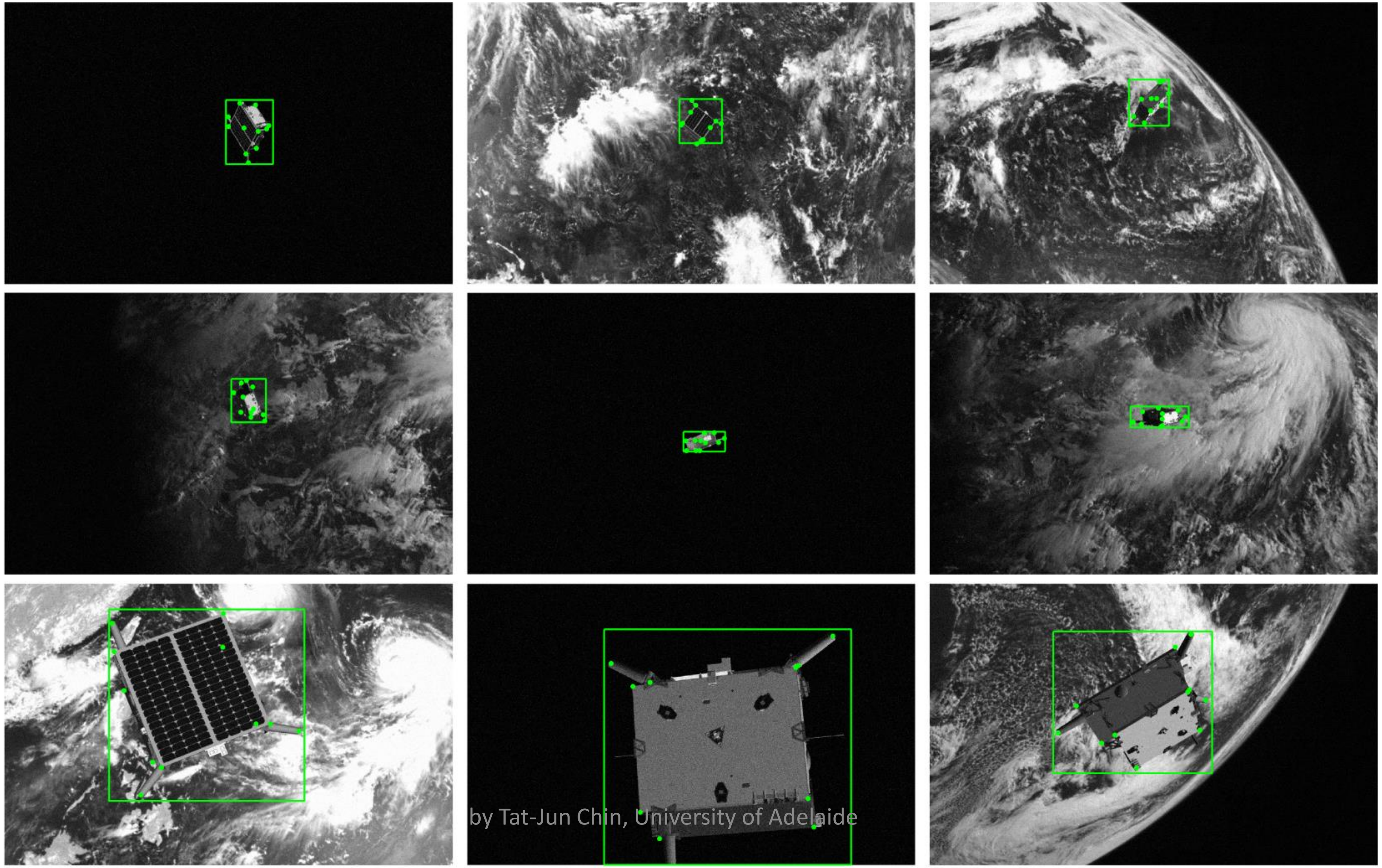
Ground truth landmarks



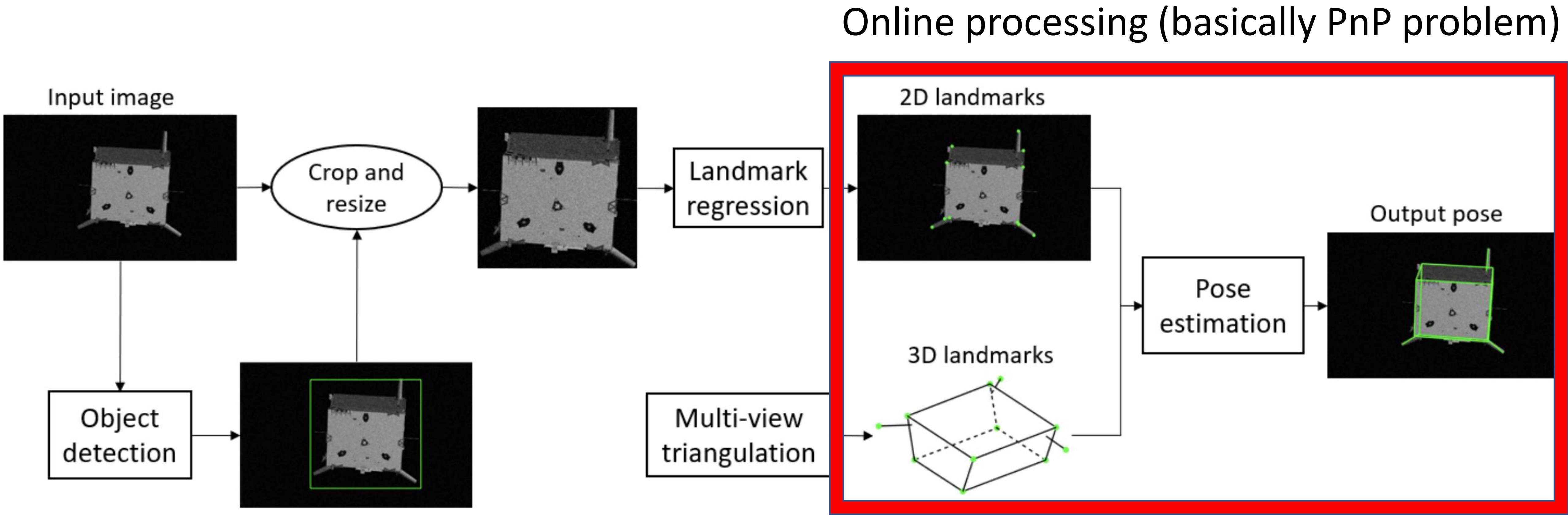
Loss function

$$\ell = \frac{1}{N}\sum_{i=1}^{N} v_i\left(h(\mathbf{z}_i) - h(\mathbf{z}_i^*)\right)^2$$

Sample results: object detection and landmark regression



Combine deep learning and geometry



Perspective-n-point (PnP)

