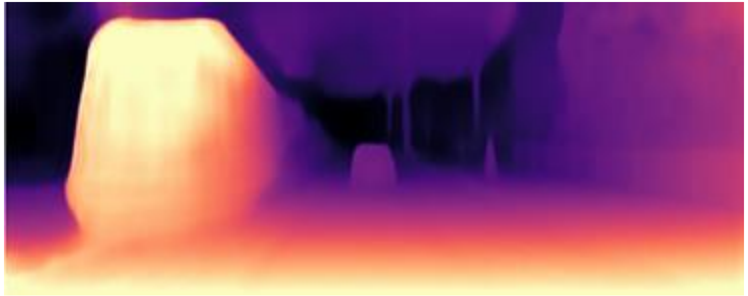
CS5670: Computer Vision

Noah Snavely

Recent work on predicting 3D geometry

RGB Image

Deep learning



Depth map

Announcements

- Final exam in class on Monday
  - Will cover material from the entire class
  - Open book / open note (please bring notes within reason)
  - Please organize yourselves so that you are seated with at least one space between yourself and your neighbor
- Quiz 4 has been graded
- Please give us feedback! Fill out course evaluations here (for bonus points!):
  - https://apps.engineering.cornell.edu/CourseEval/
- Office hours today 2-3pm in Bloomberg 365

Single-view modeling

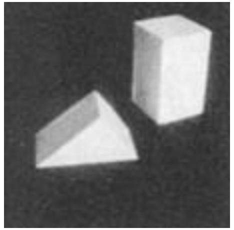
Vermeer's *Music Lesson*

Reconstructions by Criminisi et al.

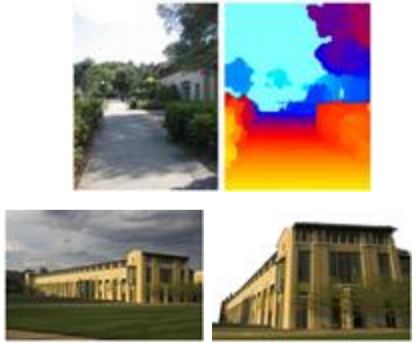
Can we use deep learning to predict geometry from a single image?

Stepping back: Astonishing progress in learning 3D perception

“Blocks world”
Larry Roberts
(1963)

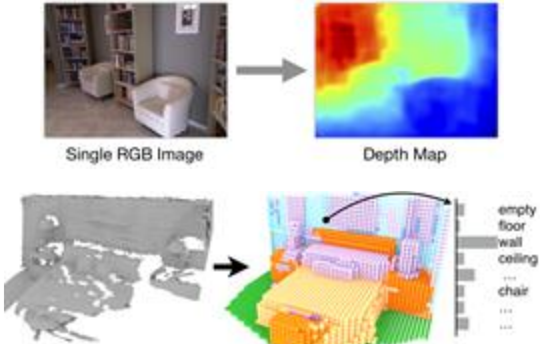


Pre-deep era
(2005)



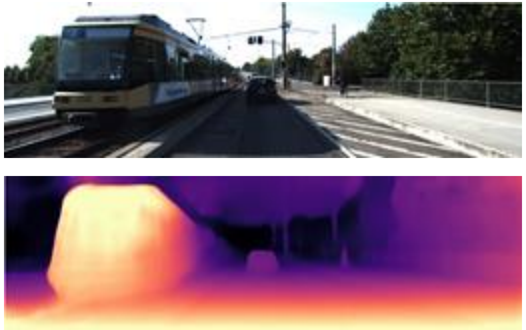
[Saxena, Chung, Ng, NIPS 2005]
[Hoiem, Efros, Hebert, SIGGRAPH 2005]

Supervised deep learning
(2014)



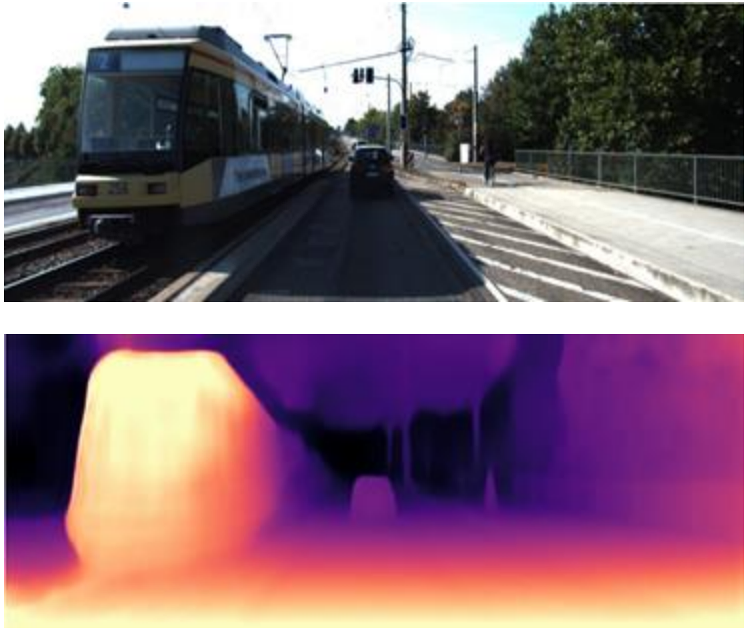
[Eigen, Puhrsch, Fergus, NIPS 2014]
[Song et al, CVPR 2017]
...
go/im2depth

Multi-view supervision
(2016)

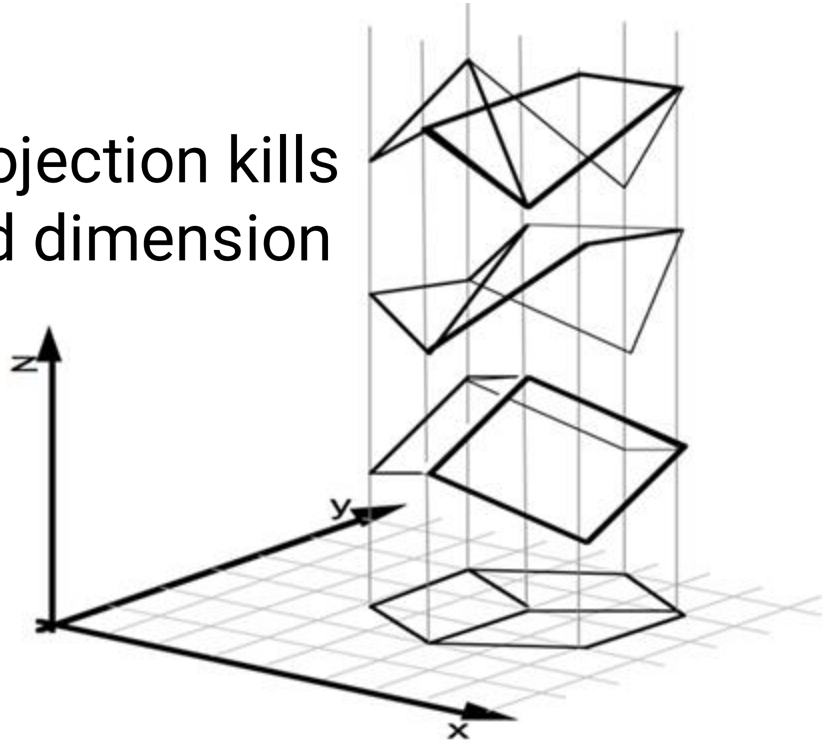


[Garg, Kumar BG, Carneiro, Reid, ECCV 2016]
[Xie, Girshick, Farhadi, ECCV 2016]
[Zhou, Brown, Snavely, Lowe, CVPR 2017]
[Vijayanarasimhan, et al., 2017]
[Godard, Mac Aodha & Brostow, CVPR 2017]
[Mahjourian, Wicke & Angelova, CVPR 2018]
...

Canonical problem: single-view depth prediction

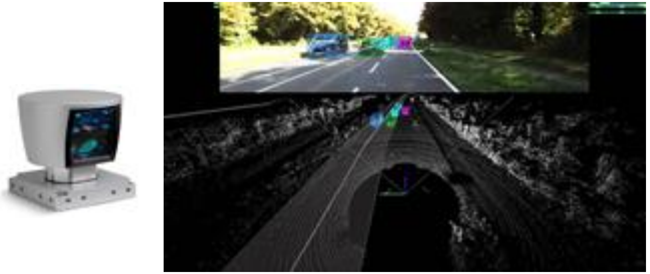


but projection kills the 3rd dimension



[Sinha & Adelson, 1993]

Training data



KITTI [Geiger et al. 2012]

NYU [Eigen et al. 2014]

Depth in the Wild [Chen et al. 2016]

