3D Vision: Theory, Application and New Trends

A Brief Tour of 3D Computer Vision

Sudipta N. Sinha

Microsoft Research, Redmond, USA

July 4, 2018

3rd SUMMER SCHOOL ON COMPUTER VISION,

BASICS OF MODERN AI, 2—7 July 2018, IIIT Hyderabad

Plan for today

3D Vision: Theory, Applications and New Trends

- *Morning*
  - *Broad overview (this lecture)*
  - *Multiple View Geometry (Chetan)*
- *Afternoon*
  - *Correspondence problems in computer vision (Sudipta)*
  - *New Trends (Chetan)*
- *Demo and Lab (Suvam and Vishnu)*

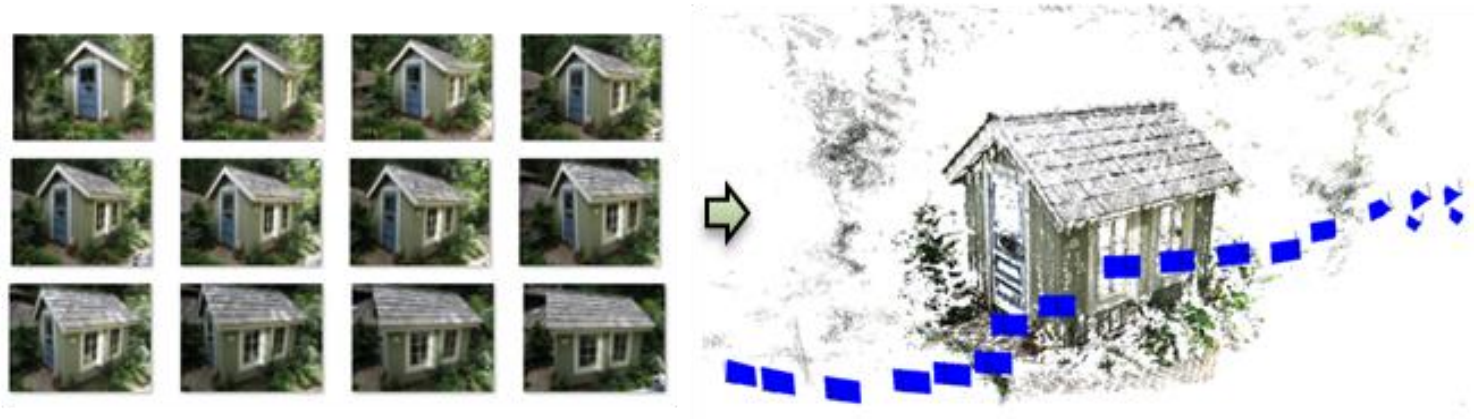
Goals of Computer Vision

- Understanding images and video by
  - *Estimating numeric aspects of the 3D scene (measurement)*

Measurement

Image-based modeling:

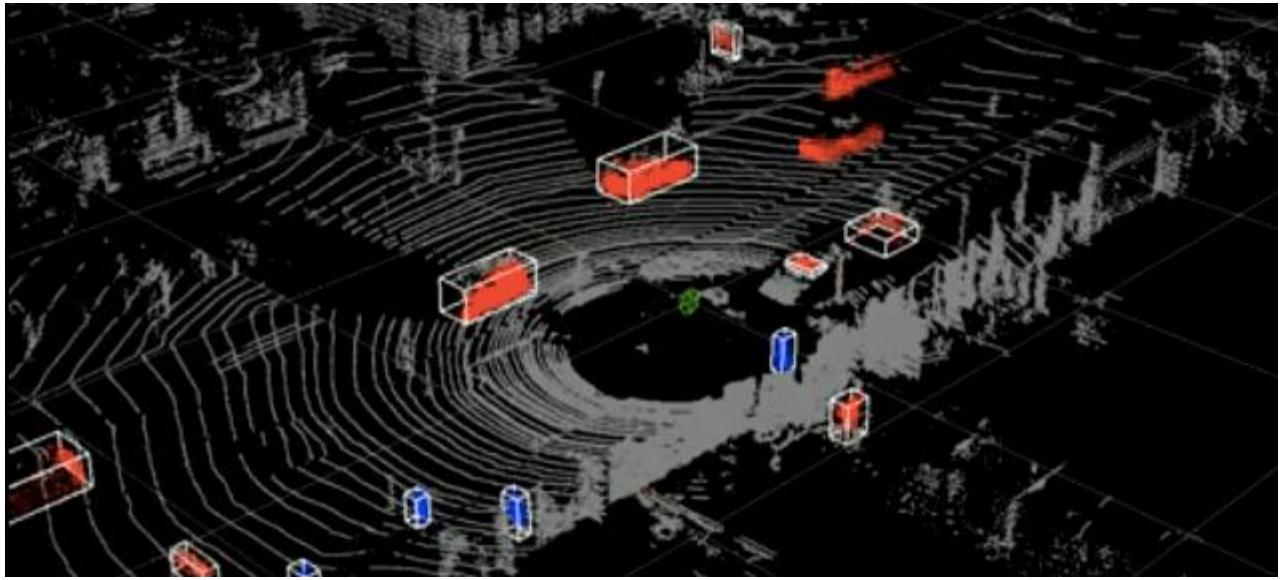
Capture scene geometry



Structure from Motion + Multi-view Stereo Reconstruction

Autonomous Driving:

Detect and track moving objects

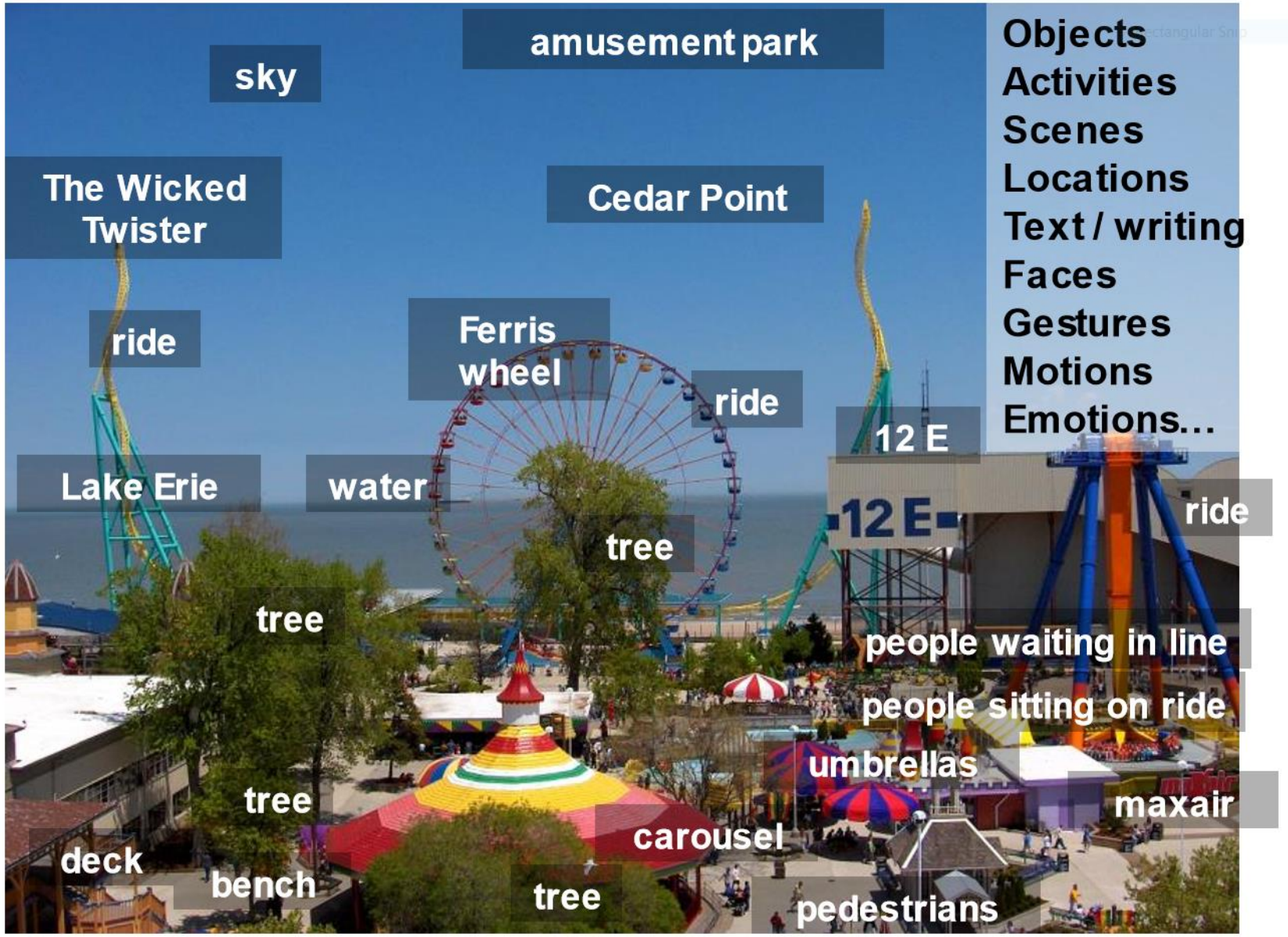


LIDAR point cloud processing from a roof-mounted sensor (Velodyne)

Goals of Computer Vision

- Understanding images and video by
  - *Estimating numeric aspects of the 3D scene (measurement)*
  - *Enabling a machine to recognize objects, people, scenes, and activities. (recognition)*

Recognition, Semantic Understanding



Source:
Kristen Grauman

Goals of Computer Vision

- Understanding images and video by
  - *Estimating numeric aspects of the 3D scene (measurement)*
  - *Enabling a machine to recognize objects, people, scenes, and activities. (recognition)*
  - *Making visual data searchable (organization)*

Goals of Computer Vision

- Understanding images and video by
  - *Estimating numeric aspects of the 3D scene (measurement)*
  - *Enabling a machine to recognize objects, people, scenes, and activities. (recognition)*
  - *Making visual data searchable (organization)*
  - *Generating novel visual output (synthesis)*

3D Computer Vision

- Measurements:
  - geometric properties of scene
  - camera and object motion
- Recover 3D models of objects and scenes from images
- Vision as Inverse Graphics
  - Reverse engineer the process that created an image of the 3D world; and then answer questions about the world.

Overview

- Why study 3D Computer Vision?
- Applications
- Basic Principles and Algorithms
- Preview: Important 3D vision tasks
  - 3D Reconstruction: Structure from Motion, SLAM
  - Camera Localization
  - Object detection and pose estimation
  - Image and Video Editing

Overview

- Why study 3D Computer Vision?
- Applications
- Basic Principles and Algorithms
- Preview: Important 3D vision tasks
  - 3D Reconstruction: Structure from Motion, SLAM
  - Camera Localization
  - Object detection and pose estimation
  - Image and Video Editing

Common Tasks in Computer Vision

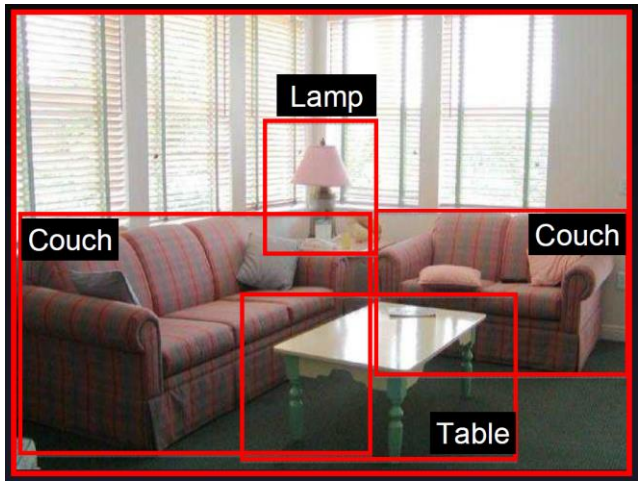
2D Recognition; Semantics Understanding

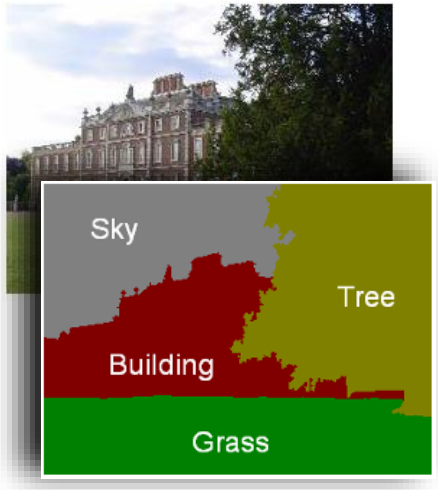
- Image categorization
- Instance recognition
- Object detection
- Semantic segmentation
- 2D human body pose estimation
- Face detection + recognition
- Emotion recognition
- Action recognition
- Image search
- Tracking
- Image captioning

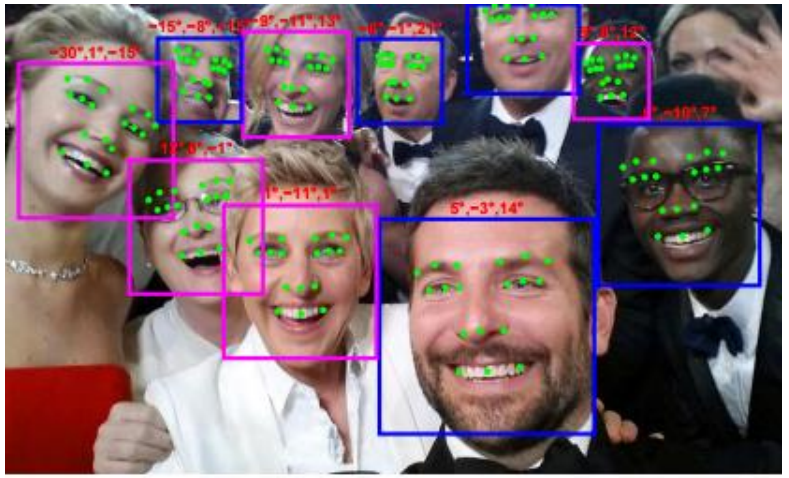
3D Reconstruction; Geometric Scene Interpretation

- Structure from motion (SfM): rigid vs. non-rigid
- Dense stereo matching
- Optic flow / scene flow
- 3D shape recovery
- Absolute and relative camera pose
- Location recognition
- Simultaneous localization and mapping (SLAM)
- Image-based modeling / 3D scanning
- BRDF estimation, illumination and reflectance
- Novel view synthesis
- Viewpoint invariant object recognition

2D Semantic Image Understanding







Object detection

- Object category
- Bounding box

Semantic Segmentation

- Object category
- Pixel labeling

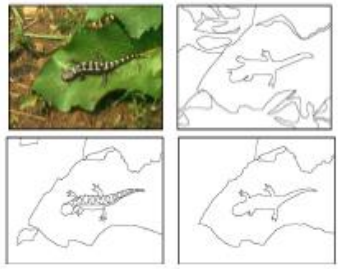
Understanding faces

- Facial landmarks
- Head pose
- gender

Human body pose

- Skeleton joint positions (2D)

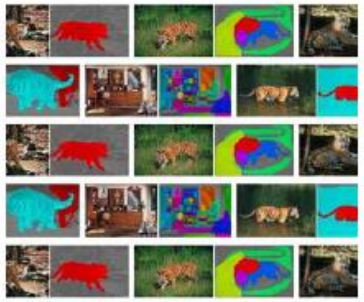
Recognition Benchmarks



BSD (2001)

Caltech 101 (2004), Caltech 256 (2006)

