3D Reconstruction from Multi-View Video Streams

Collaborators at UW:


Kalyani Marathe


Mahtab Bigverdi


Sadjyot Gangolli


Nishat Khan


Linda Shapiro


Ranjay Krishna

Collaborators at Amazon:


Ariel Gordon


Michael Wolf

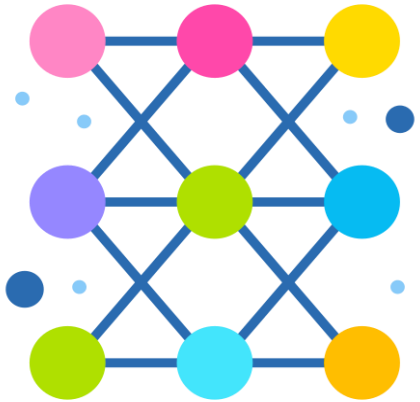
"Amazon's Sparrow is designed to identify and handle millions of warehouse [...] inventory items."

Problem formulation: Input multi-view images -> output depth

Input

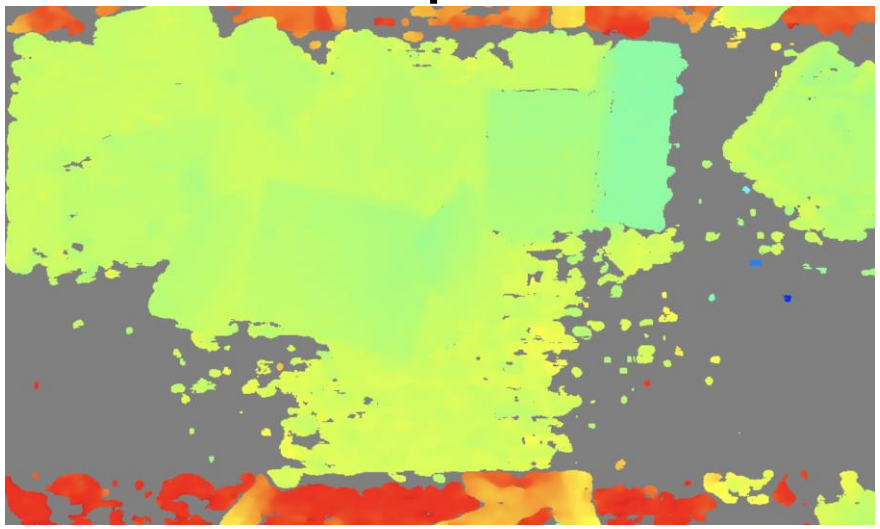
1-3 RGB images of the conveyor belt + packages