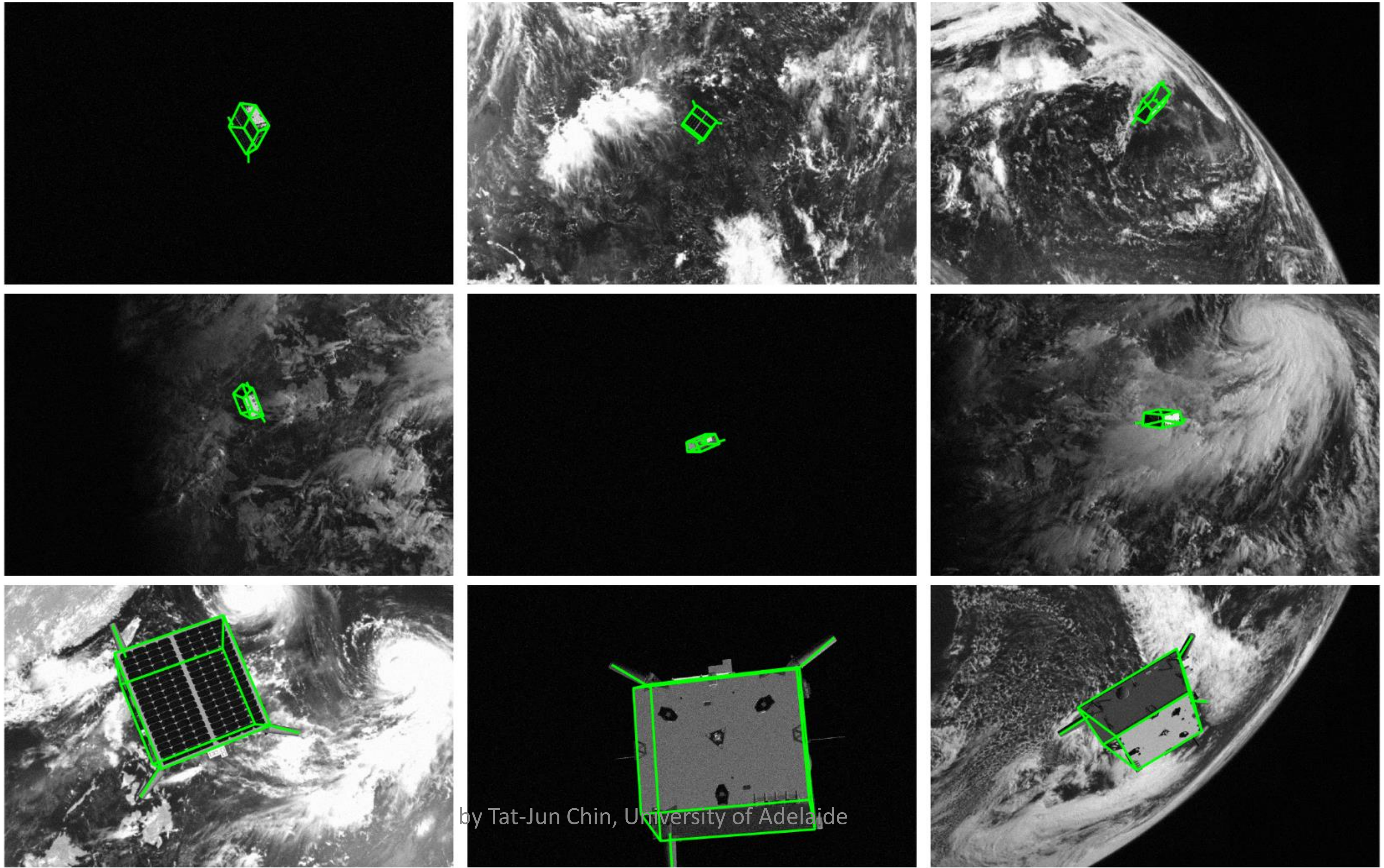
- Given 2D observations $\{(x_i, y_i)\}_{i=1}^{N}$ of a set of landmarks $\{\mathbf{X}_i\}_{i=1}^{N}$, solve for the camera pose $\mathbf{p}$

$$\min_{\mathbf{p}} \sum_i \|(x_i, y_i) - f(\mathbf{X}_i \mid \mathbf{p})\|_2^2$$

where $f(\mathbf{X}_i \mid \mathbf{p})$ computes the pinhole projection of 3D point $\mathbf{X}_i$ onto image with camera pose $\mathbf{p}$.