Direct, real-world training data is limited for geometric problems

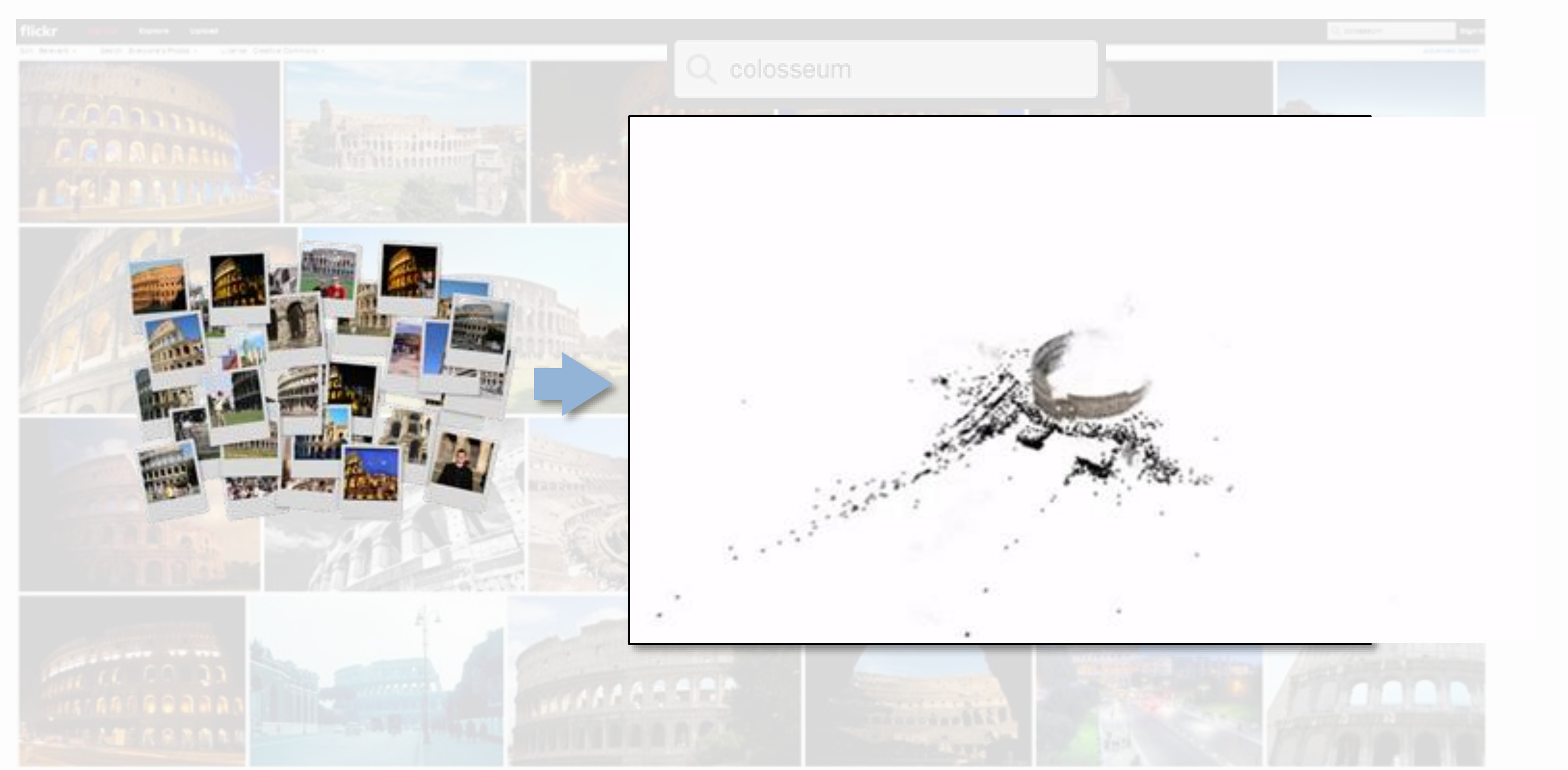
How can we gather more diverse data?

Can we learn 3D from simply observing all the images / videos on the Internet?

Training: Multiple views

Testing: Single Image

Another source of training data



[Snavely, Seitz, Szeliski. *Photo Tourism*. SIGGRAPH 2006]

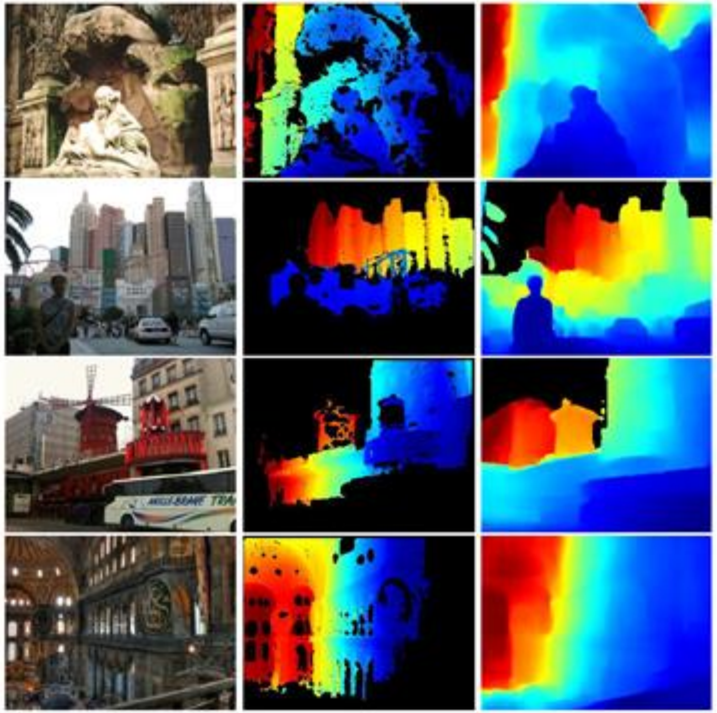
MegaDepth dataset

>130K (RGB, depth map) pairs

- generated from 200+ landmarks
- reconstructed with SfM + MVS using COLMAP [Schoenberger et al]

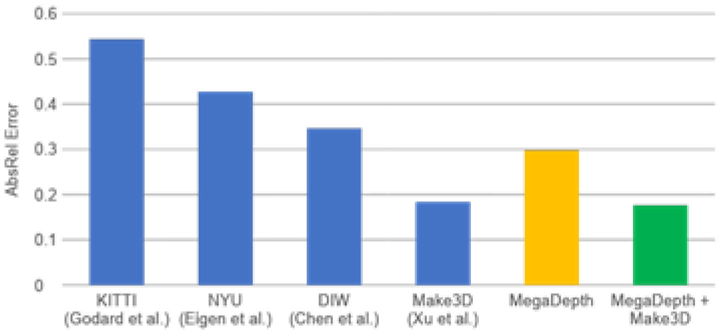
[Zhengqi Li and Noah Snavely. *MegaDepth: Learning Single-View Depth Prediction from Internet Photos*. CVPR 2018]