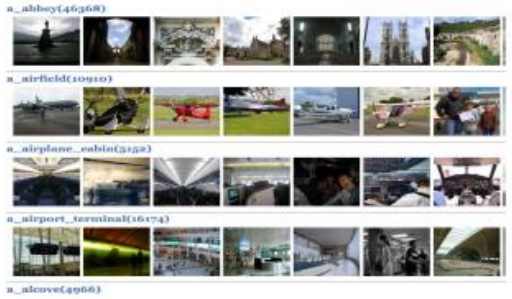
PASCAL (2007-12)



LabelMe (2007)

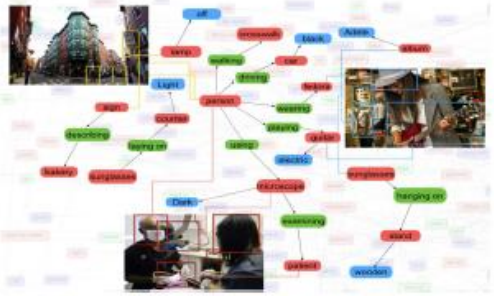
ImageNet (2009)

SUN (2010)



Places (2014)

MS COCO (2014)



Visual Genome (2016)

Learning in Vision

Supervised Learning:

Given training set $\{(x_i, y_i)\}$, $i = 1 \ldots N$, where $x_i \in X, y_i \in Y$ and X and Y are the inputs and outputs, learn a parametric function

$$f(x, \boldsymbol{\theta}) : X \rightarrow Y$$

Find $\boldsymbol{\theta}^*$ that minimizes the loss on the training set

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ L(\theta, X, Y) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} l(f(x_i; \boldsymbol{\theta}), y_i)$$

Function family: Linear classifier, SVM, random forests, convolutional neural nets

Loss: Exact form of L( ), additional loss terms to encode other priors

Optimizer: Greedy method, convex optimization, SGD

Learning in Vision

Supervised Learning:

Given training set $\{(x_i, y_i)\}$, $i = 1$ ... inputs and outputs, learn a parame

$$f(x, \boldsymbol{\theta}) : X \rightarrow Y$$

Find $\boldsymbol{\theta}^*$ that minimizes the loss on t

$$f(x; \boldsymbol{\theta}) = F^K(F^{K-1}(F^{K-2}(\cdots F^1(x)\cdots)))$$

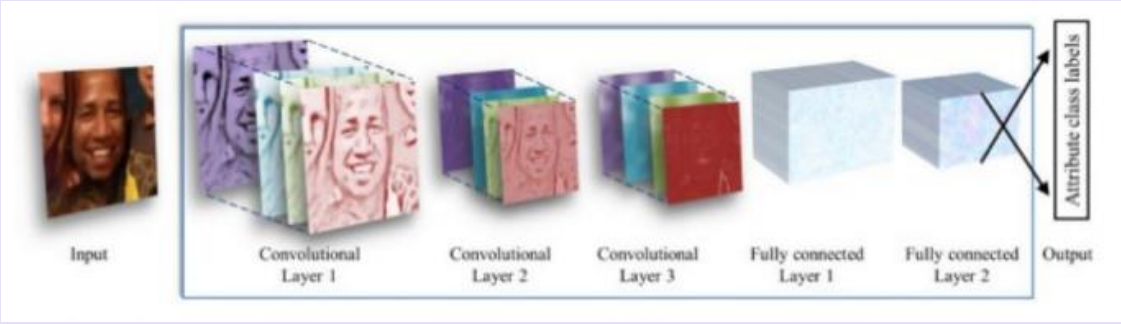


image → prediction

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ L(\theta, X, Y) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1} l(f(x_i; \boldsymbol{\theta}), y_i)$$

Function family: Linear classifier, SVM, random forests, **convolutional neural nets**

Loss: Exact form of L( ), additional loss terms to encode other priors

Optimizer: Greedy method, convex optimization, SGD

3D reconstruction and applications

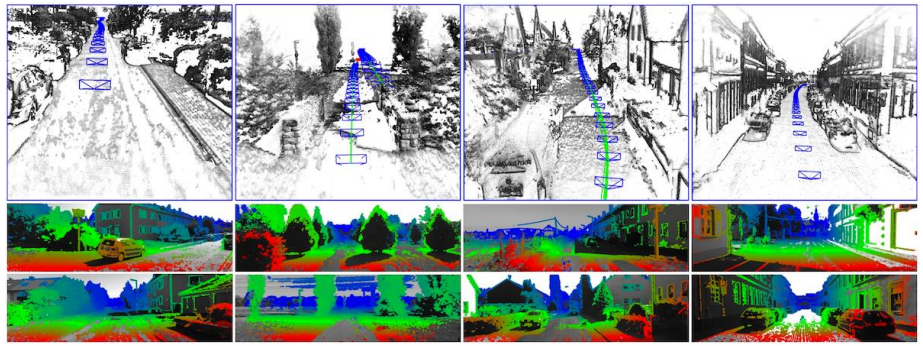
Image stitching

Structure from motion

3D scanning, depth sensing

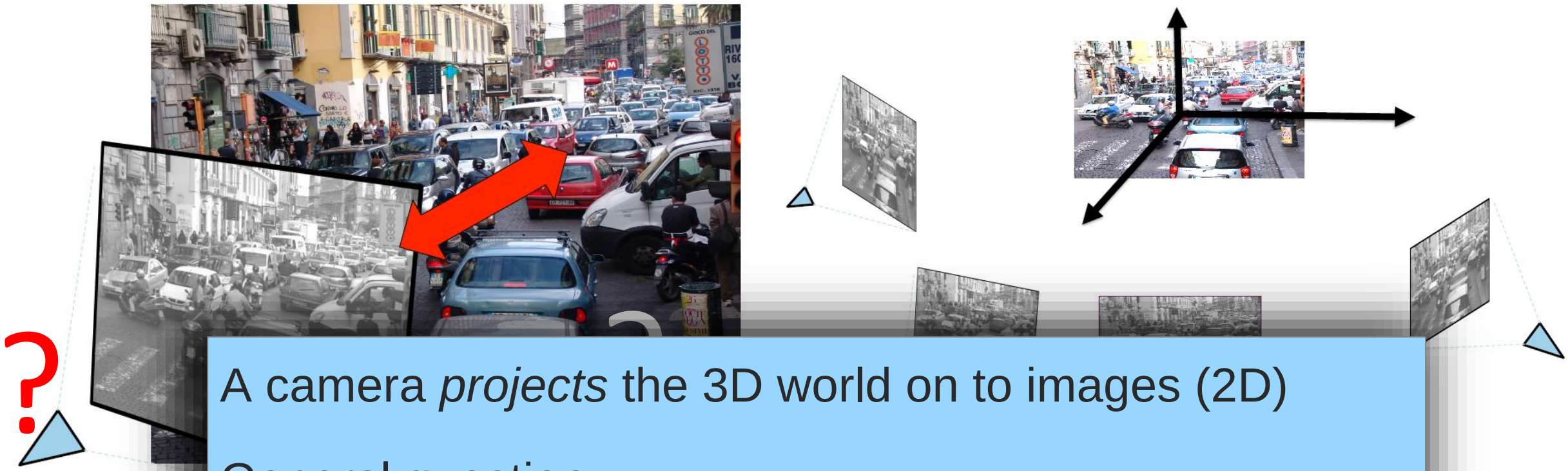
Augmented reality

360º cameras and video



Visual SLAM +
Autonomous navigation

Geometric 3D Computer Vision

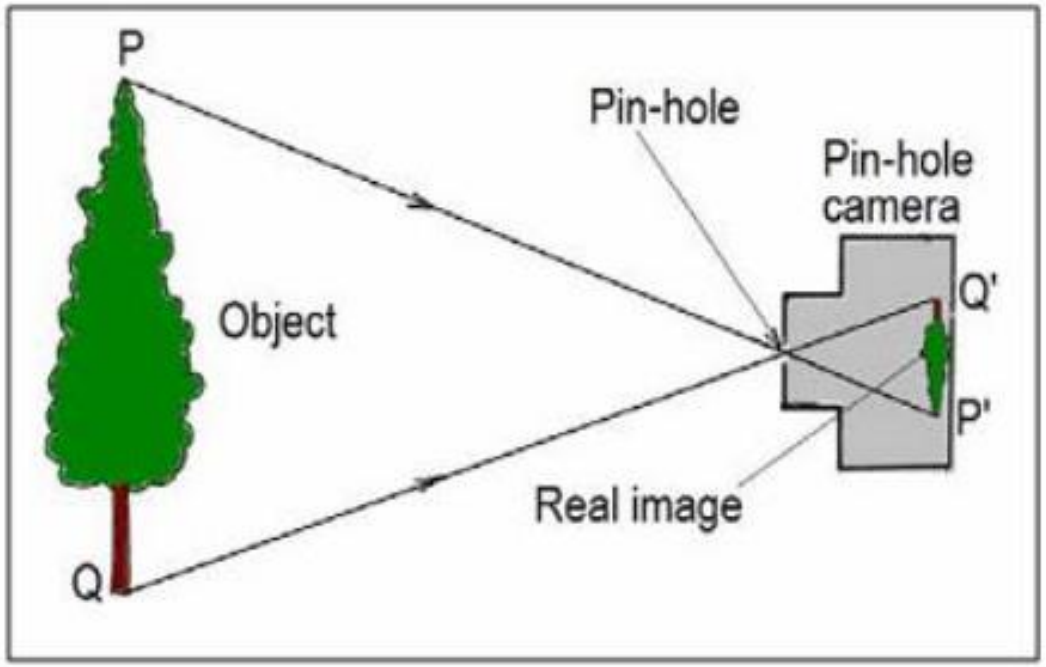


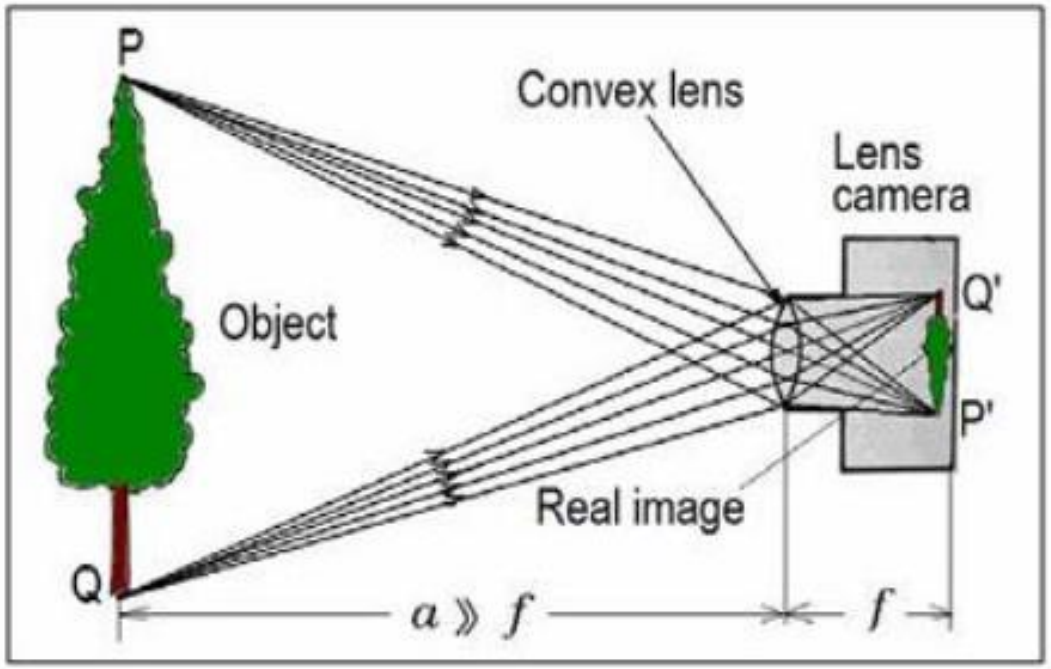
A camera *projects* the 3D world on to images (2D)

<u>General question:</u>

Given one or more images, what can you tell about the geometry of the underlying scene or the cameras.

Image Formation





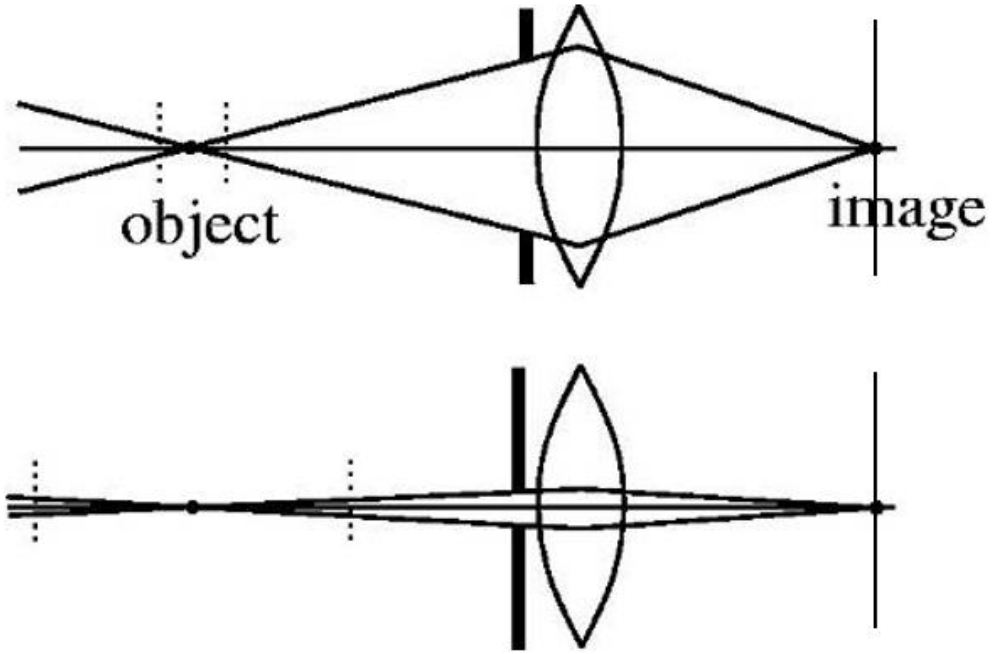
- Pin-hole camera model
- Larger pinhole; add a lens to get a sharp image

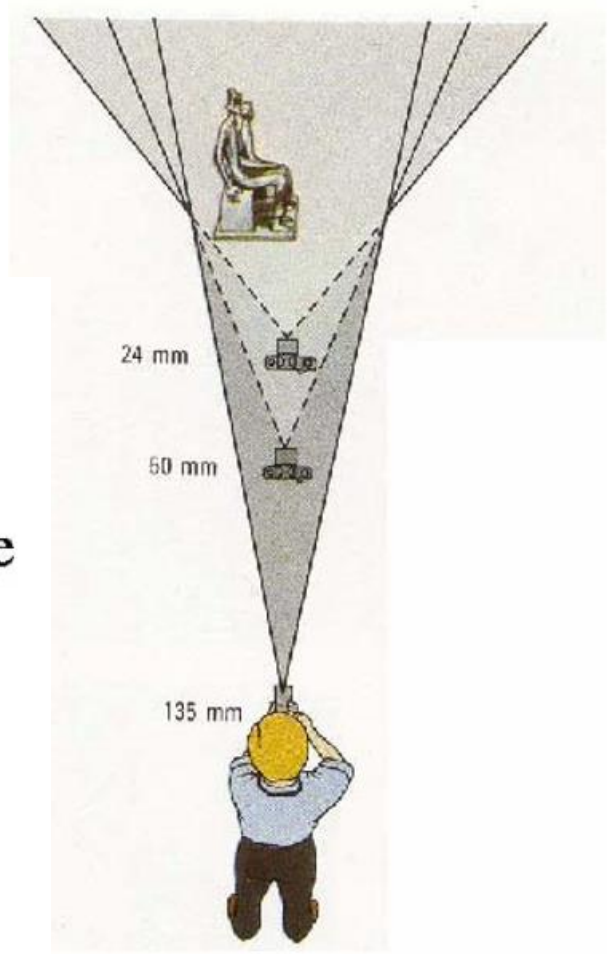
Image Formation

- Perspective Projection
  - 3D → 2D
  - Occlusions
- Information lost
  - Depth (or distance)
  - Angles
- Depth Recovery
  - 2.5D representation