Model



Deep learning based
Adaptable
Real time
Low cost

Output



Depth map from the robotic arm
Accurate enough for robotic manipulation

Environment Challenges

Millions of packages

Dynamic Environment

Changes in lighting conditions

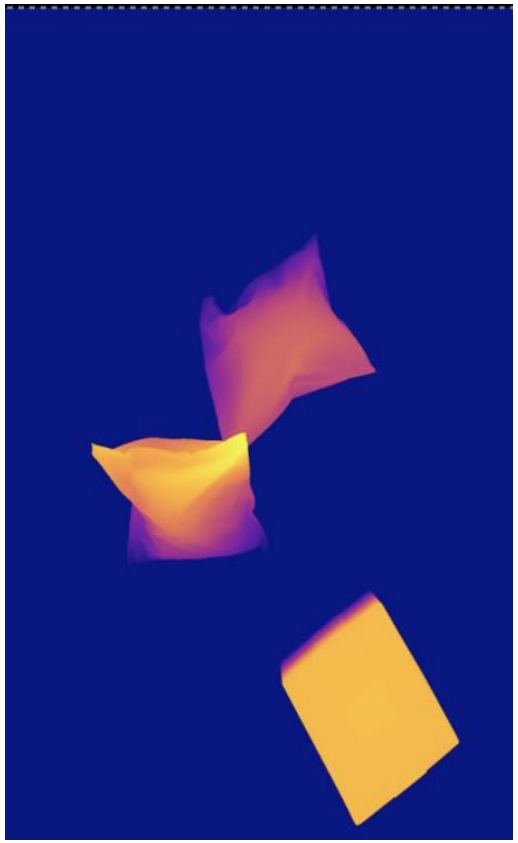
Object Challenges

New package types

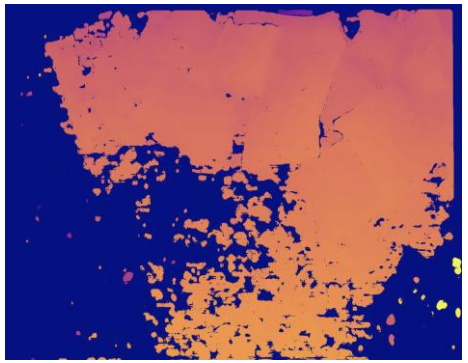
Art on packages

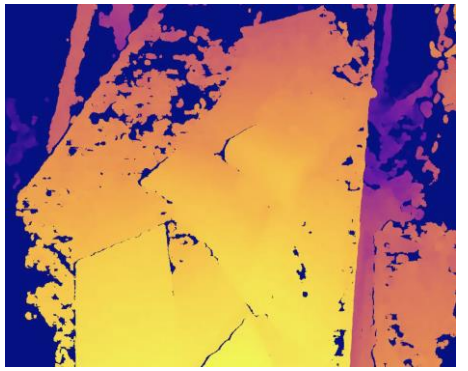
Transparent packages

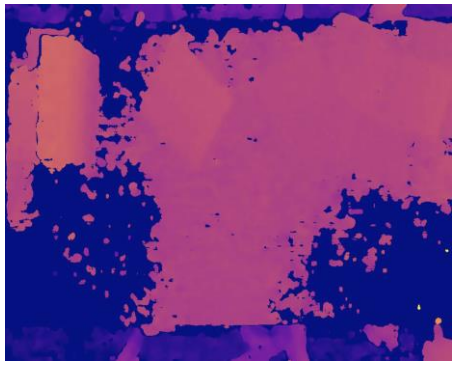
The data we have is small



50k samples of Synthetic dataset –
RGB images of conveyor and ground truth depth maps