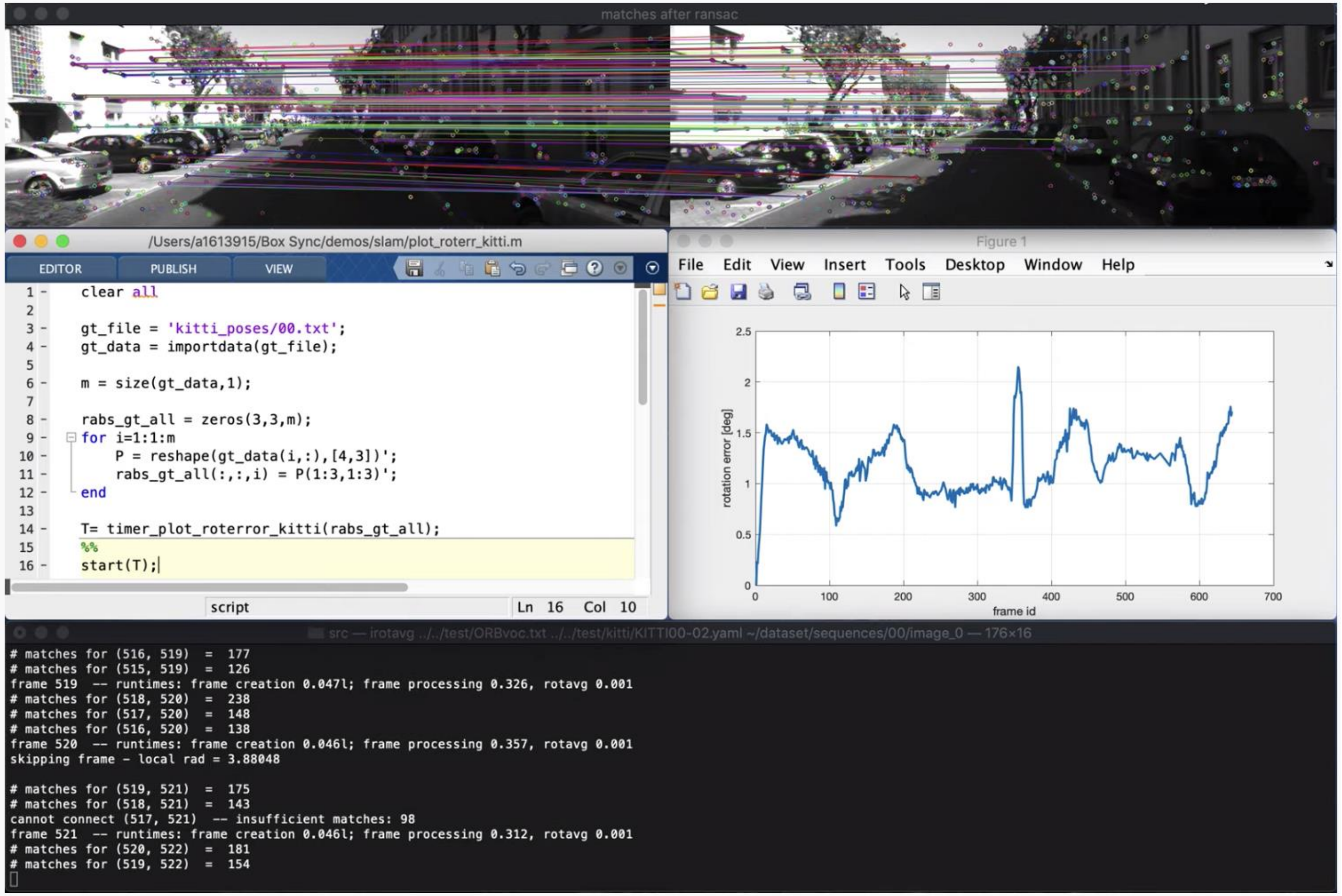
- Typical algorithm: get initial solution using algebraic least squares, then refine using nonlinear least squares (Levenberg-Marquardt).

Sample results: pose estimation (displayed as projection of 3D wireframe model)