MegaDepth-trained prediction results

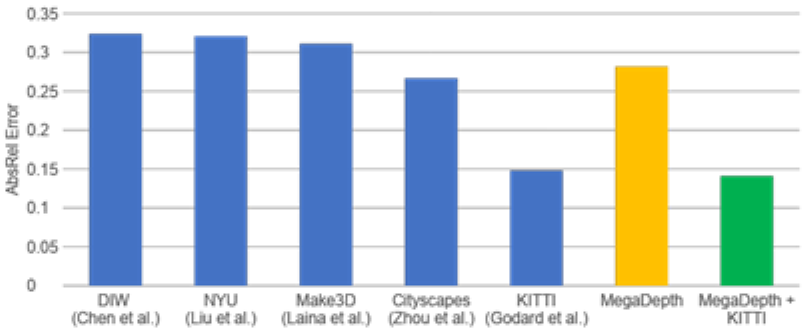


Input Ground truth Predicted depth

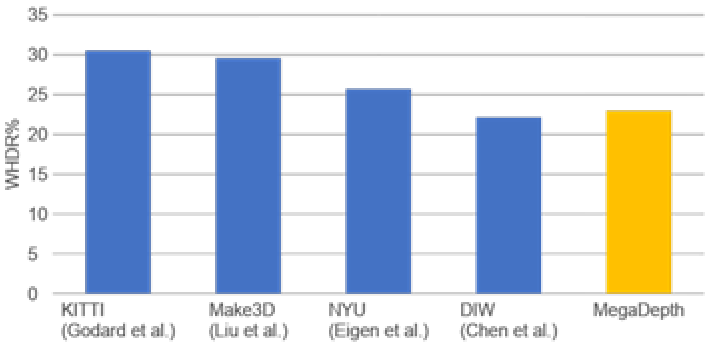
Internet data generalizes well



Train on X, test on Make3D

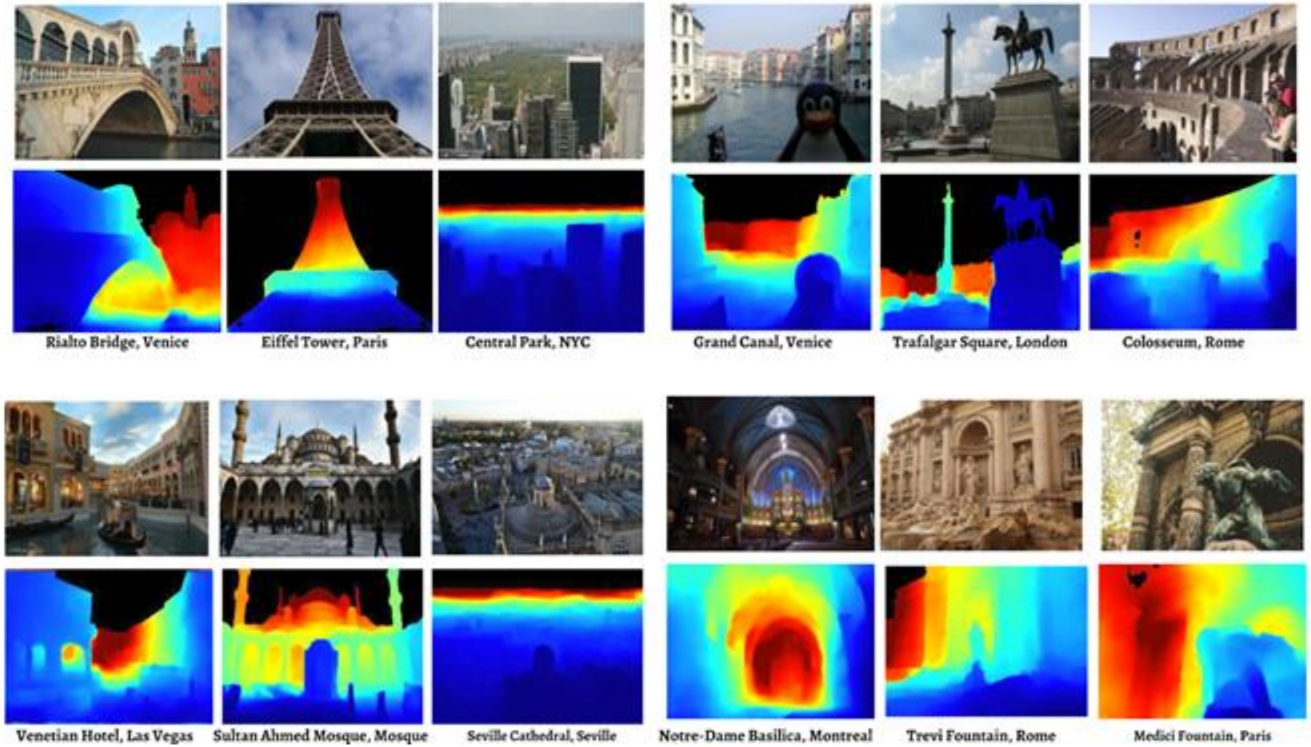


Train on X, test on KITTI



Train on X, test on DIW

More depth prediction results



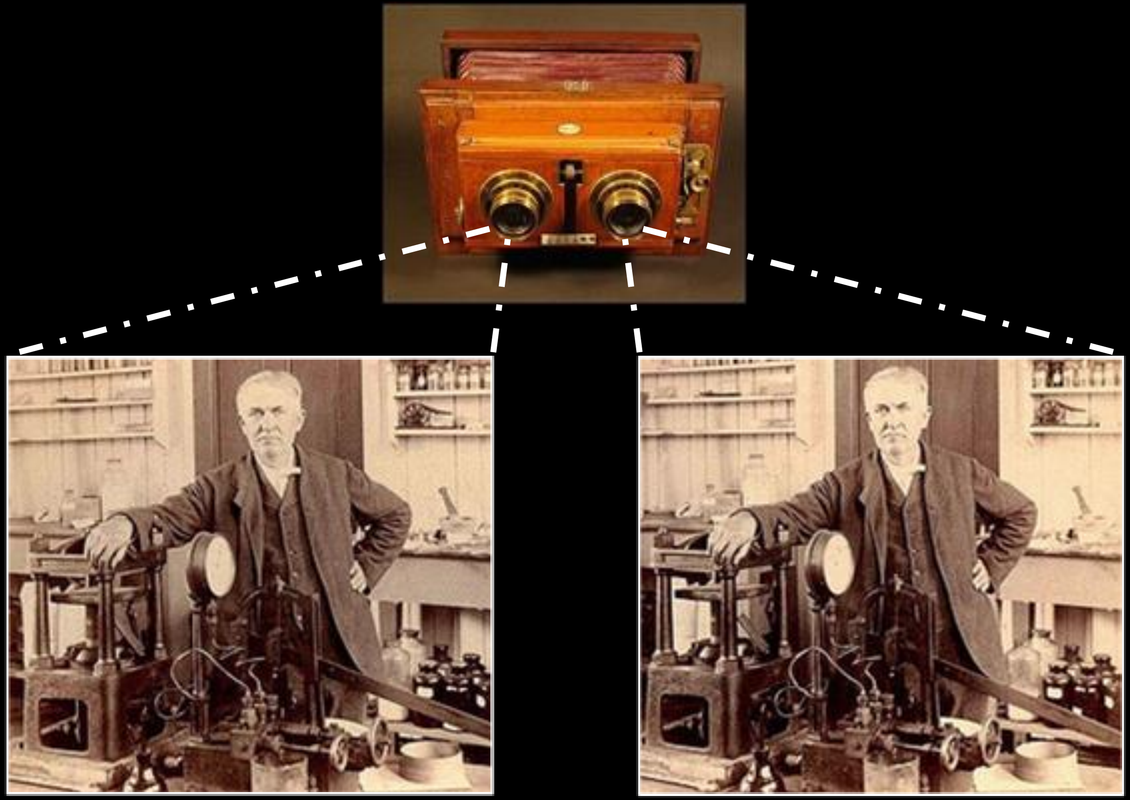
MegaDepth dataset

- All data, including images, SfM reconstructions, and depth maps available at

bitly.com/megadepth

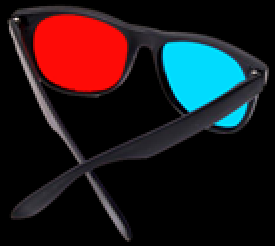
- Reconstructions also useful for other tasks, e.g. learning feature correspondence