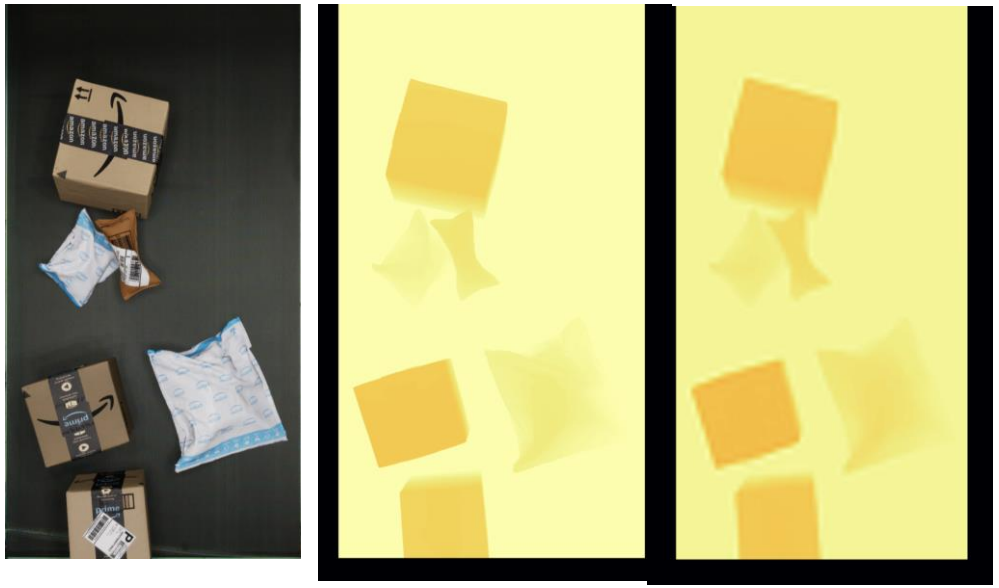




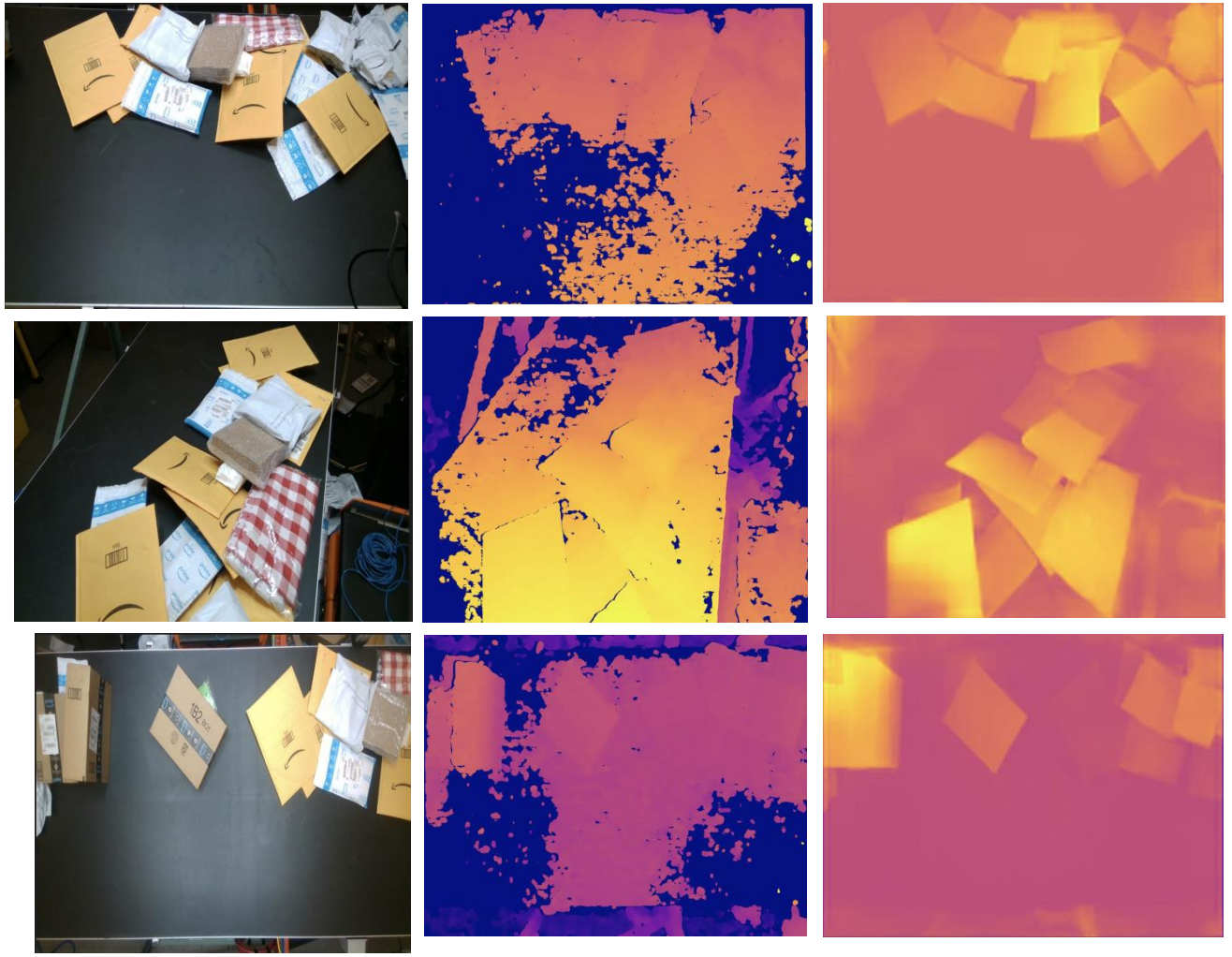


120 sets of real dataset -
RGB images from 3 viewpoints and ground truth depth maps obtained from sensors
T = 0, t= 1, t= 2.... , t = 119

Our attempts work well for synthetic data



RGB, Ground truth and Predicted depth maps
(Synthetic data)



RGB, Ground truth and Predicted depth maps
(Real data)