Image Formation

- Choice of the Lens
  - Field of View/Focal length
  - Depth of Field





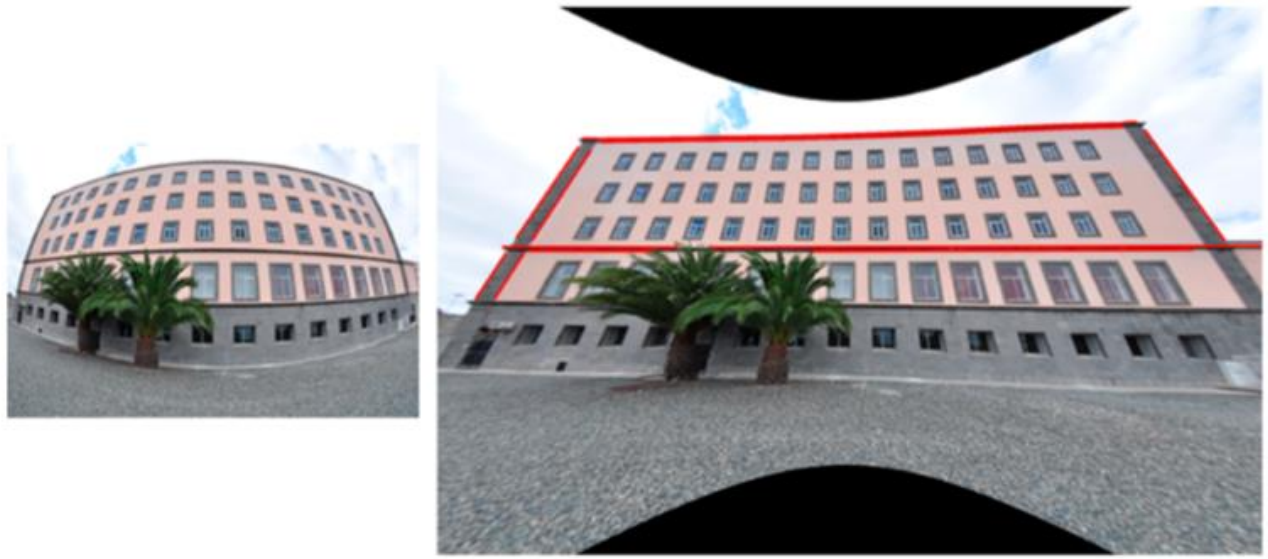
Large FOV, small f
Camera close to car

Small FOV, large f
Camera far from the car

Source: Rob Fergus; Alyosha Efros

Image Formation

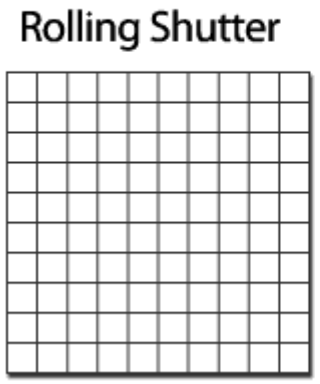
- Radial lens distortion
- Fish eye lens

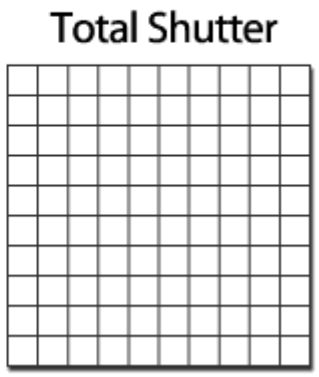


Source: Wu+ 2017, Optical Engineering

Image Formation

- Rolling Shutter vs. Global Shutter





- Jello effect and other kind of rolling shutter distortions.
- Common in low-end smartphone cameras.

Recognition vs. 3D Reconstruction

- Aspects of image formation (camera lens models, 3D to 2D projections) are typically ignored when solving image recognition or pattern classification tasks.
- However, modeling the image formation process is crucial for 3D reconstruction and related tasks.
- The camera model parameters must be accurately obtained during a pre-calibration phase or must be recovered during the reconstruction process.

Overview

- Why study 3D Computer Vision?
- Applications
- Basic Principles and Algorithms
- Preview: 3D vision tasks and algorithms
  - 3D Reconstruction: Structure from Motion, SLAM
  - Camera Localization
  - Image and Video Editing
  - Object detection and pose estimation

Applications that need 3D Vision

- Robotics
- 3D/4D Aerial Maps
- Augmented Reality, Virtual Reality
- 3D photography, Special Effects for Films
- Gaming, Entertainment
- Image and Video Editing
- Image Forensics
- 3D Telepresence

Robotics

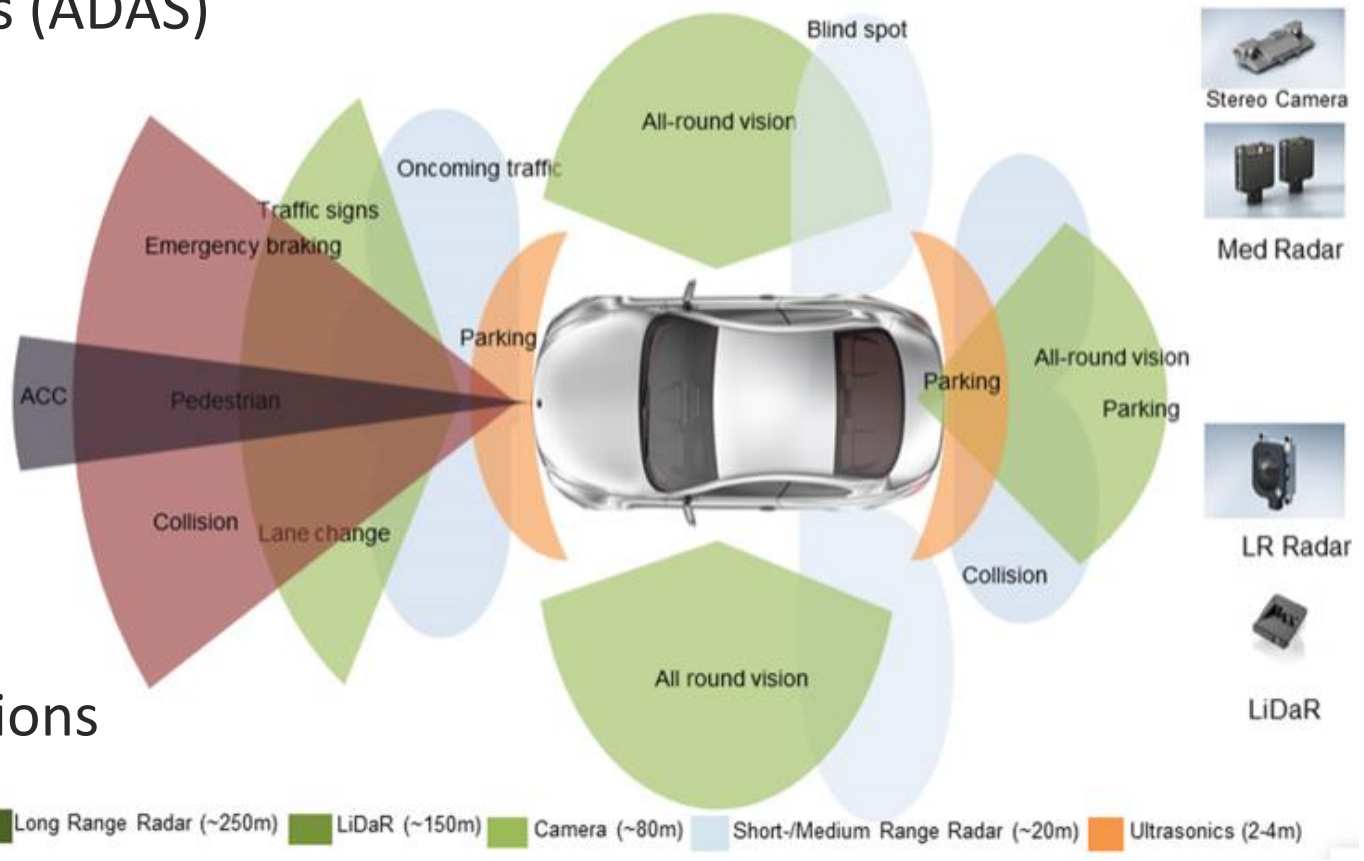
Enable machines to perceive scenes, objects, humans and other entities. Then, make decisions and interact with the 3D world.

- Autonomous vehicles
- UAVs and Drones
- Home Robots, Assistants
- Industrial Robots
- Robotic Arms

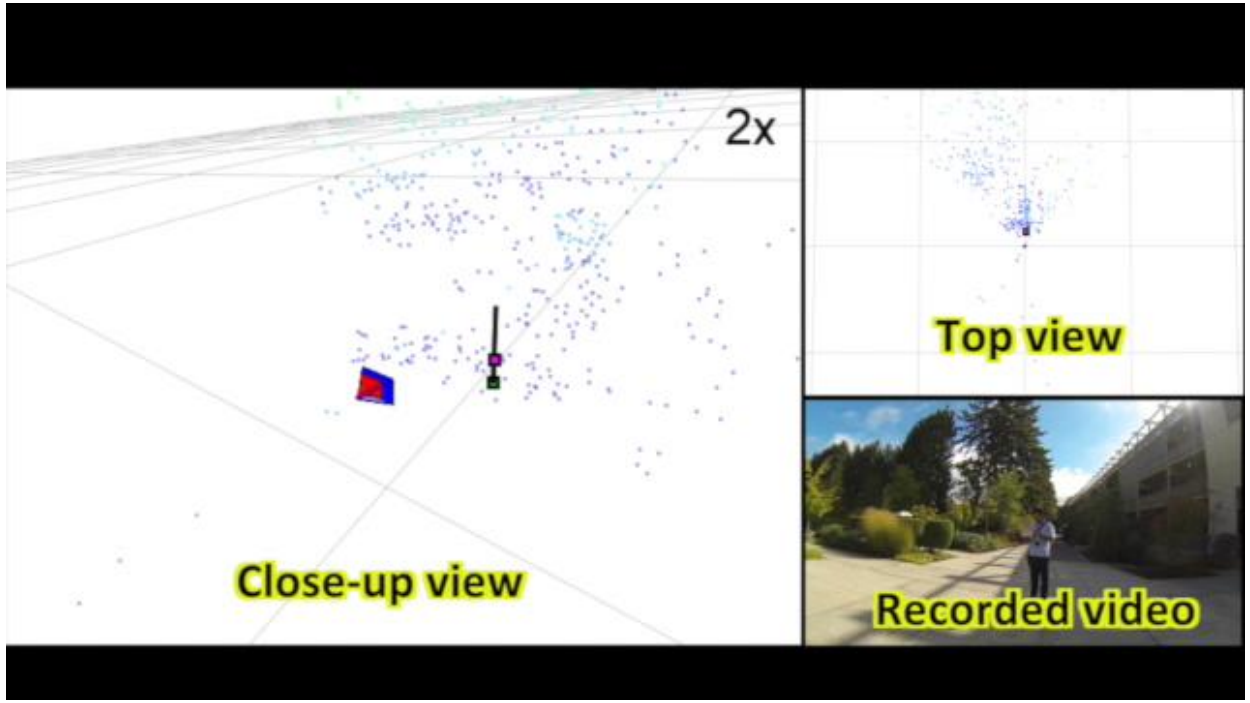
Autonomous Driving

Advanced Driver Assistance Systems (ADAS)

- Lane following
- Localization
- Object Detection (recognition)
  - 2D, 3D, bounding boxes
  - Pixel-level segmentation
- Finding drivable surface
- Recognizing road signs
- Make near term “future” predictions
  - vehicles, bicyclists
  - pedestrians,
- Detect obstacles; Automatic braking
- Unexpected situations (traffic light not working, road work, closure, detours



Unmanned Aerial Vehicles and Drones



Lim and Sinha, ICRA 2015

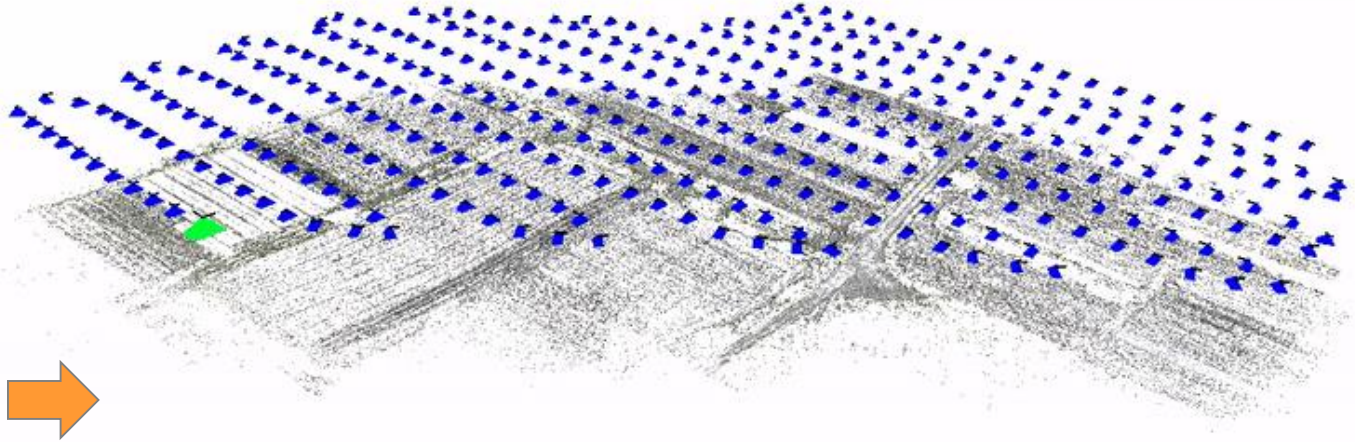
- Photography, cinematography; robotic flying camera
- Search and rescue; active search and detection

Vision-based Autonomous Navigation in GPS-denied Environments

Forster et al 2017, *SVO: Semi-Direct Visual Odometry for Monocular and Multi-Camera Systems*, IEEE Transaction on Robotics

Drone Imagery in Precision Agriculture

3D point cloud reconstruction
(using aerial 3D photogrammetry)