Stereo Photography



Stereo Photography

Viewing Devices



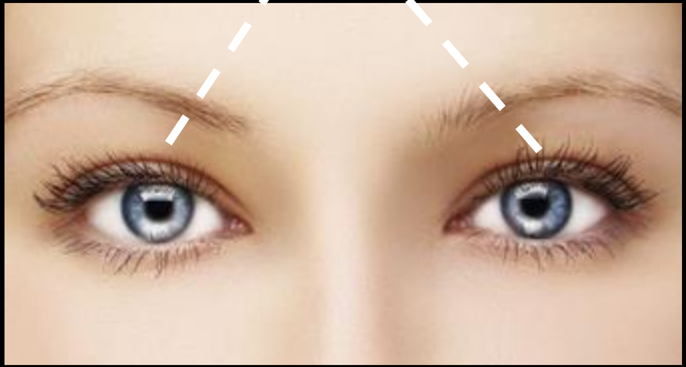
Stereo Photography

Queen Victoria at World Fair, 1851

Stereo Photography

Issue: Narrow Baseline

~6.5 cm



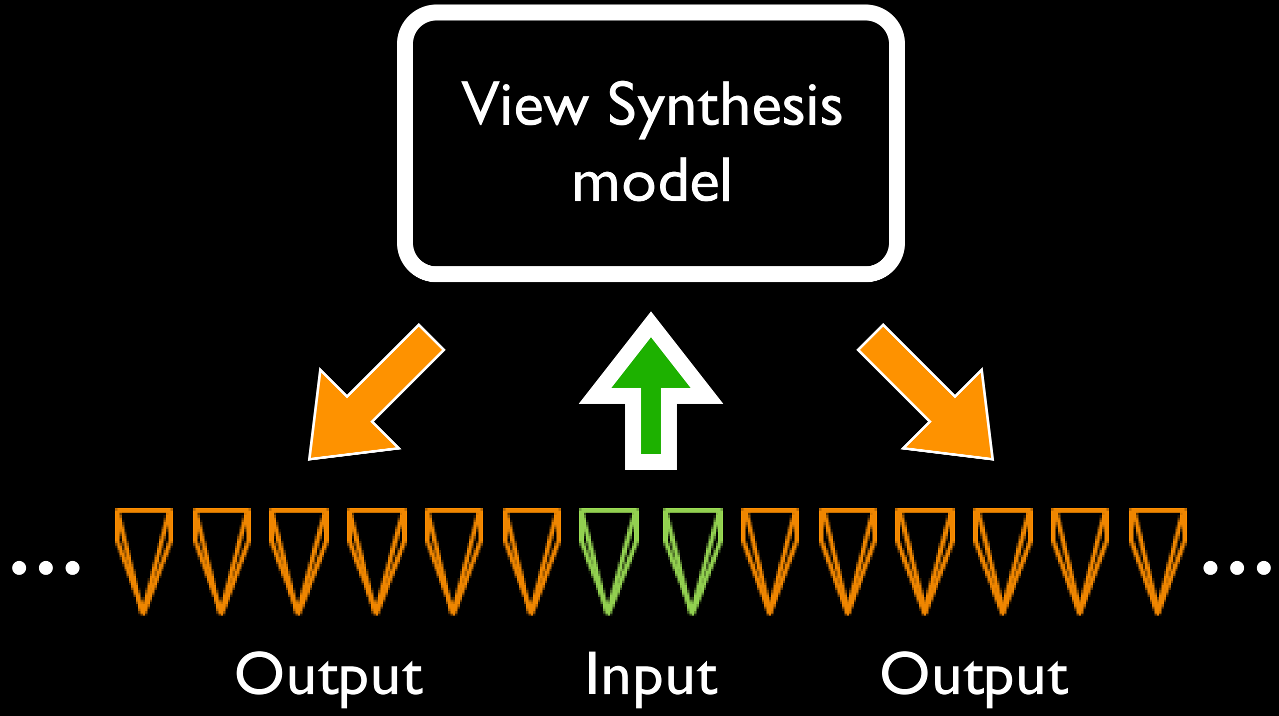
~1.5 cm


Left
MAZZER


Right
MAZZER


MAZZER

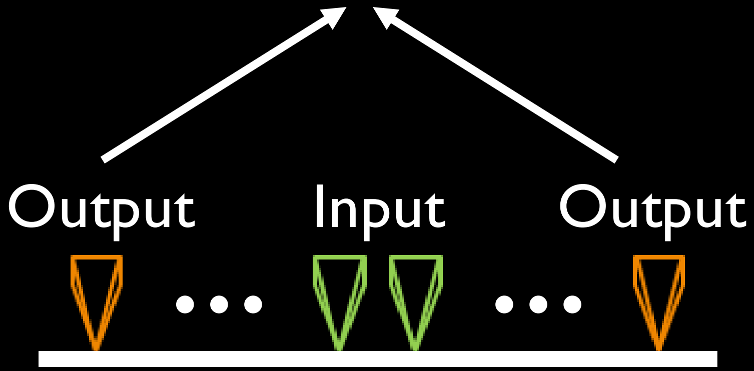
Problem Statement



Challenges

Extrapolation

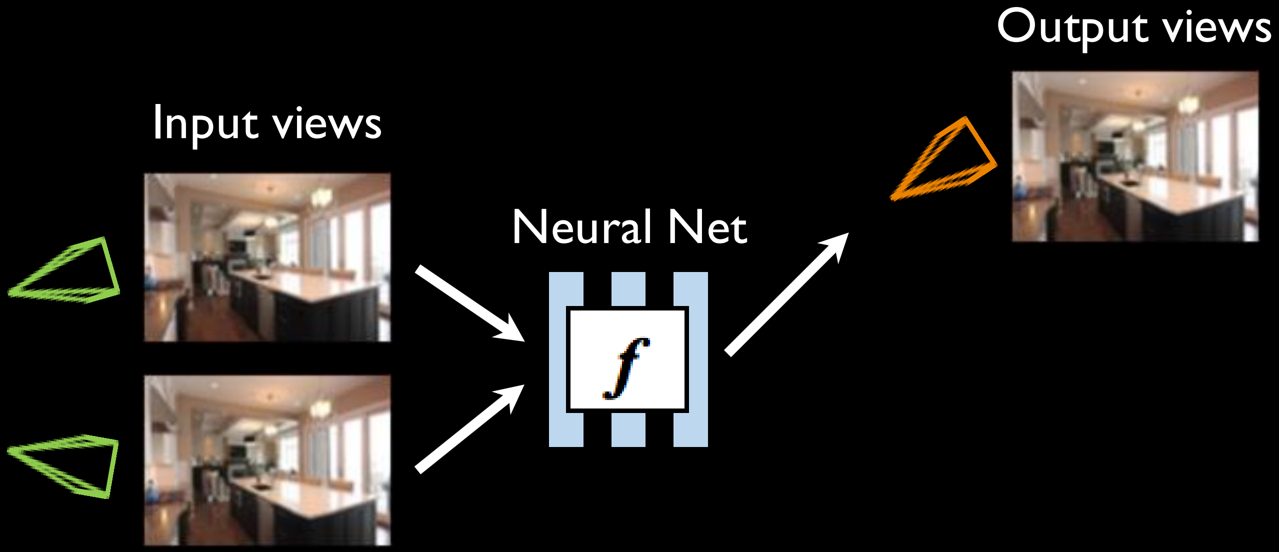
Large disocclusion



Non-Lambertian Effects

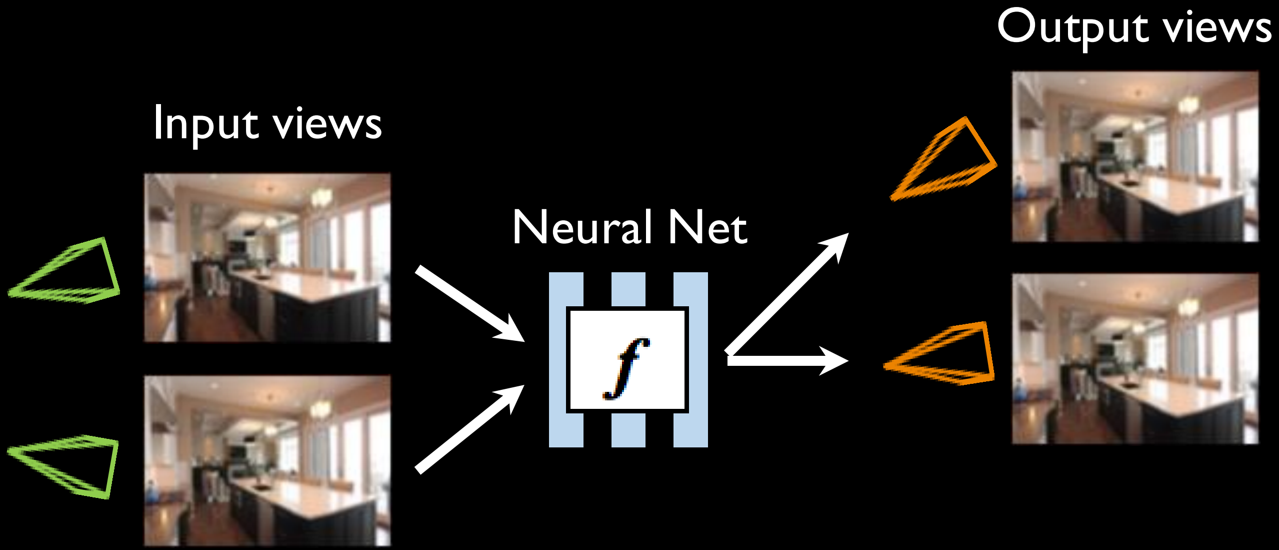
Reflections, transparencies, etc.

Prior Methods: No Shared Scene Representation



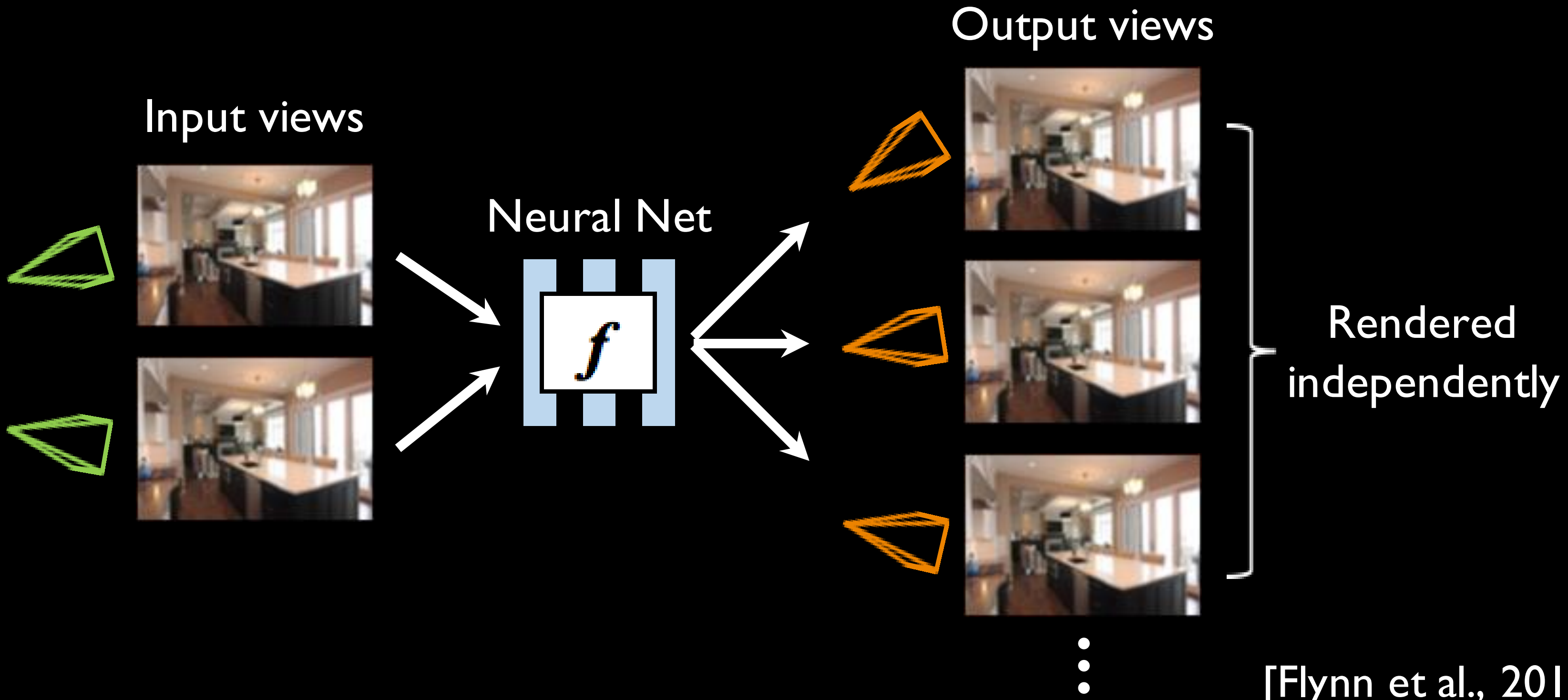
[Flynn et al., 2015]
[Kalantari et al. 2016]