Potential future work

Exploit temporal continuity



Joint estimation of pose and dense 3D model from single image

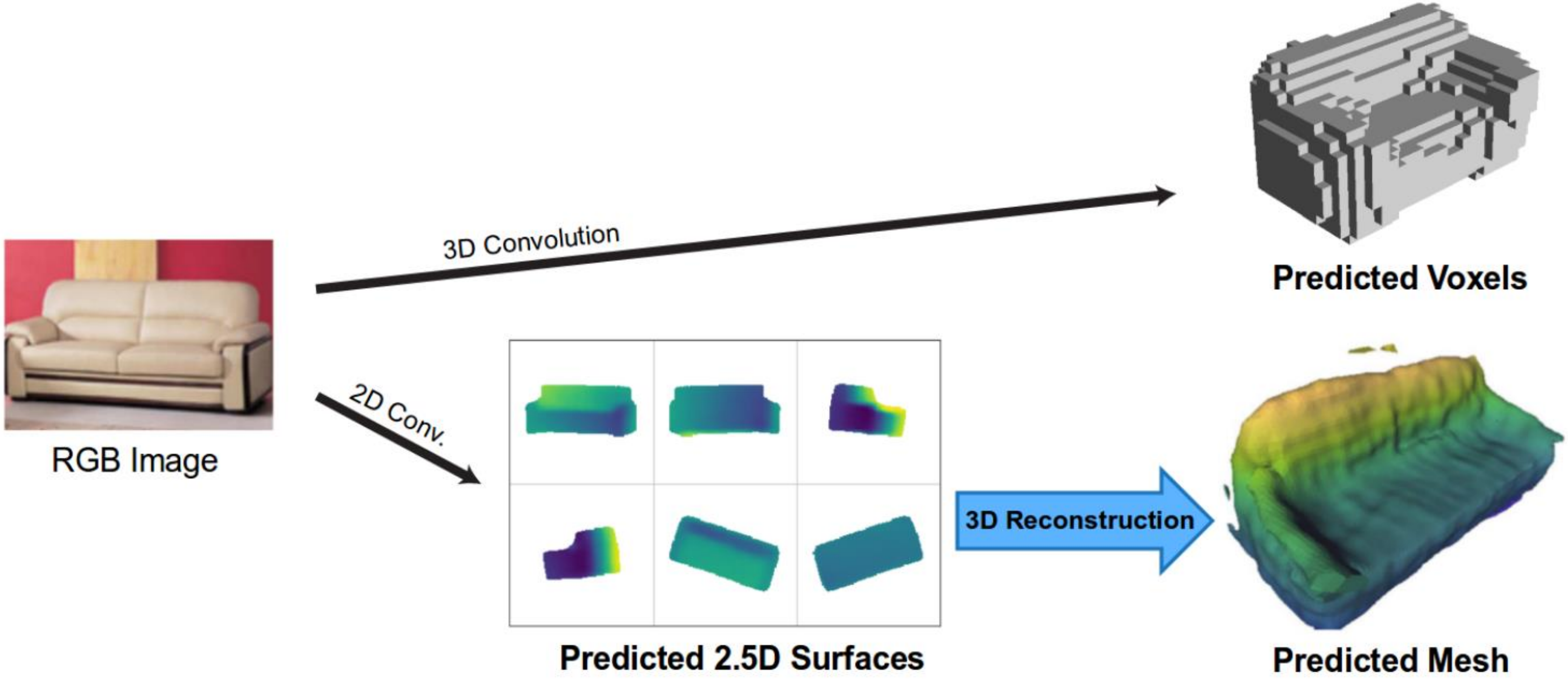
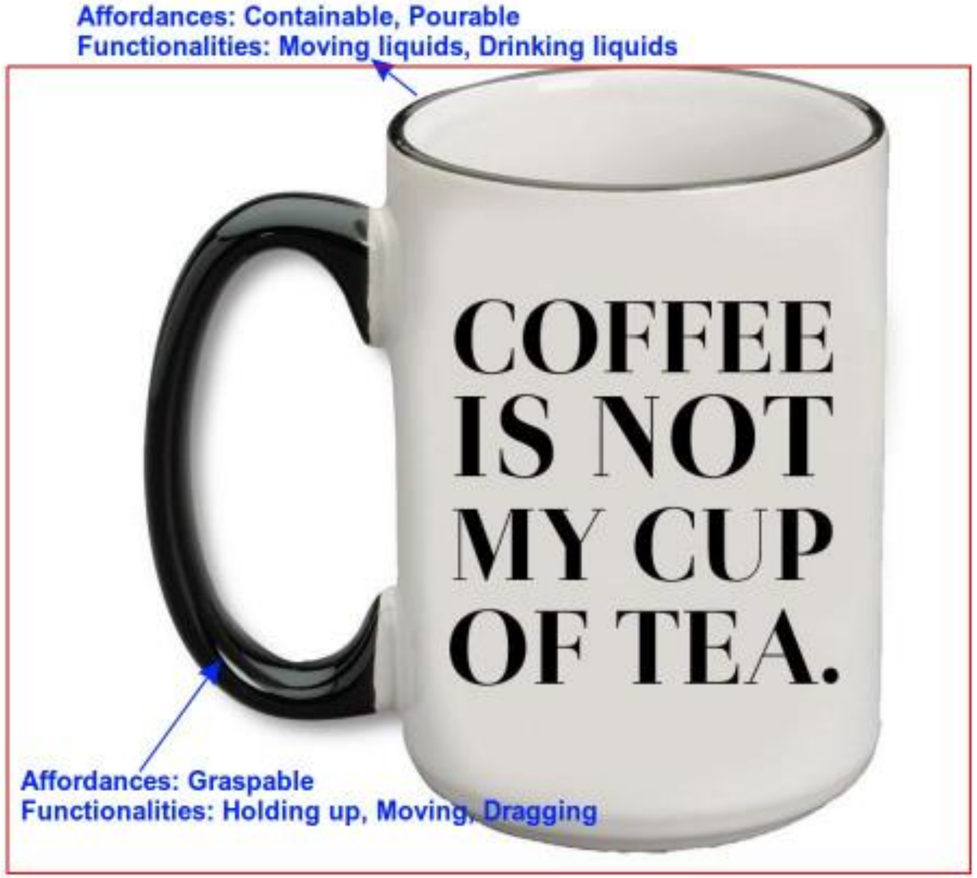


Figure from Shin et al., CVPR 2018.

Cope with intra-class variations

- Challenges:
  - Can the same set of landmarks be used for different satellites?
  - Can the object detector cope with intra-class variations?

Visual affordance estimation



Figures from Hassanin et al., Visual Affordance and Function Understanding: A Survey.

Thank you!