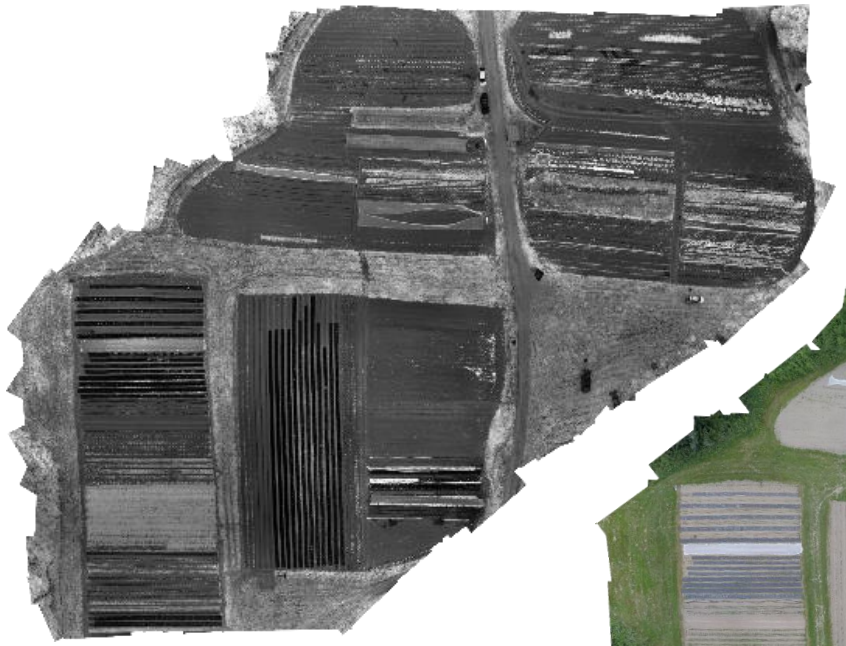


Cameras corresponding to video keyframes shown in blue

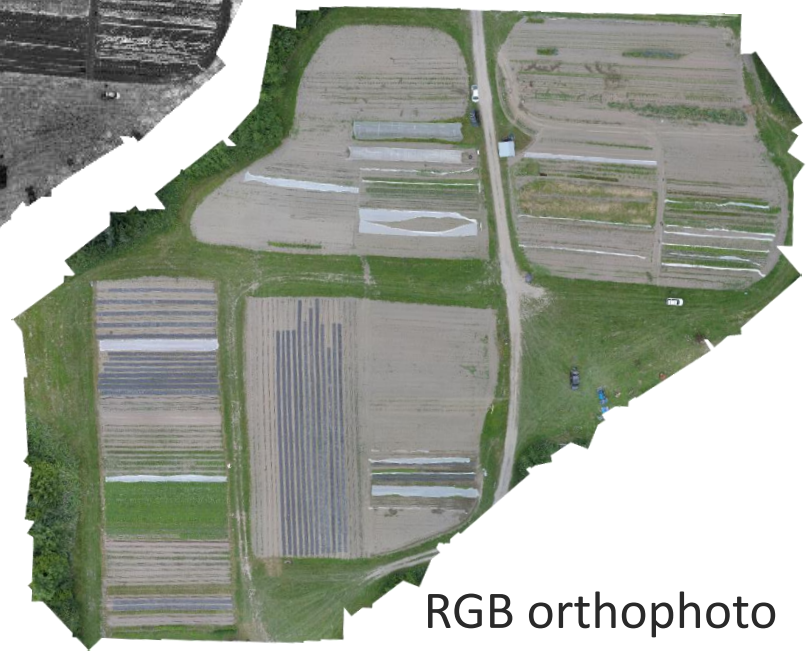
- 3D/4D mapping in agriculture, mining
- Aerial imaging for environment monitoring

Drone Imagery in Precision Agriculture

- Several advantages over satellite imagery
- High-resolution orthomosaic, RGB, multispectral



NIR orthophoto



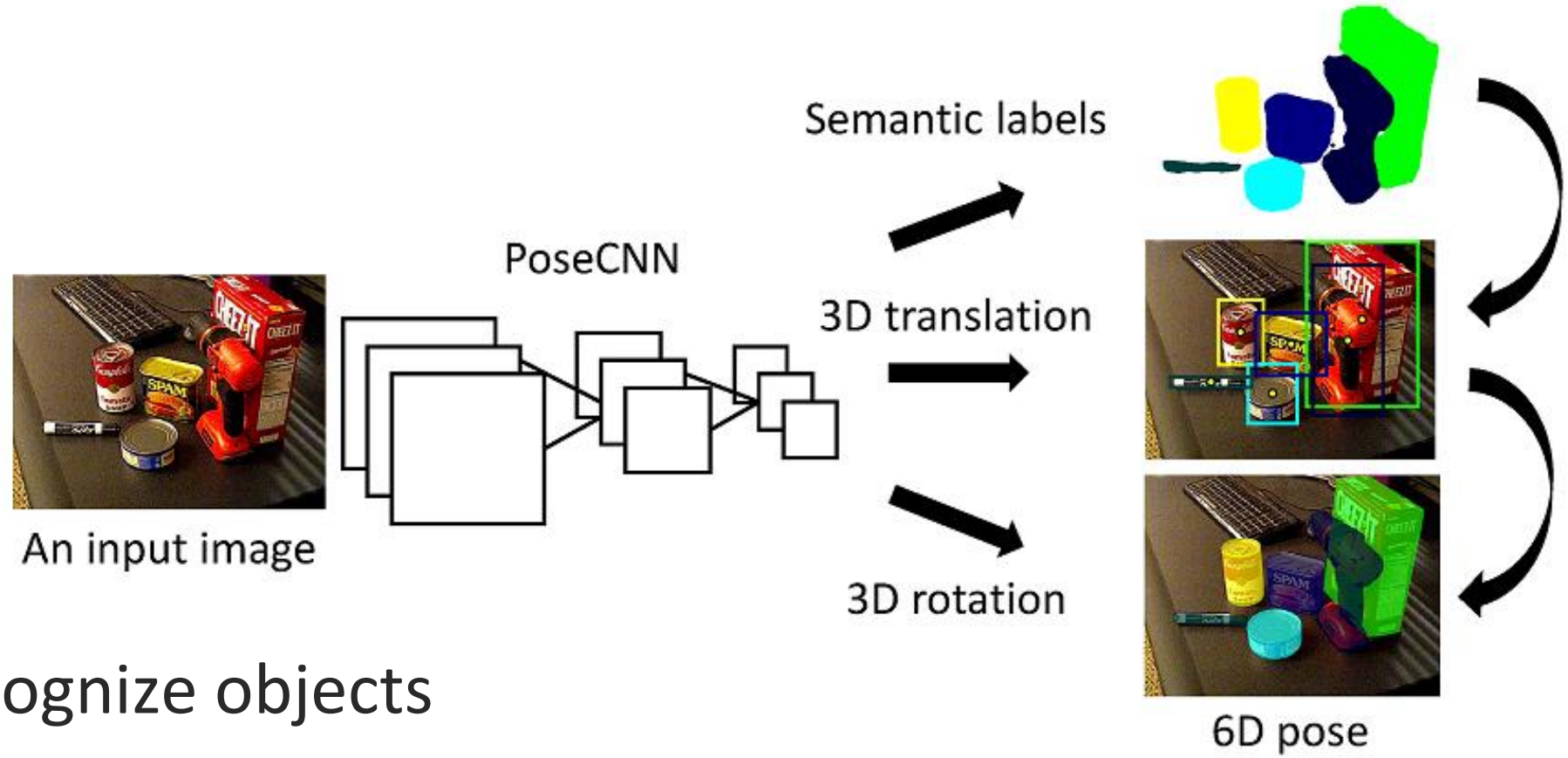
RGB orthophoto

Carnation, WA

near Bangalore, India

4D Reconstruction for Precision Agriculture

Pose Estimation for Robotic Manipulation



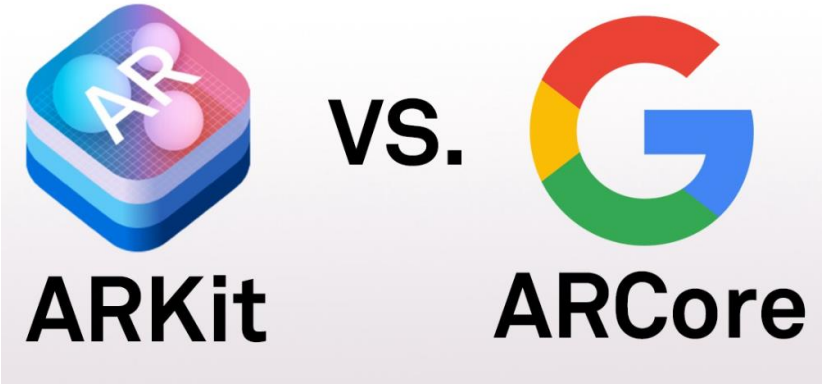
- Recognize objects
- Predict their 6-dof pose
- Active Vision; View Planning

Xiang et al 2018, *PoseCNN*
Robotics, Science and Systems

Augmented Reality

Microsoft HoloLens

MagicLeap

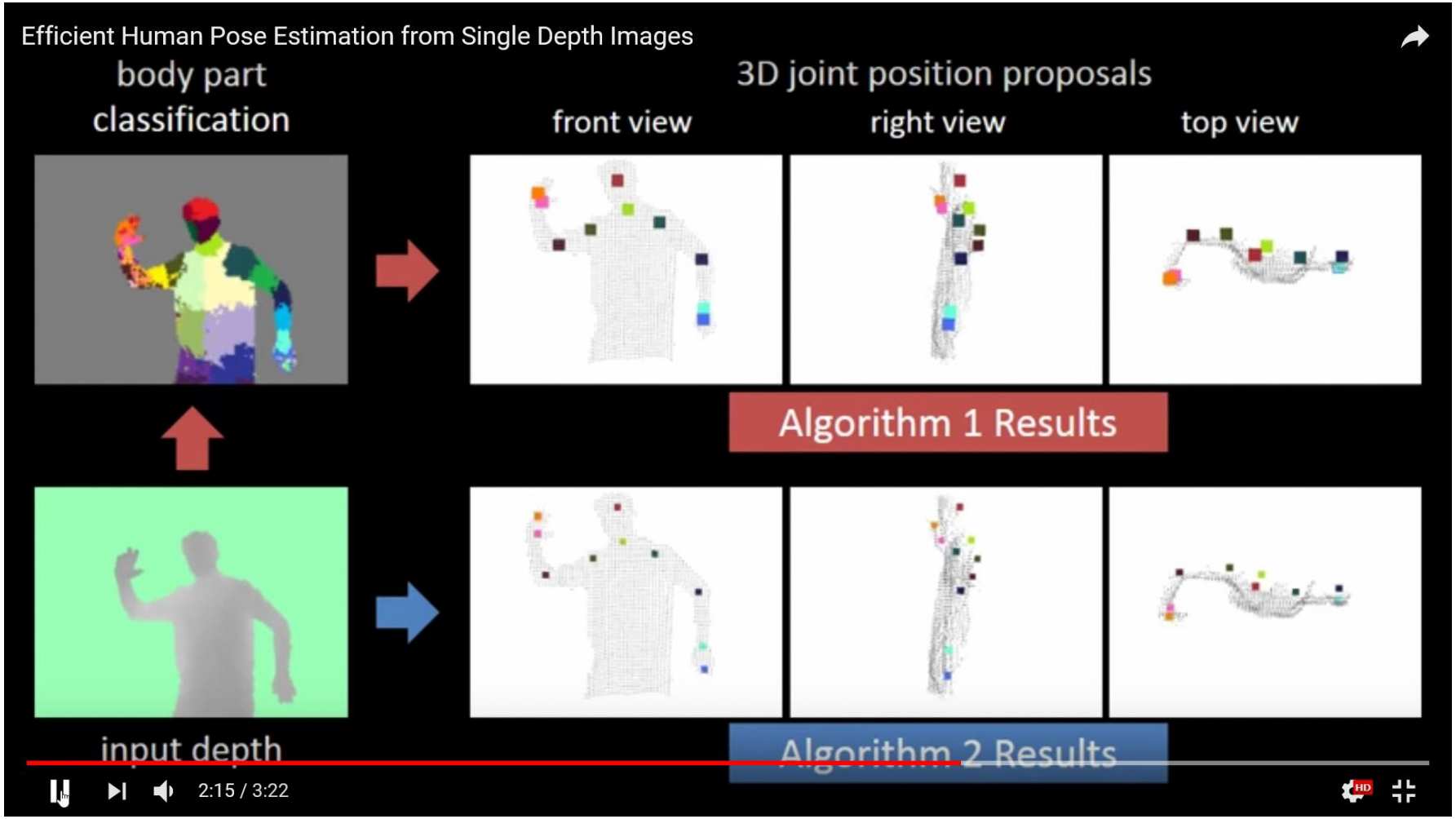


3D Vision challenges

- Camera / Head Tracking
- Spatial 3D mapping
- Hand tracking
- Object detection, pose estimation
- Light estimation and relighting

Gaming, Human Computer Interface

Human pose estimation from a depth image (Microsoft Kinect)

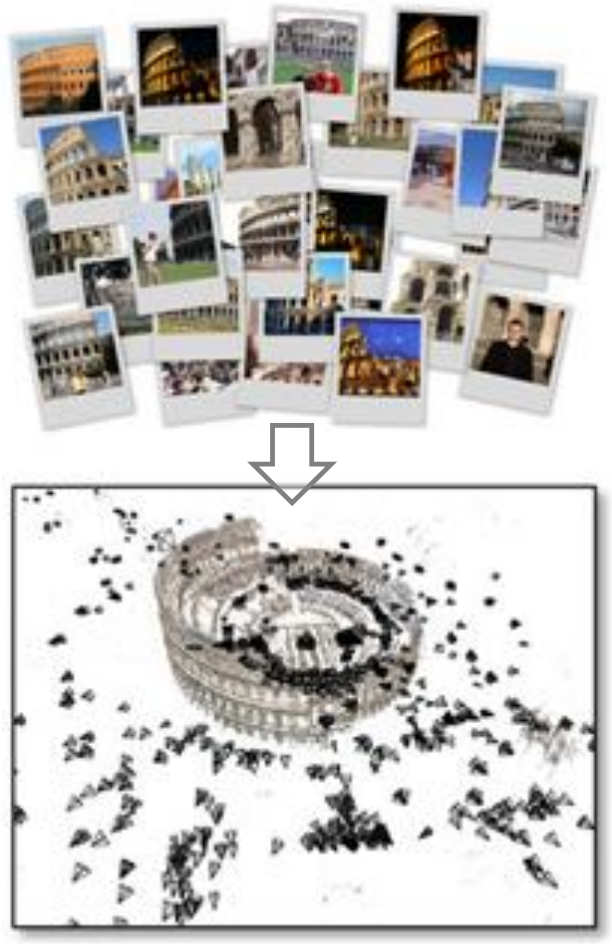


3D Photography

- Image stitching
  - Pure Camera rotation
  - Geometric alignment
  - Photometric alignment
- PhotoTourism

  [Snavely et al. 2006]
- Microsoft Photosynth

Image stitching



Structure from motion

3D scanning, 4D performance capture

- Dou et al. 2016, *Fusion4D: Real-time Performance Capture of Challenging Scenes,* in SIGGRAPH

Overview

- Why study 3D Computer Vision?
- Applications
- Basic Principles and Algorithms
- Preview: 3D vision tasks and algorithms
  - 3D Reconstruction: Structure from Motion, SLAM
  - Camera Localization
  - Image and Video Editing
  - Object detection and pose estimation

Ill-posed Inverse Problem

- 3D Reconstruction is an ill-posed problem
  - many shapes consistent with the same input image.
- Need more information to disambiguate
  - Multiple images, smoothness assumption (Regularization)
- Find the shape, geometry most consistent with assumptions