Prior Methods: No Shared Scene Representation



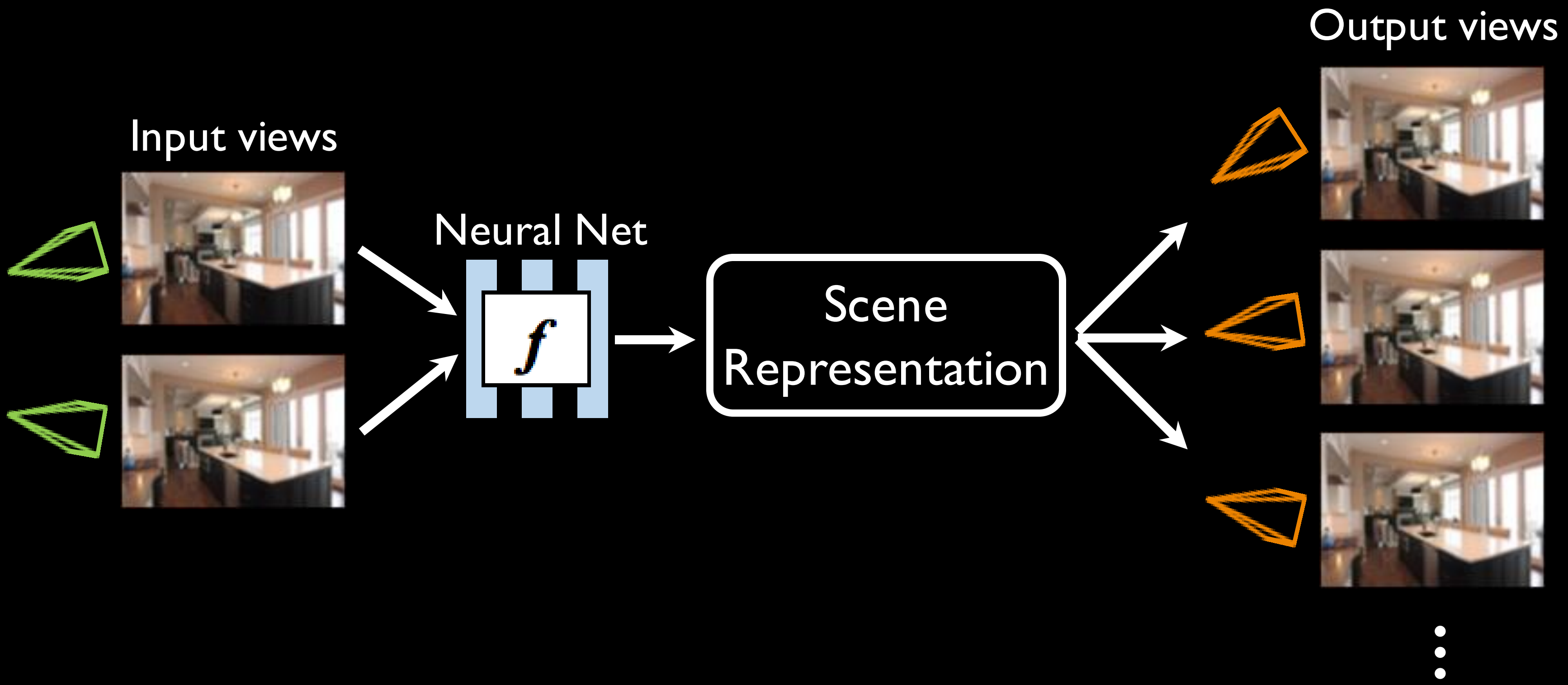
[Flynn et al., 2015]
[Kalantari et al. 2016]

Prior Methods: No Shared Scene Representation



[Flynn et al., 2015]
[Kalantari et al. 2016]

Ours: Shared Scene Representation



Stereo Magnification: Learning View Synthesis using Multiplane Images

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, Noah Snavely

SIGGRAPH 2018

Multiplane Camera (1937)

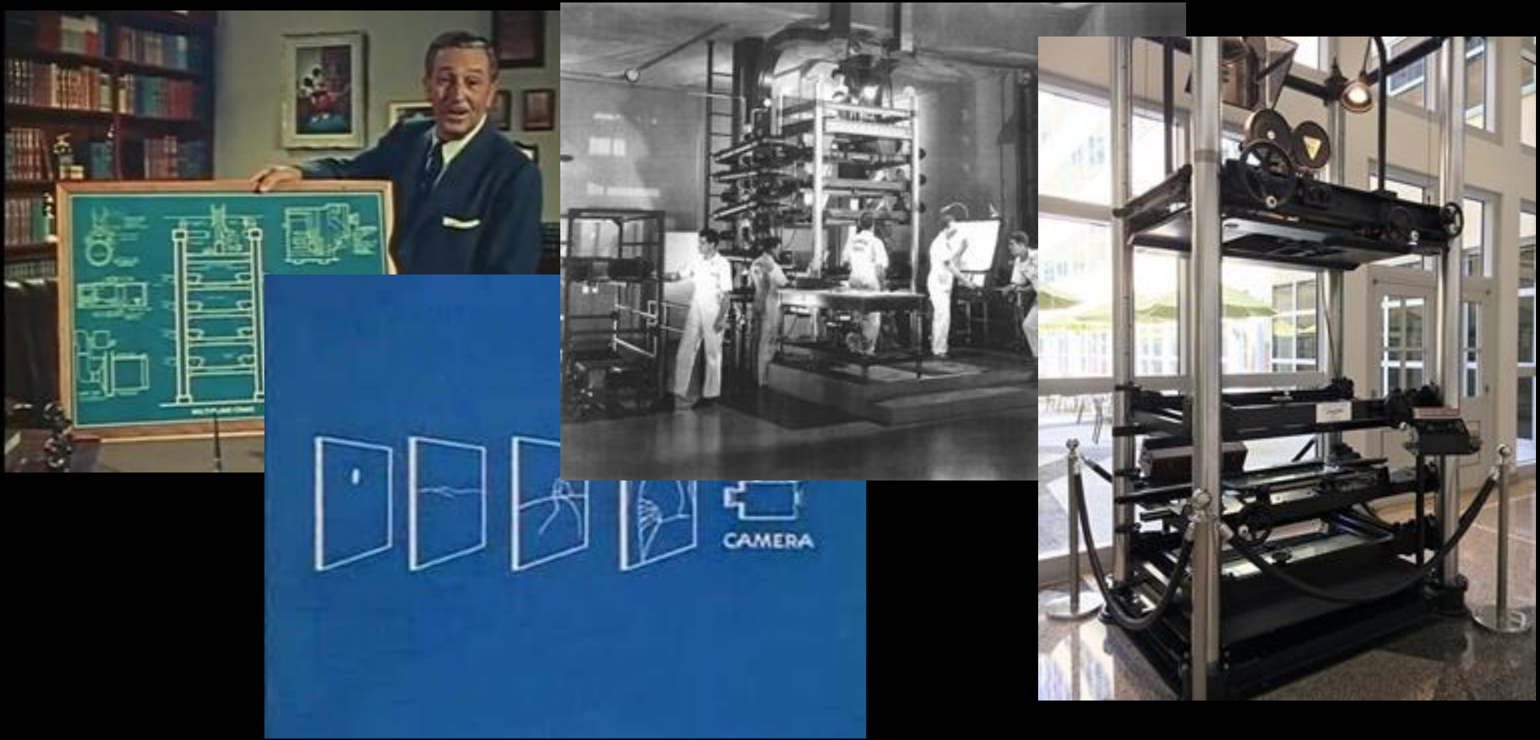
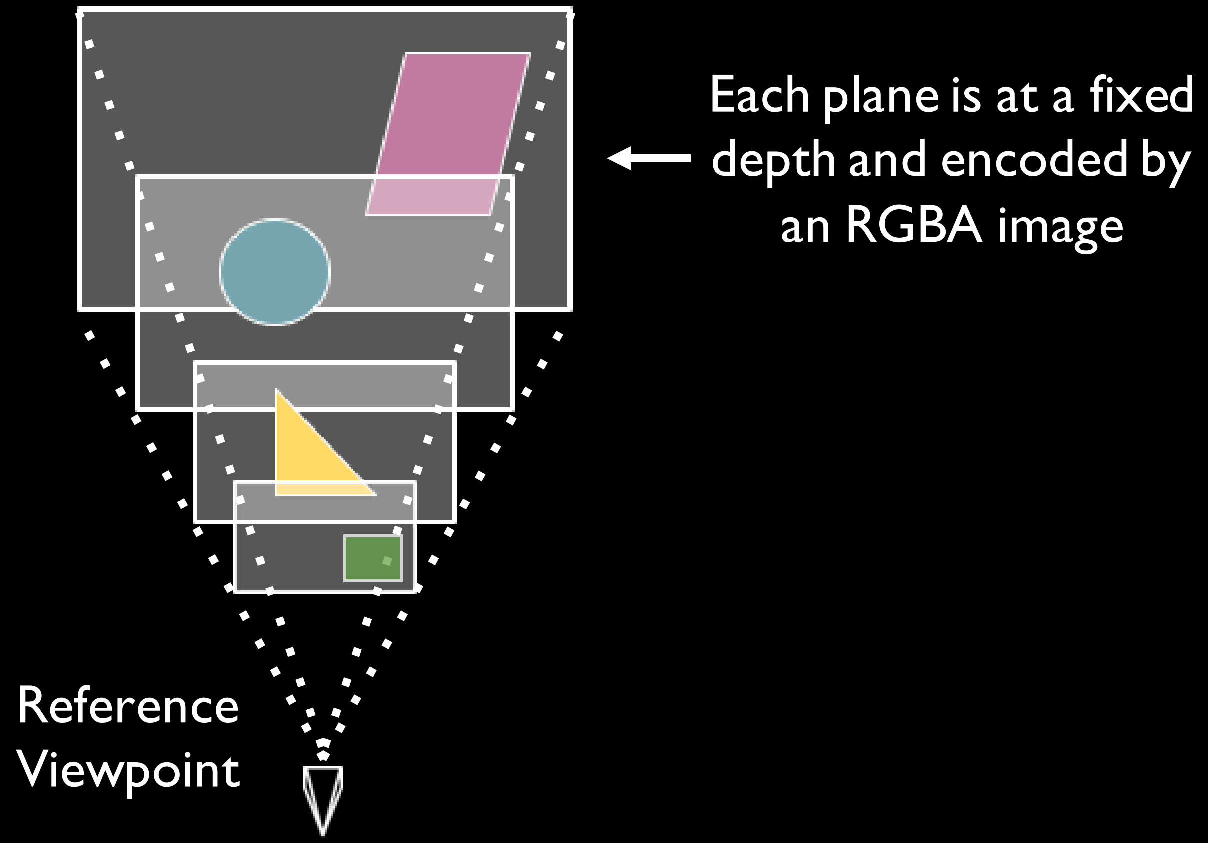