NewCRFs, Yuan et al, CVPR 2022

Much larger errors for real images than for synthetic images

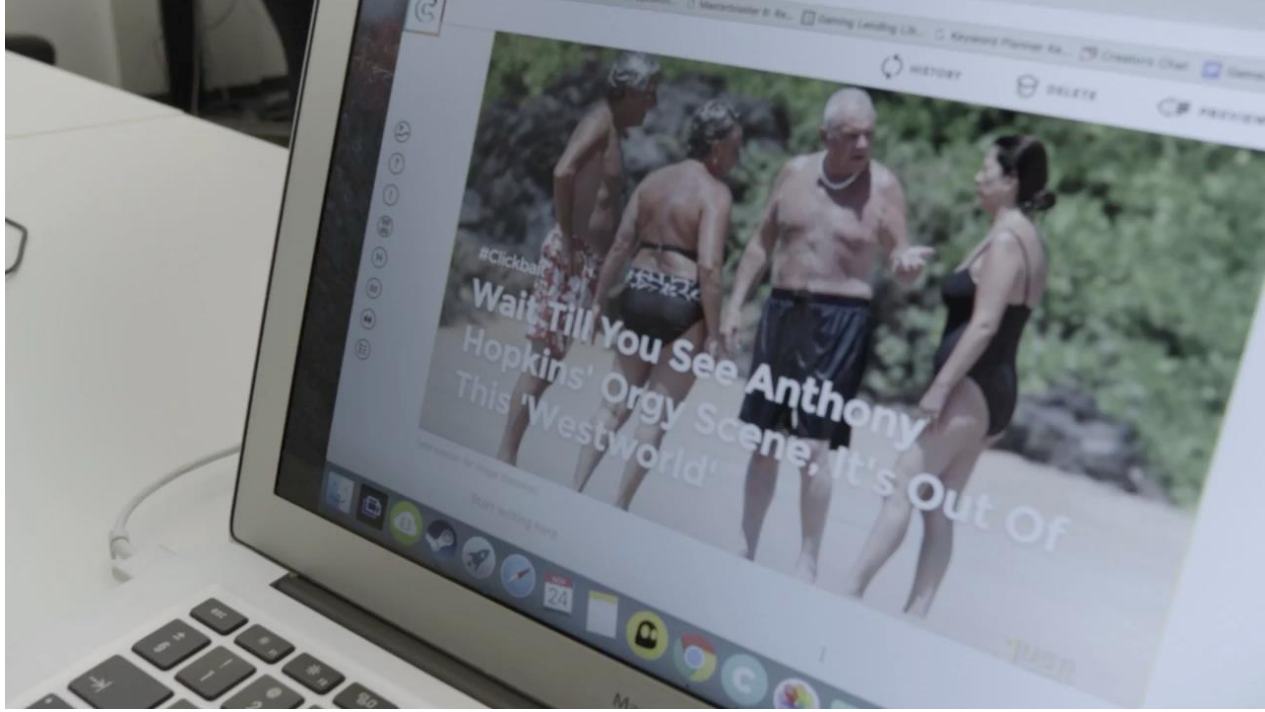
Test dataset	Sample Image	Scale Invariant Logarithmic Error
Held out synthetic test data		0.40
Real data (top view)		12.75
Real data (side view)		19.43
Real data (front view)		26.68

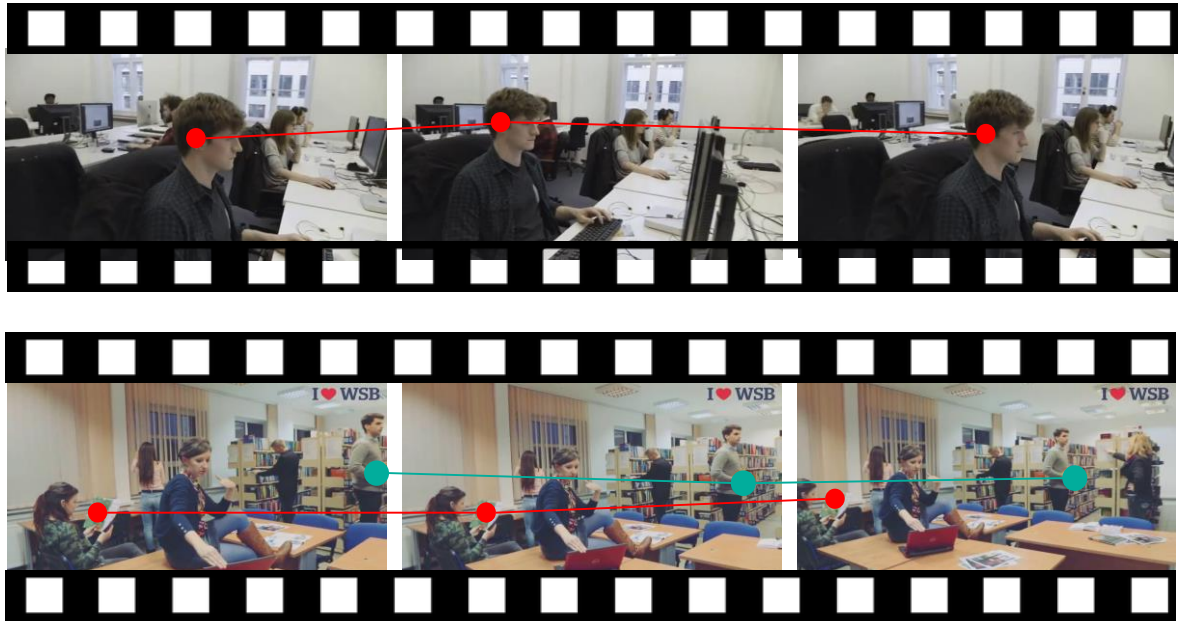
Why current models fail?

- Synthetic data is too easy, and not representative of real world challenges
- We are developing domain adaptation methods to improve synthetic to real world images.

- All existing state of the art depth models use an underlying feature extractor trained on ImageNet.
- We are developing new methods that will learn the 3D structure.

Extracting pixel correspondences from thousands of online videos





Calculating the pixel correspondences

- Use SIFT and other pixel features
- RANSAC to calculate Homography between pairs of frames

Current Work

- Three Ras are working on the project
- They are working on data collection, model training, and evaluation.
- We will hear about it from them.