Shape from X

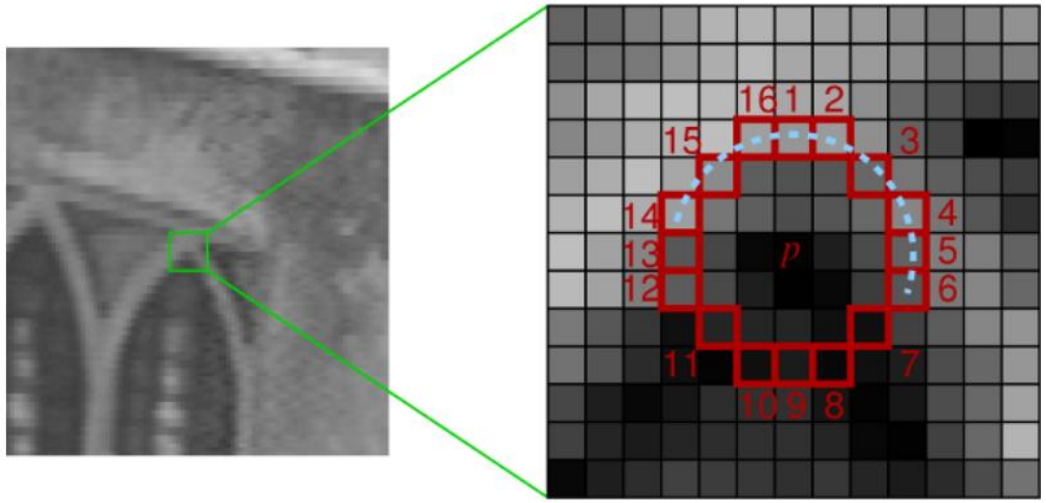
Passive methods

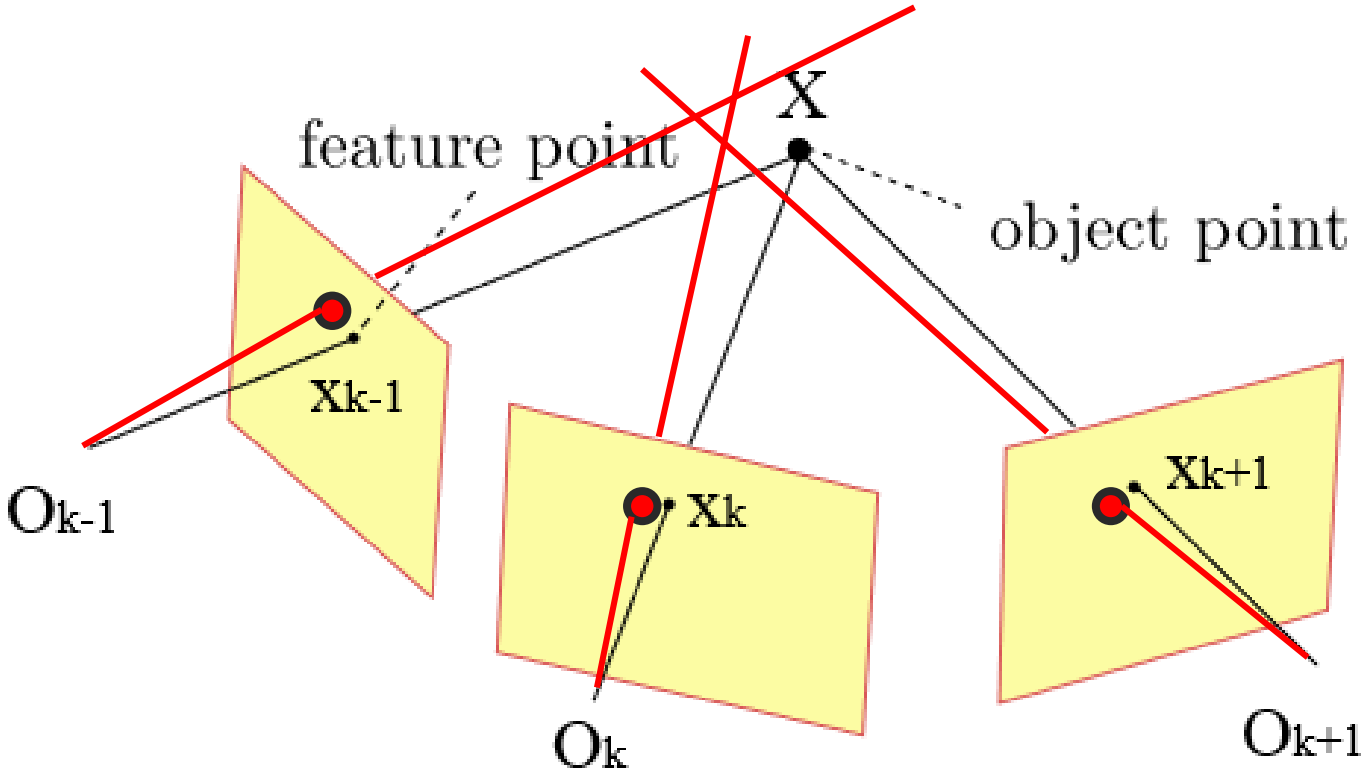
- Structure from motion
- Stereo matching
- Multi-view stereo
- Shape from shading
- Shape from silhouettes

Active Methods

- Structured light
- Time of flight
- Photometric stereo

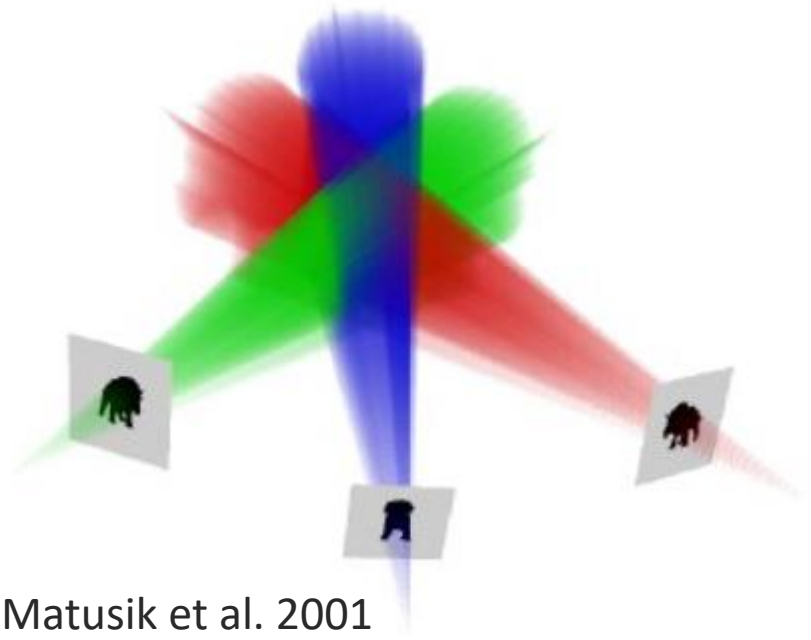
Shape From Multiple View Triangulation





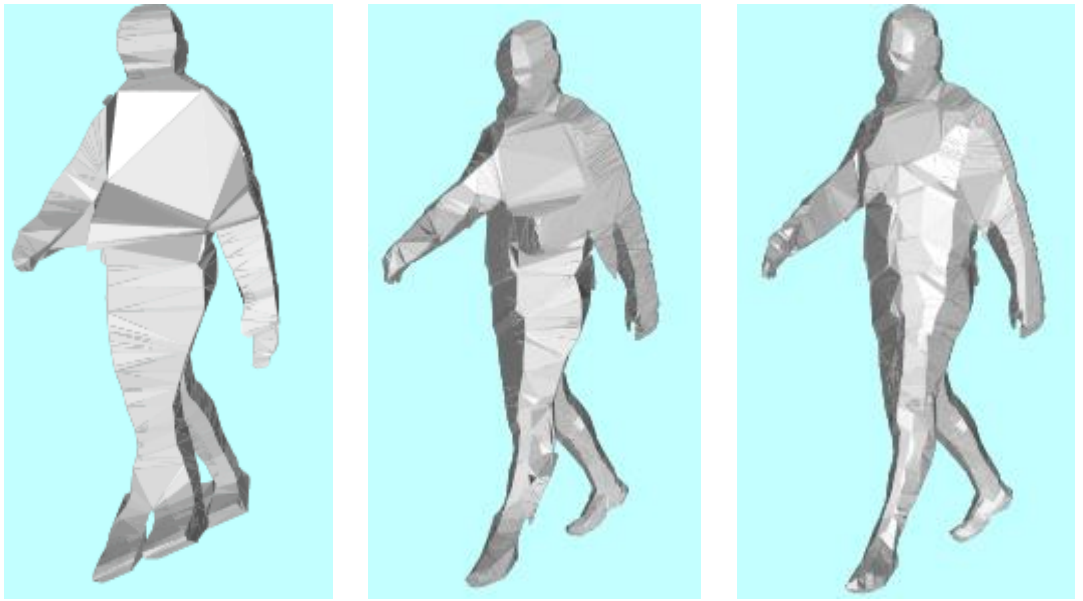
- FAST keypoint detector, [Rosten and Drummond, 2006]
- Given “matched” keypoints in multiple views, obtain the 3D point by *triangulation* of rays.
  - Camera Calilbration (and Pose) is known.
- Effect of noise – rays don’t actually intersect !

Shape from Silhouette



Matusik et al. 2001

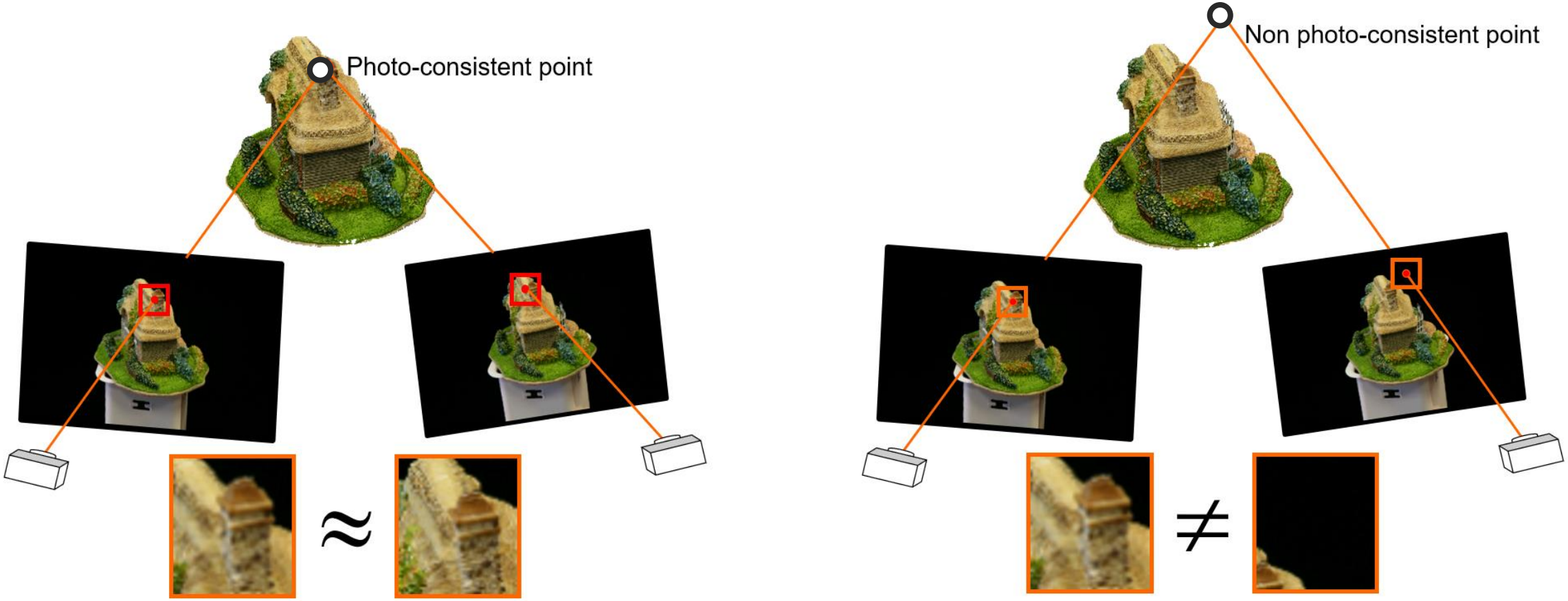
Franco and Boyer 2003



2 views 4 views 6 views

- Input: silhouettes (segmentation) in calibrated images taken from multiple viewpoints.
- Back-project "silhouette cones" and intersect them in 3D to obtain the *visual hull. It is always* an overestimate of the true shape.
- More viewpoints help, but concavities cannot be recovered.

Photo Consistency



For correct depth hypothesis, projected points (or patches) have similar appearance, i.e. they are *more photo-consistent*.

Courtesy: Carlos Hernandez

Multiple View 3D Reconstruction

Hernandez et al. 2004

Vogiatzis et al. 2007

Shan et al. 2013

Images captured with same camera; or taken from Flickr

- Dense 3D reconstruction (scenes, objects, …)
- Mesh with texture maps (Lambertian assumption common)

Key Elements of 3D Reconstruction Methods

- Cameras
  - Calibrated vs. uncalibrated
  - One moving camera, multiple cameras
- Structure
  - Representation (point clouds, meshes, implicit volumes ...)
  - Rigid vs. dynamic scenes
  - Deformable vs. articulated objects
  - Model-based (templates) vs. model-free methods
  - Semantic Reasoning using machine learning techniques

Overview

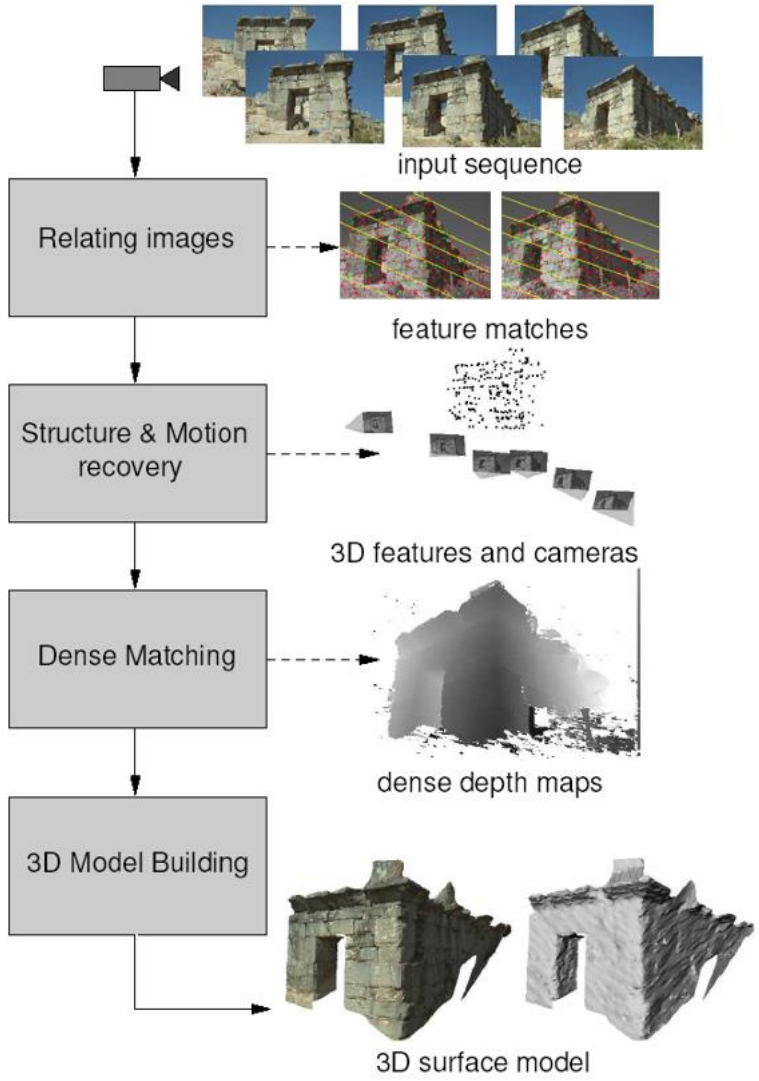
- Why study 3D Computer Vision?
- Applications
- Basic Principles and Algorithms
- Preview: 3D vision tasks and algorithms
  - 3D Reconstruction: Structure from Motion, SLAM
  - Camera Localization
  - Image and Video Editing
  - Object detection and pose estimation