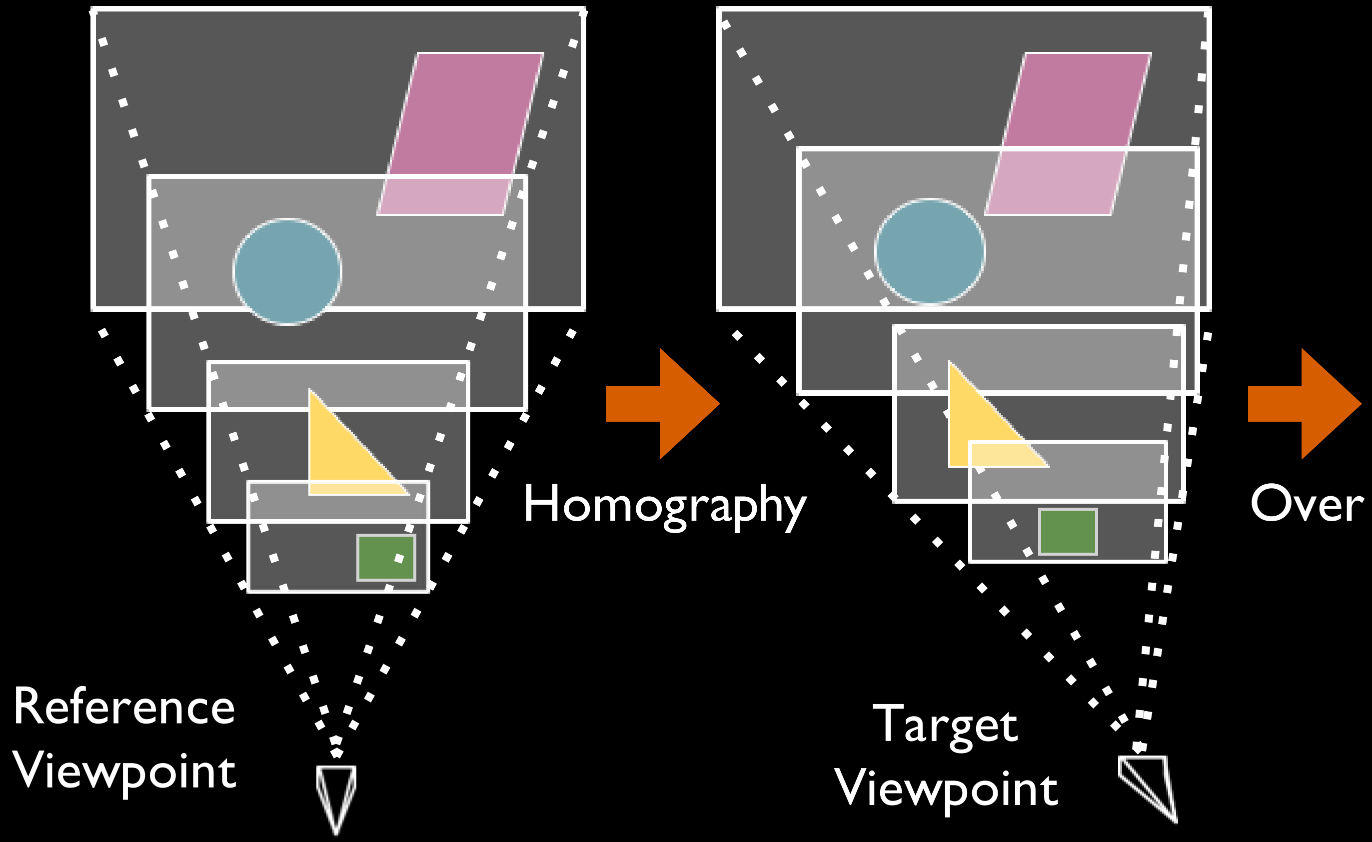
Image credits: Disney

https://www.youtube.com/watch?v=kN-eCBAOw60 (from 1957)

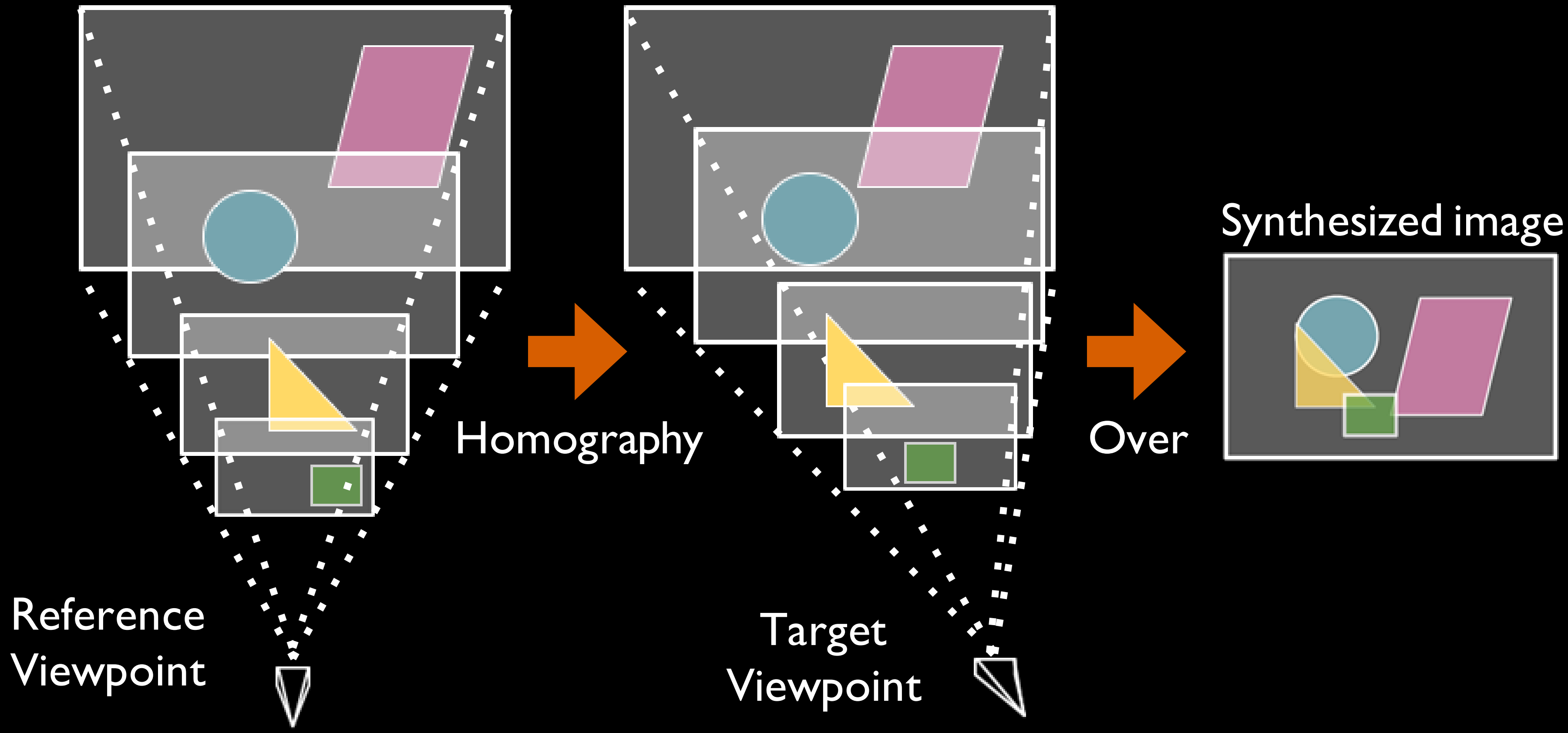
Multiplane Images (MPIs)