Structure from Motion, SLAM

3D reconstruction pipeline

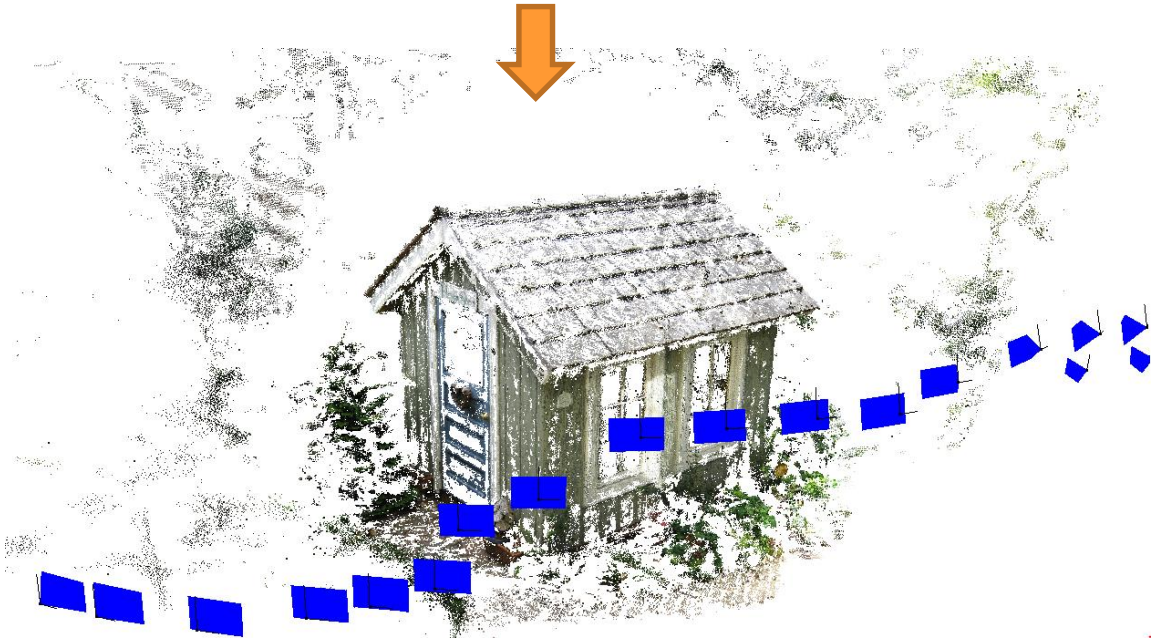
- Feature Extraction and Matching
- Structure From Motion
  - Relative pose problem
  - Absolute pose problem
- Bundle Adjustment
- Dense Stereo Matching
- Depth-map Fusion
- Texture-mapping

> Textured 3D Model



Structure From Motion

Sparse Pixel Correspondence



Structure (3D points) and Motion (camera pose)

Structure From Motion

Given observations $U = \{x_{ij}\}$, $x_{ij} \in R^2$

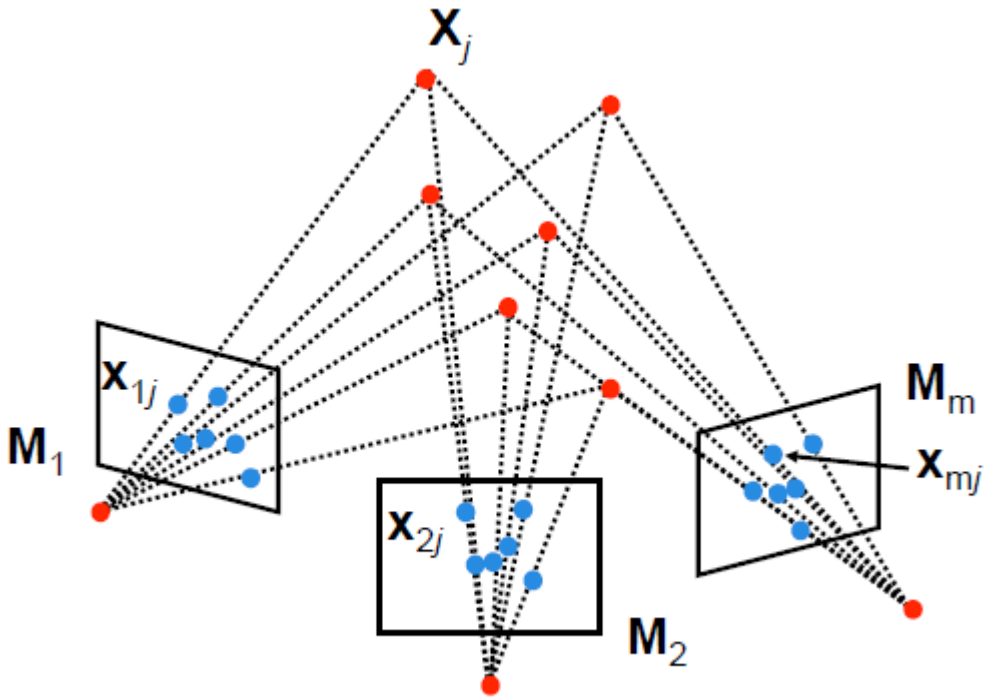
of n 3D points in m cameras,

recover the 3D points P and cameras C,

$$P = \{X_j\},\ X_j \in R^3, j \in [1 \ldots n]$$

$$C = \{M_i\},\ i \in [1 \ldots m]$$

where, $M_i = [R_i \mid t_i] \in R^{3\ x\ 4}$ denotes the pose (3D rotation, position) of the i-th camera.



Approach: minimize the re-projection error in all the images

Structure From Motion on Internet Photos

COLMAP [Schonenberger+ CVPR 2016]

- **ROME:** 75K images
- 21K cameras, 5+ million points
- Running Time: 3 hours (with GPUs)

Bundle Adjustment based on Ceres Solver
http://ceres-solver.org/

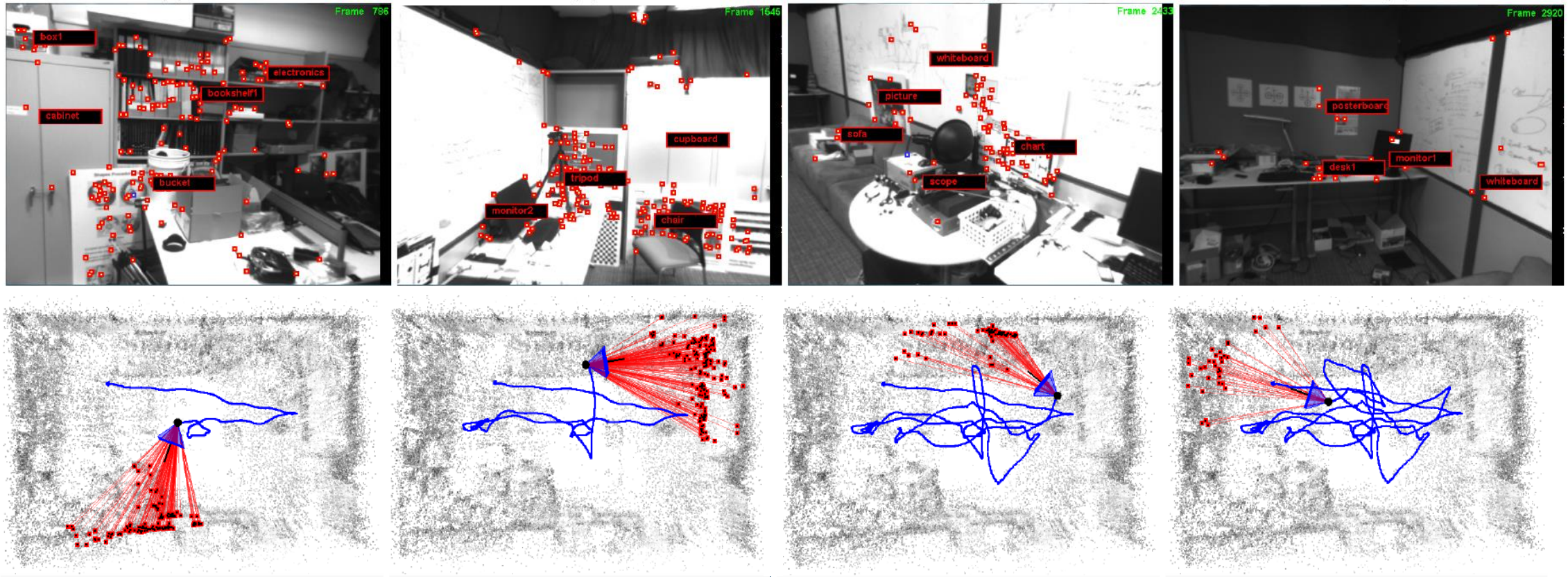
Simultaneous Localization and Mapping

Mur-Artal et. al. 2015, "ORB-SLAM: a versatile and accurate monocular SLAM system." *IEEE Transactions on Robotics* 31.

Image-based Localization

Image-based 6-DoF Camera Localization

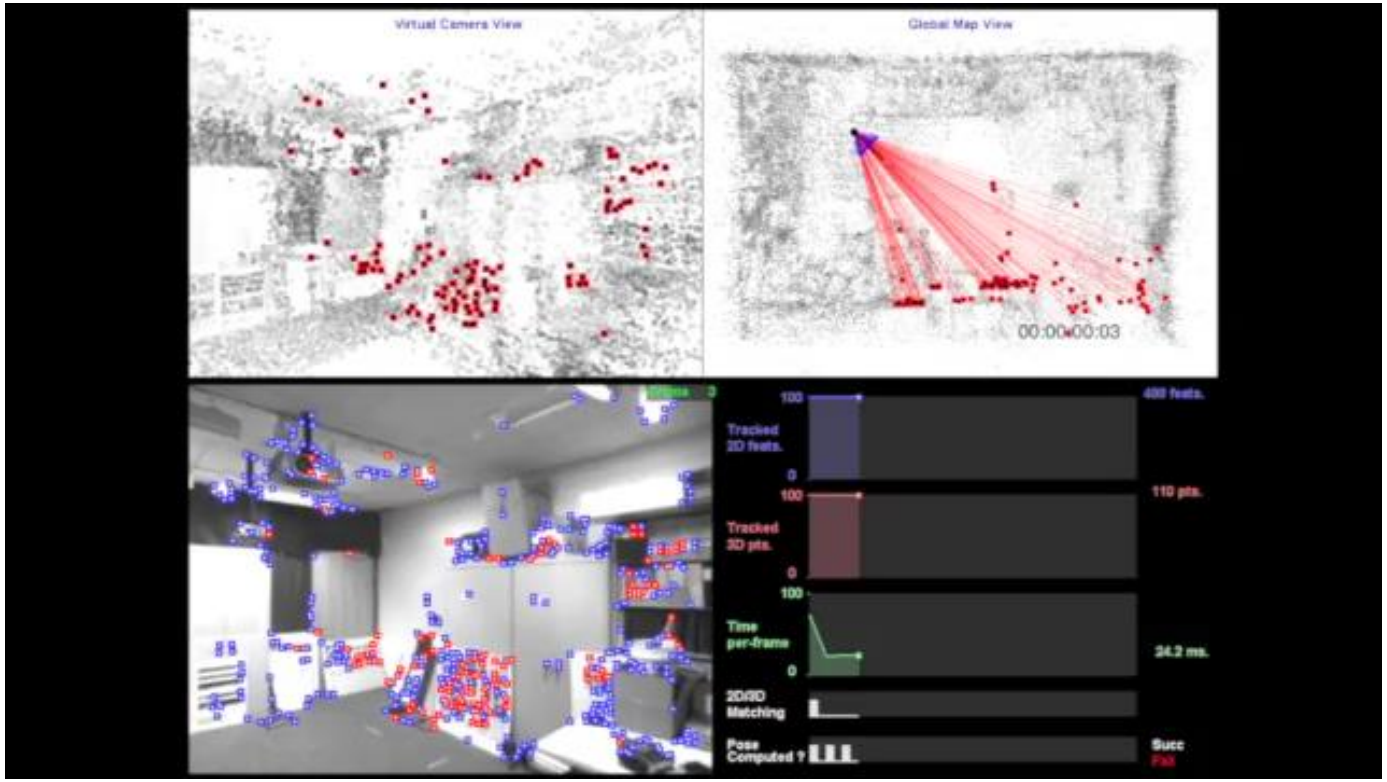
Lim et. al. 2015, "Real-time monocular image-based 6-DoF localization “,
The International Journal of Robotics Research (IJRR).



Goal: Compute **exact position and orientation** of query image relative to 3D scene model.

Image-based 6-DoF Camera Localization

Lim et. al. 2015, "Real-time monocular image-based 6-DoF localization “, *The International Journal of Robotics Research (IJRR).*



Goal: Compute **exact position and orientation** of query image relative to 3D scene model.

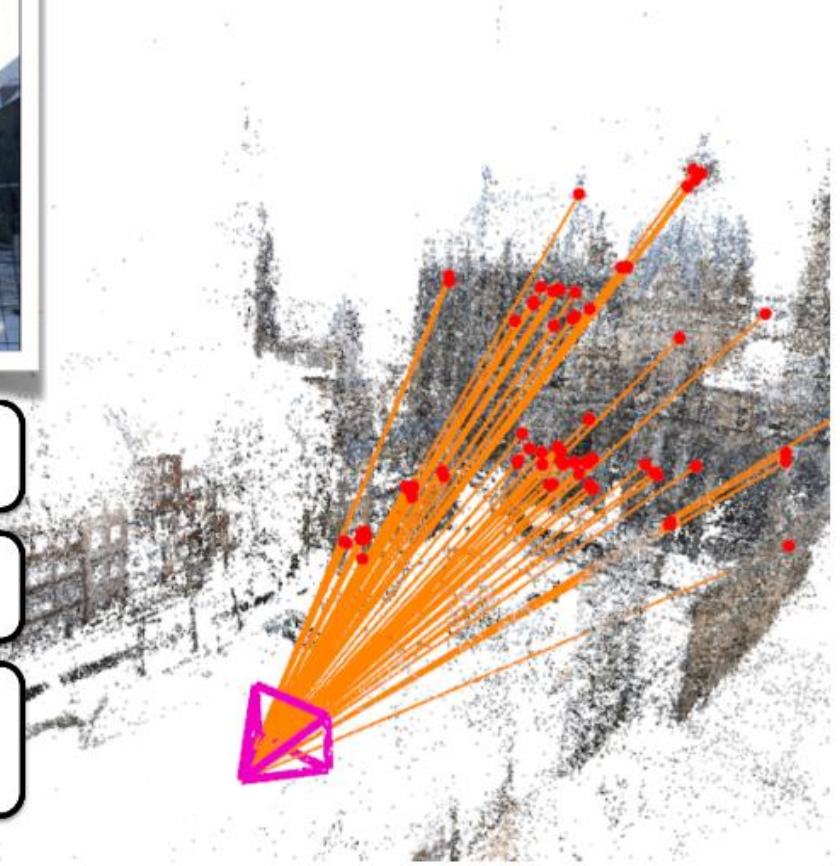
6-DoF Camera Pose Estimation

Feature-based approach:

Extract Local Features

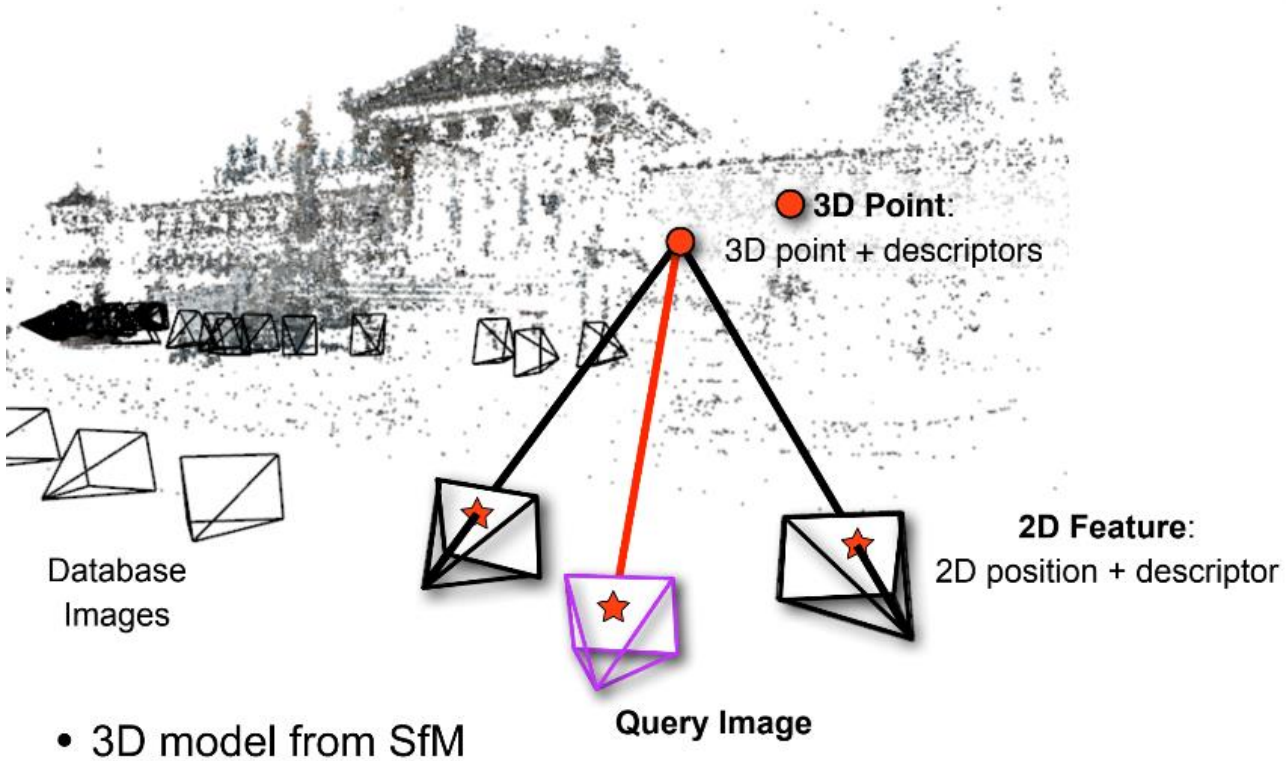
Establish 2D-3D Matches

Camera Pose Estimation:
RANSAC + n-Point-Pose Algorithm

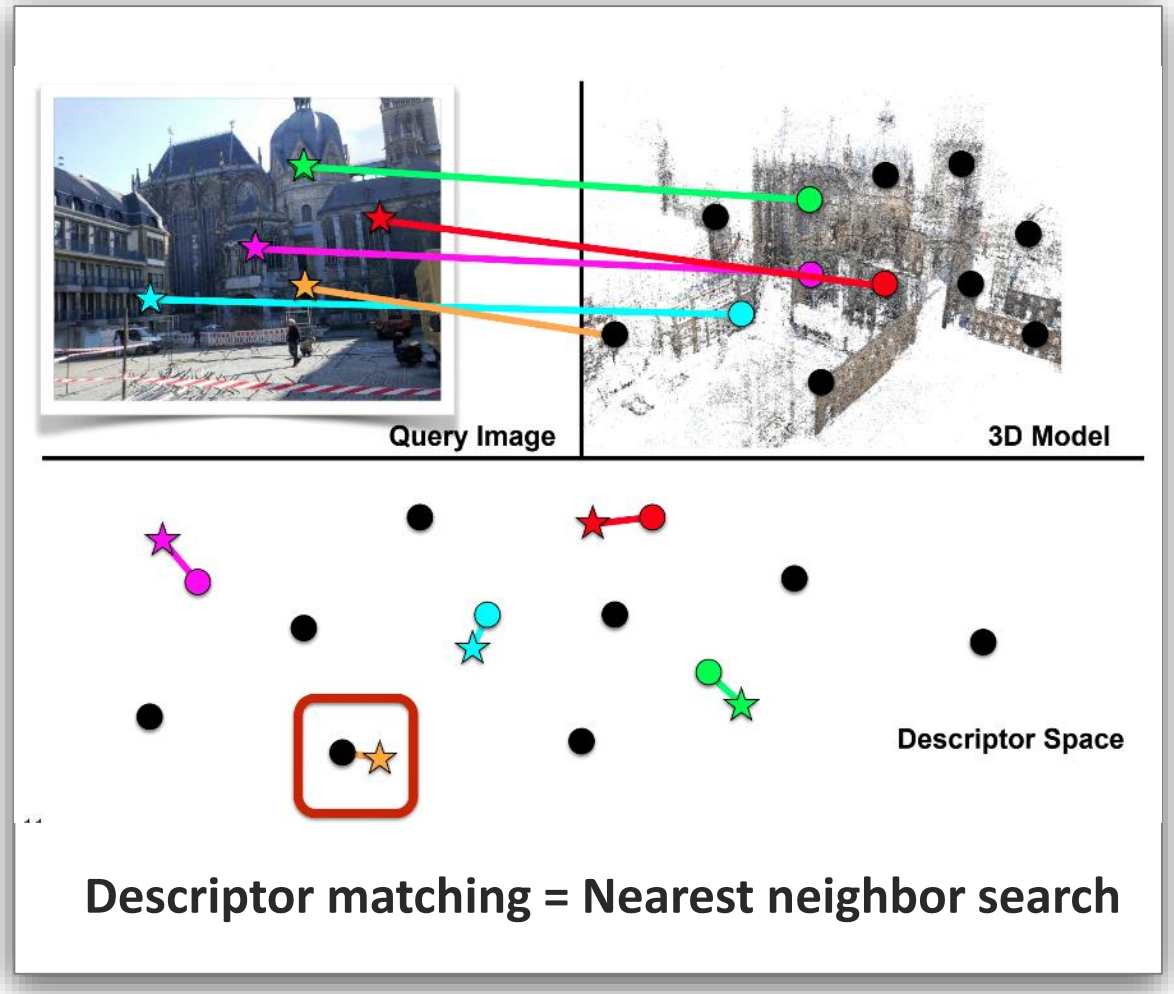


6-DoF Camera Pose Estimation

1. Establish 2D-3D Matches



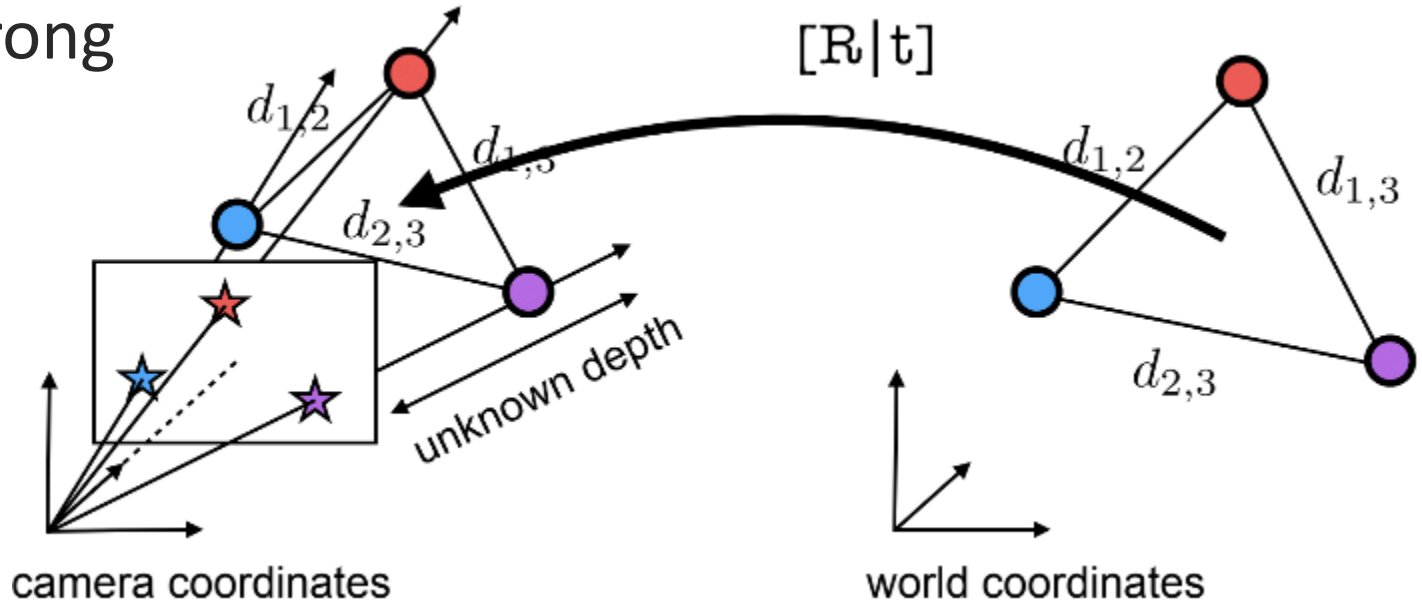
- 3D model from SfM
- 2D-3D correspondences from (SIFT) **descriptor matching**



6-DoF Camera Pose Estimation

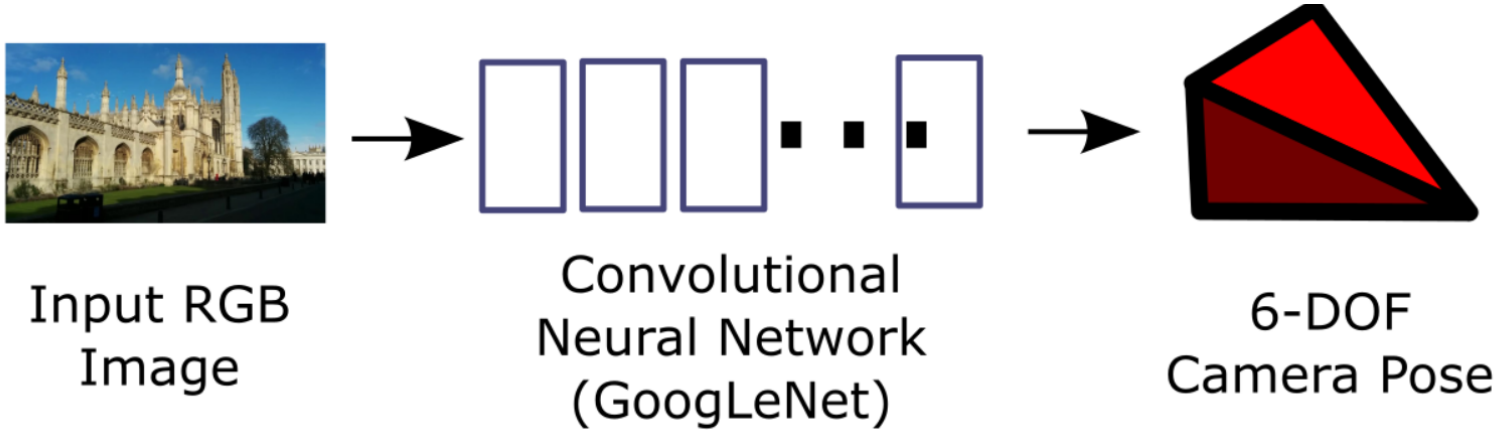
1. Camera Pose Estimation

- 2D – 3D matches can be wrong
- Robust estimator needed



- RANSAC
  - Minimal solvers
  - 3 point solver (P3P)
  - 4th point to disambiguate
- Non linear Optimization

6-DoF Camera Pose Estimation



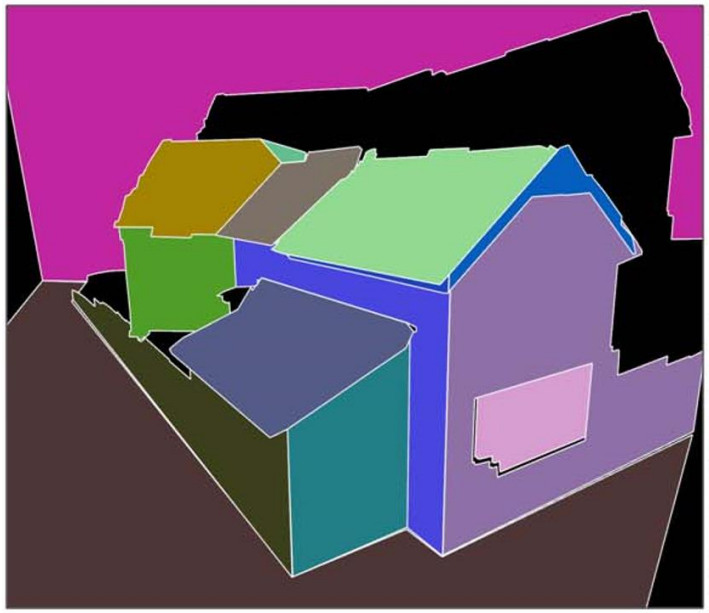
PoseNet [Kendall+ ICCV 2015]

- End to end 6-DoF Camera Pose Regression using CNNs
- more robust to blur, drastic light changes, fast (on GPU)
- **Not yet as accurate as feature based methods**

  (see CVPR 2017 visual localization tutorial)

Novel View Synthesis

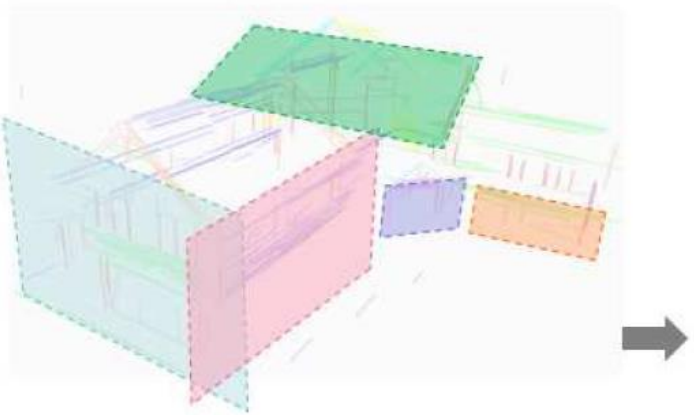
Novel View Synthesis



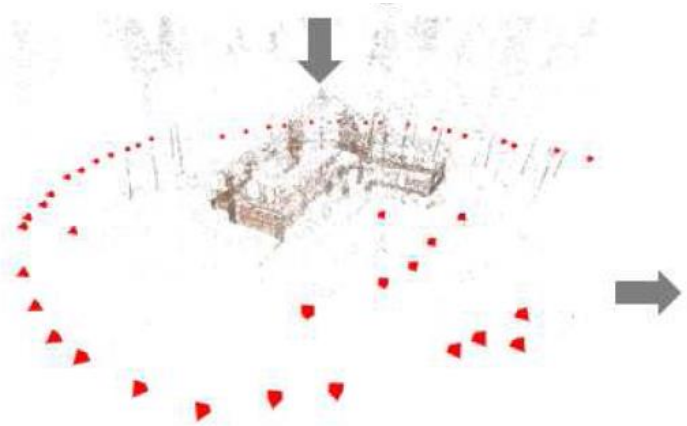
- Image with Piecewise Planar Depth Map; camera is calibrated.
- Novel view synthesized by warping source image via depth map; black pixels are holes (must be inpainted).