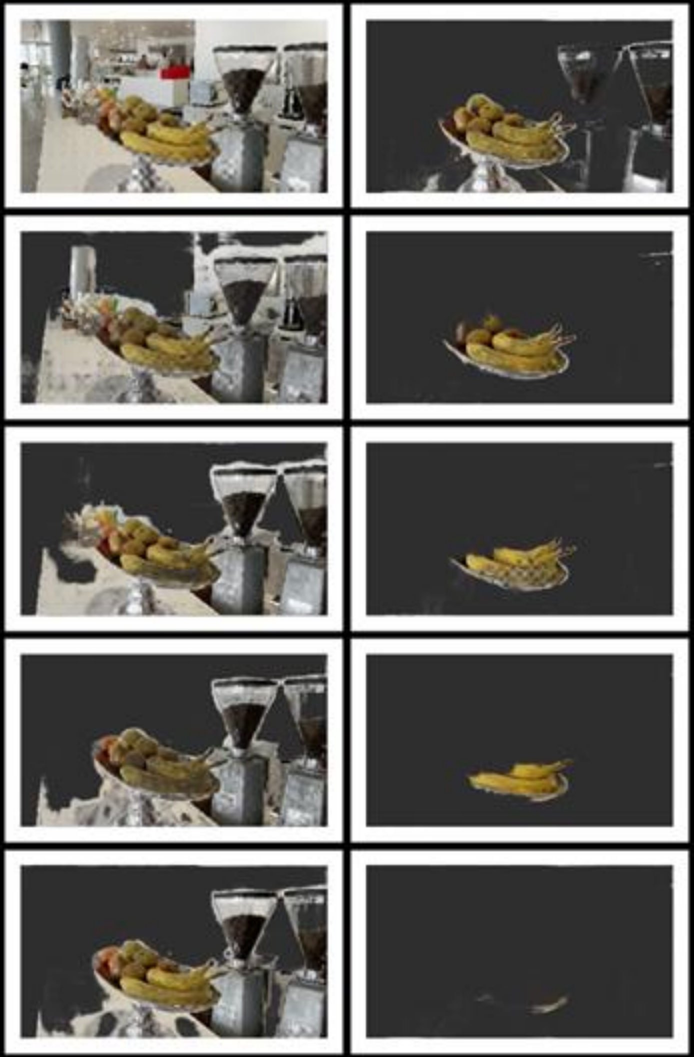


View Synthesis using Multiplane Images

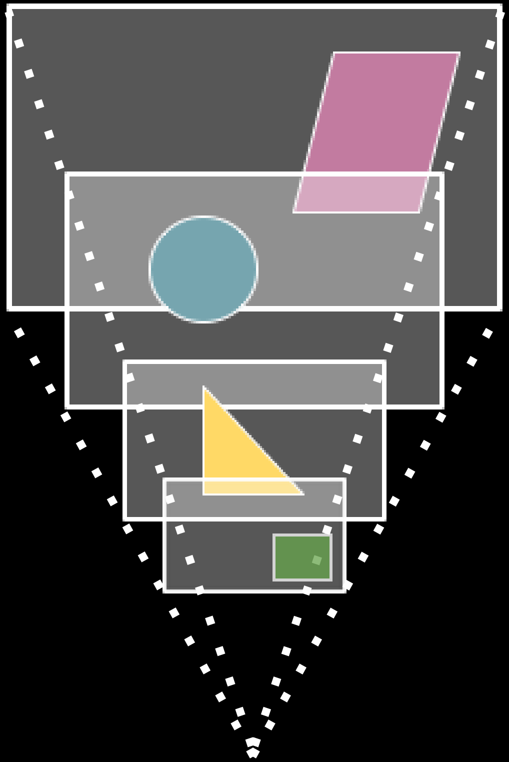


View Synthesis using Multiplane Images



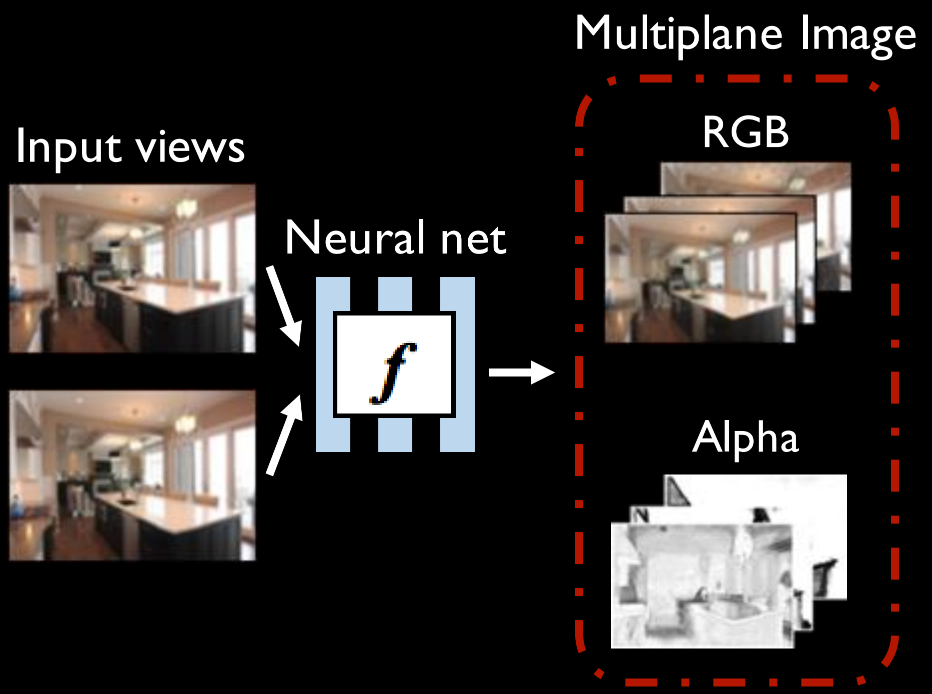


Properties of Multiplane Images

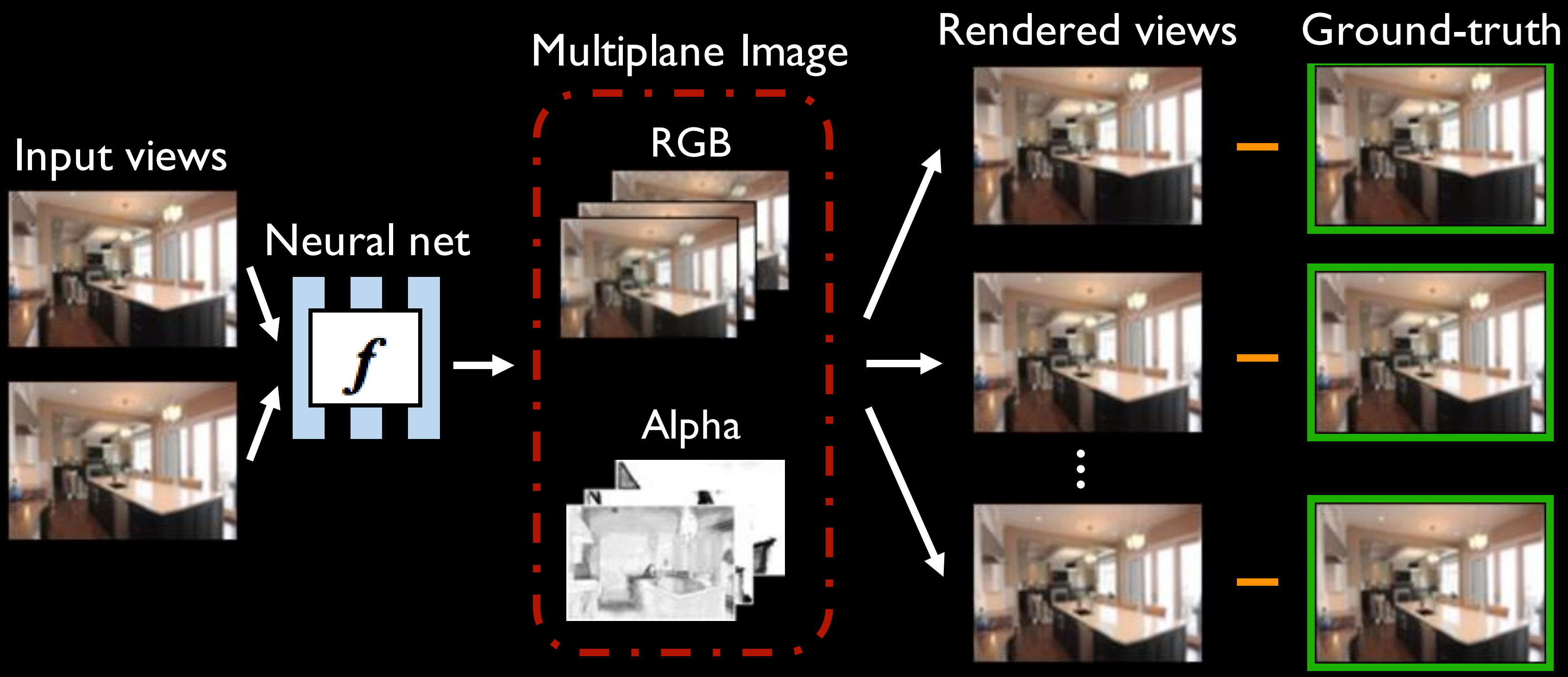


- Models disocclusion
- Models soft edges and non-Lambertian effects
- Efficient for view synthesis
- Differentiable rendering