Stereo matching with planar priors

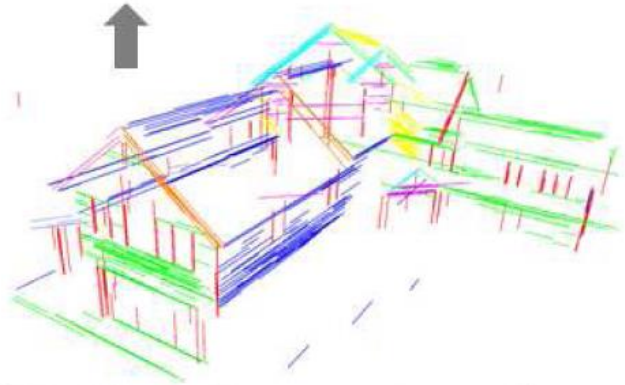
Sinha, Steedly and Szeliski, ICCV 2009



Multiple Plane Detection



Structure from motion

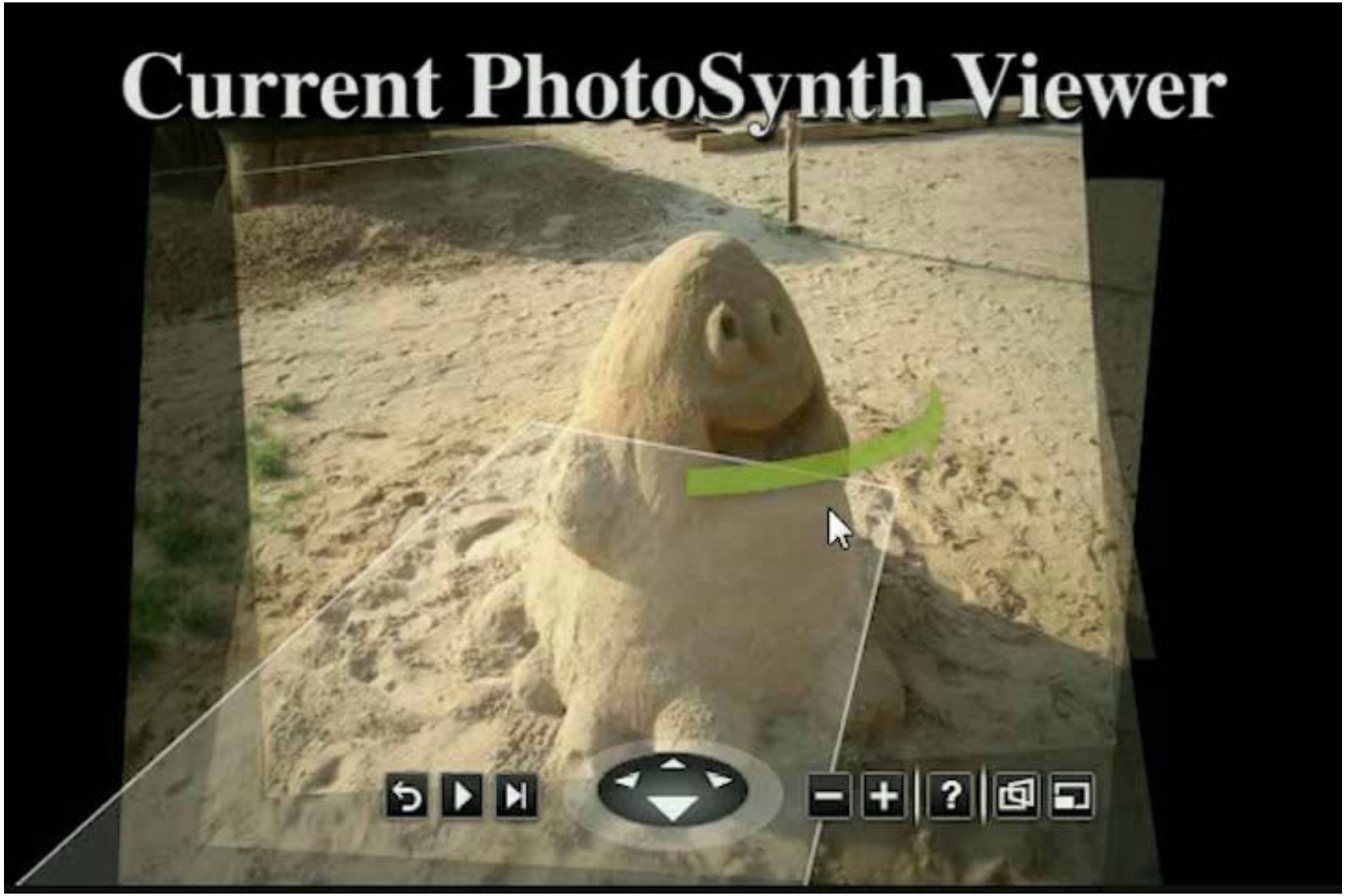


3D Line Reconstruction

Markov Random Field (MRF) optimization

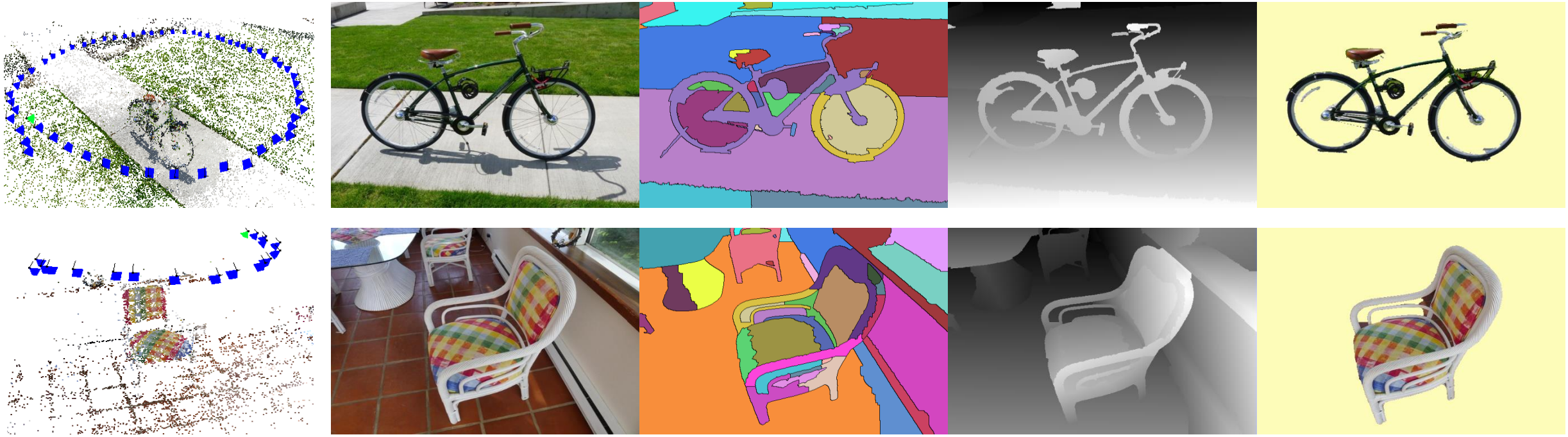
Novel View Synthesis

Novel View Synthesis



Stereo and Segmentation with Planar Priors

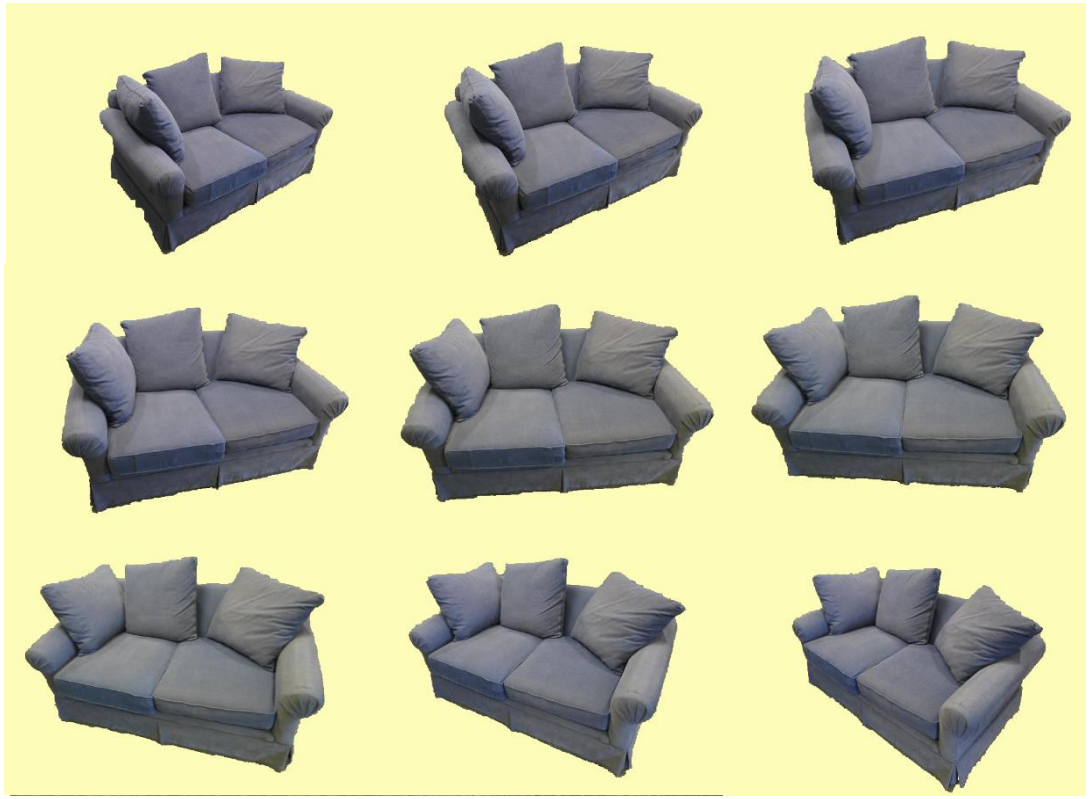
Kowdle+ 2012, "Multiple View Object Cosegmentation using Appearance and Stereo Cues“, in ECCV.



- Multiple views of an object; cameras calibrated using structure from motion
- First estimate a piecewise planar depth map from each view
- Object of interest cue
- Multi-view consistency constraints used; segmentation is automatic

Stereo and Segmentation with Planar Priors

Kowdle+ 2012, "Multiple View Object Cosegmentation using Appearance and Stereo Cues“, in ECCV.



Object 6-DoF Pose Estimation

3D Recognition, 2D-3D Model Alignment

In a single RGB image, recognize the object; predict 3D position and orientation.

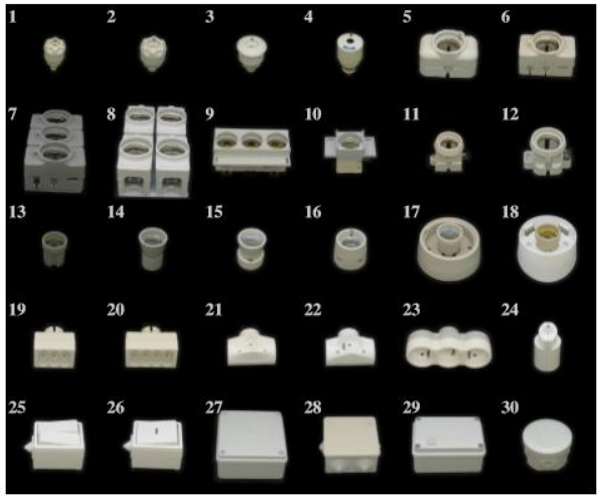
Lowe 2001

Rothganger+ 2005

Lepetit+ 2005

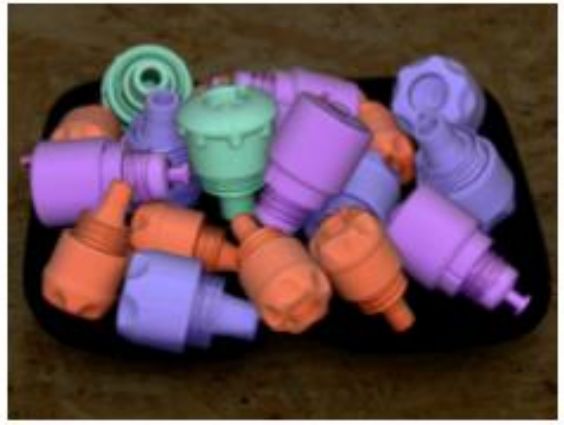
- 2D—3D matching (model is now a small object)
- PnP pose estimation
- Works for textured, distinctive objects

Texture-less Object 6D Pose Datasets



LINEMOD [2012]

15 objects



T-LESS [2017]

30 objects

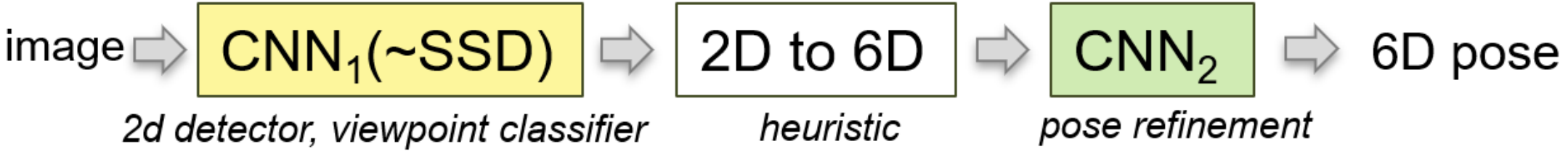
YCB-VIDEO [2018]

21 objects

Deep 6-DoF Object Pose Estimation

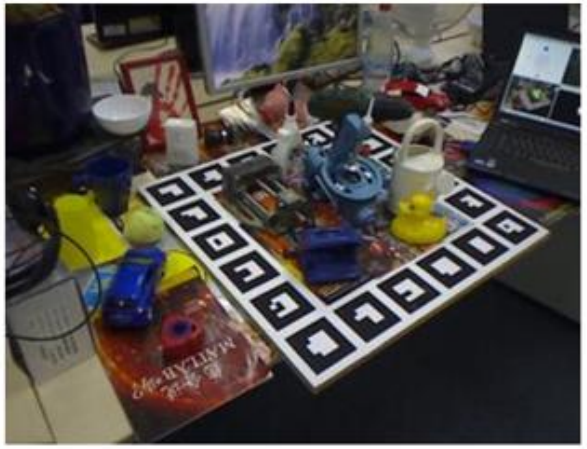
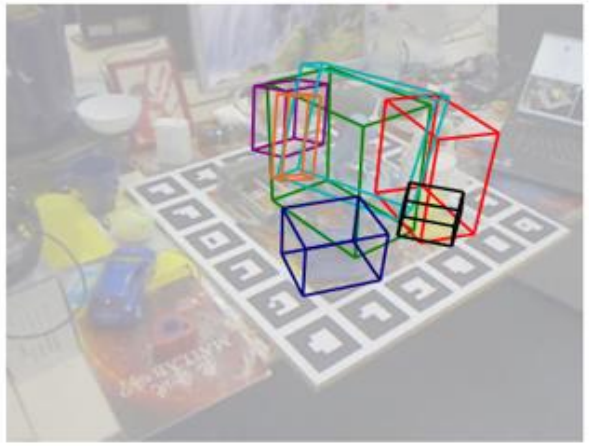
- BB8 [Rad and Lepetit 2017]

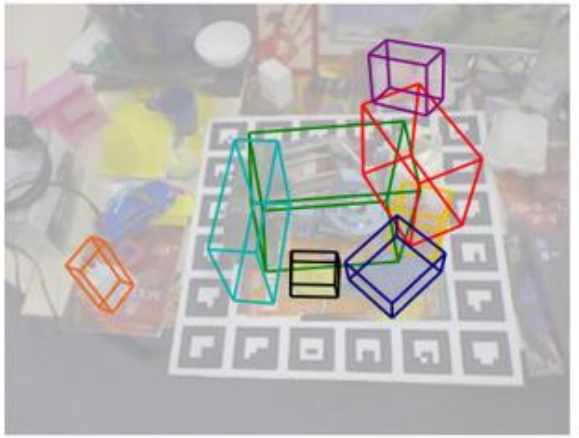
- SSD-6D [Kehl+ 2017]



- Single shot pose [Tekin+ 2018]

Single Shot 6D Object Pose (Tekin+ 2018)



- Inspired by the YOLO Object detector
- CNN predicts 2D projections of 3D bounding box corners. Then, run PnP solver on just eight 2D--3D matches.
- Accurate, fast (50-90 fps);
- Detects multiple objects in one pass.

Summary

- 3D Computer Vision: Applications and Insights
- Core ideas behind 3D reconstruction algorithms
- Sneak Preview:
  - Structure from Motion,
  - SLAM
  - Camera Localization
  - Object detection and pose estimation
  - Novel View Synthesis

Next Session:
Multiple View Geometry