Learning Multiplane Images

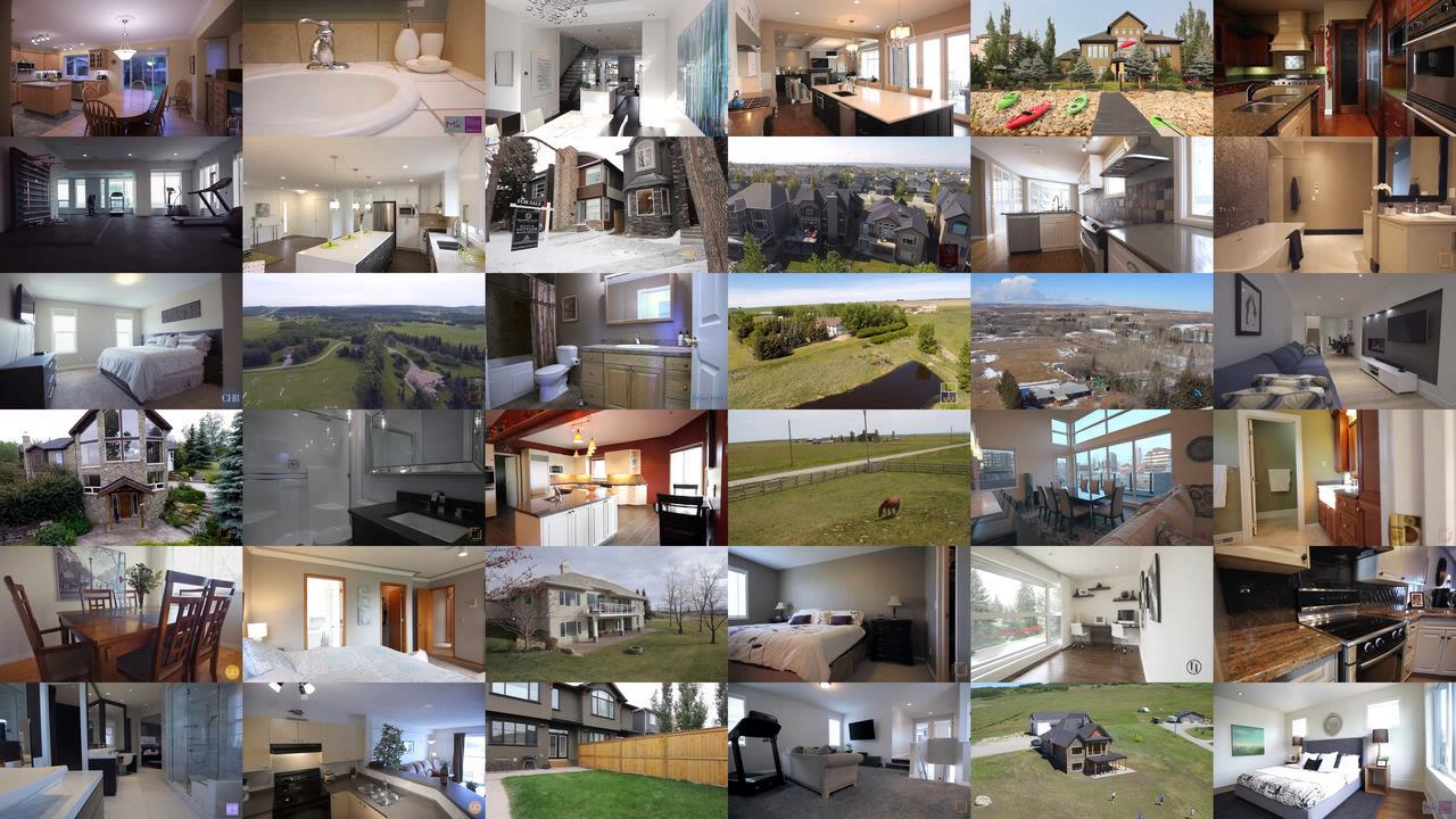


Learning Multiplane Images

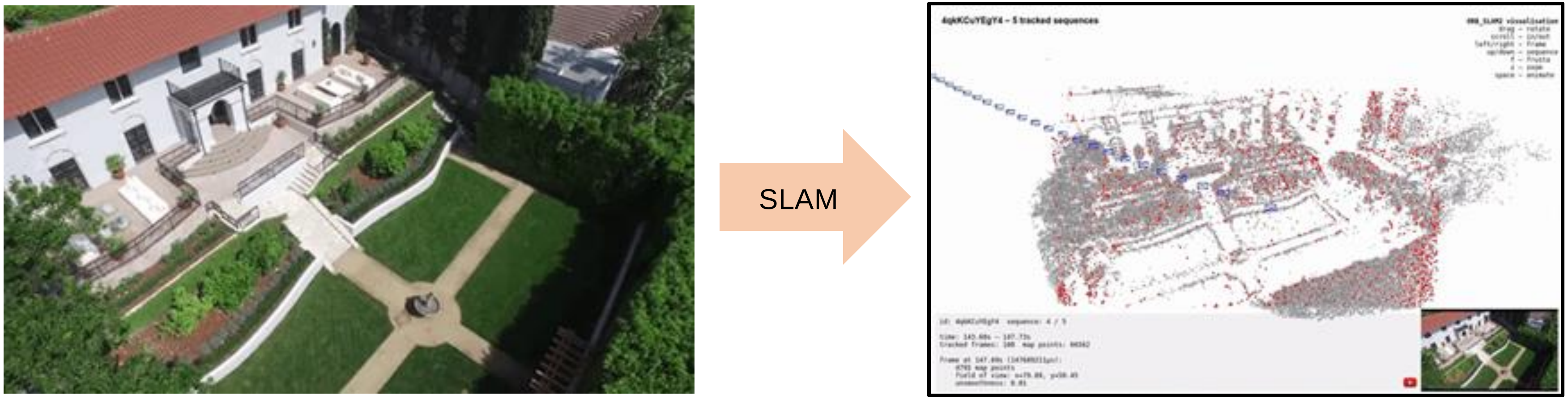


Training Data

Need massive set of triplets with known camera poses


FOR SALE

RealEstate10K

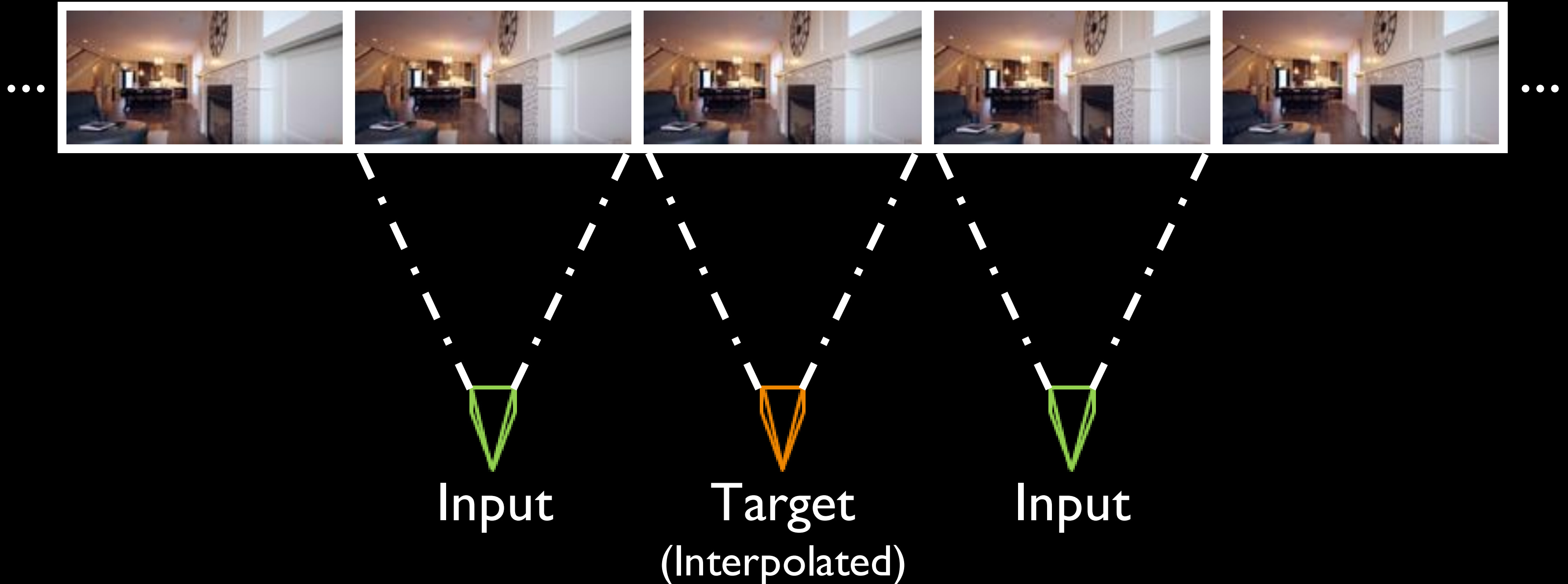


10 million frames from **80,000 video clips** from **10,000 videos**

https://google.github.io/realestate10k/

Sampling Training Examples

Sampling Training Examples



Results


Left

Right

Output

Image 1

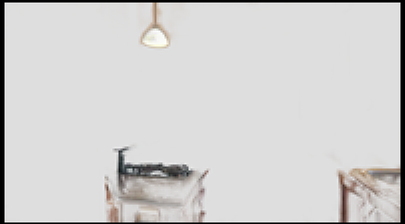
Image 2


Output

Reference input view

Plane 0

Plane 9



Plane 13

Plane 16

Plane 24

Plane 26

Extrapolating Cellphone Footage

1.4 cm

→

6.3 cm


Right input image

Learning 3D geometry: Key Ingredients

- Use the right representation (*e.g., Multi-plane Images*)
- Train on lots of data (*e.g., Internet videos*)
- Train using a widely available source of supervision — *other video frames*
  - This idea of **multi-view supervision** has been very active in 3D vision for the past few years
  - Predict from one frame, test by projecting into another and computing a **reprojection loss**

Limitation: Dynamic Scenes

- So far, our training data assumes rigid scenes
- Otherwise, SfM / SLAM will fail, as will reprojection loss
- But most scenes have moving and non-rigid objects

Learning Depths of Moving People by Watching Frozen People

Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, Bill Freeman

CVPR 2019 (to appear)

https://www.youtube.com/watch?v=fj_fK74y5_0

Takeaways

- Harness the power of multi-*view supervision* for 3D learning
- The Internet is an amazing source of training data full of surprising images and videos
- Representations are important! Layers are one nice approach, but the best representation is elusive
    - Should be expressive, efficient, good for learning, etc...

Future directions

- Train on much more varied (noisier) data (all of YouTube?)
- Much larger view extrapolations (requires better inpainting in disoccluded regions)
- Predicting richer representations from a single view
  - Towards full **inverse graphics:** image to shape, materials, and geometry

Thank you!


Richard Tucker


Zhengqi Li


Tinghui Zhou


John Flynn


Graham Fyffe


Shubham Tulsiani


David Lowe


Matt Brown

Questions?