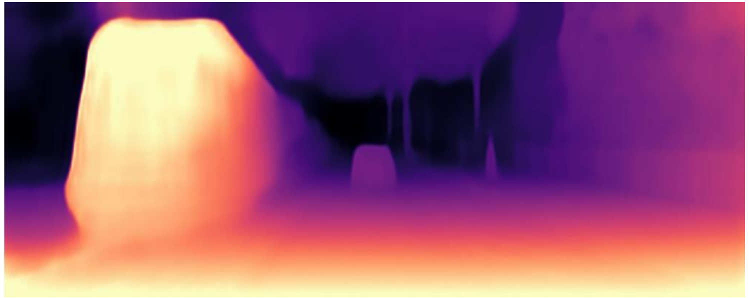
CS5670: Computer Vision

Learning 3D Geometry

RGB Image

Deep learning



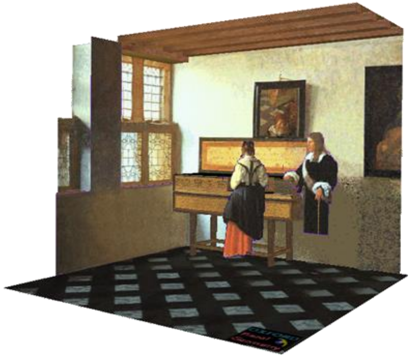
Depth map

Announcements

- Please give us feedback! Fill out course evaluations here (for bonus points!):
  - https://apps.engineering.cornell.edu/CourseEval/
- Project 5 due Friday at 11:59pm
- Take-home final exam to be released May 11
- Monday: course wrap up (last lecture of class)

Single-view modeling

Vermeer's *Music Lesson*

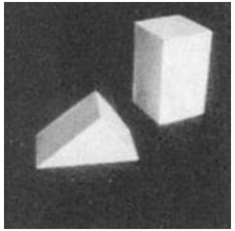


Reconstructions by Criminisi et al.

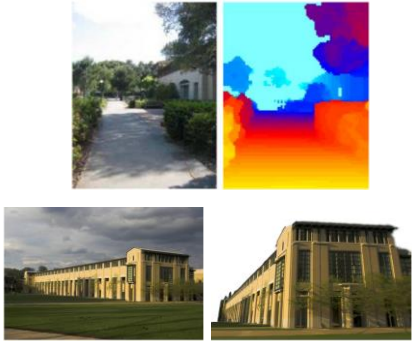
Can we use deep learning to predict geometry from a single image?

Astonishing recent progress in learning 3D perception

“Blocks world”
Larry Roberts
(1963)

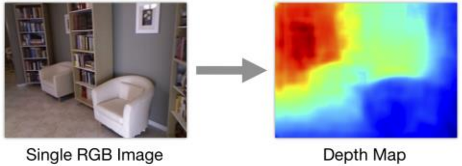


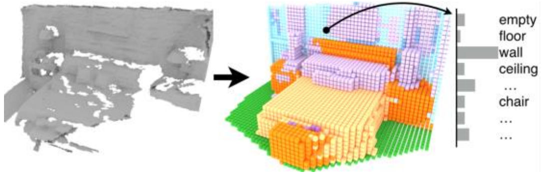
Pre-deep era
(2005)



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

Supervised deep learning
(2014)



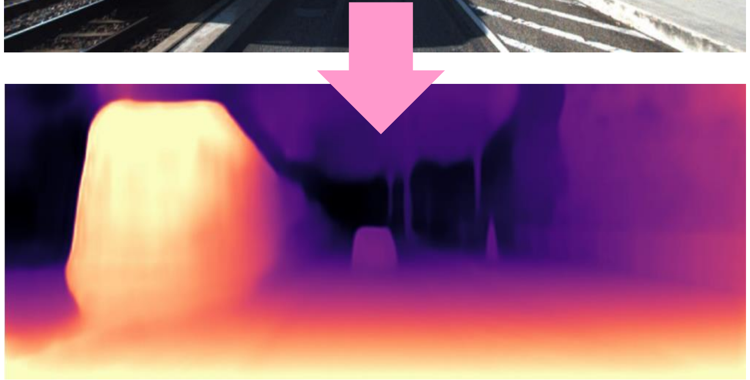


[Eigen, Puhrsch, Fergus, NIPS 2014]
[Song et al, CVPR 2017]
...
go/im2depth

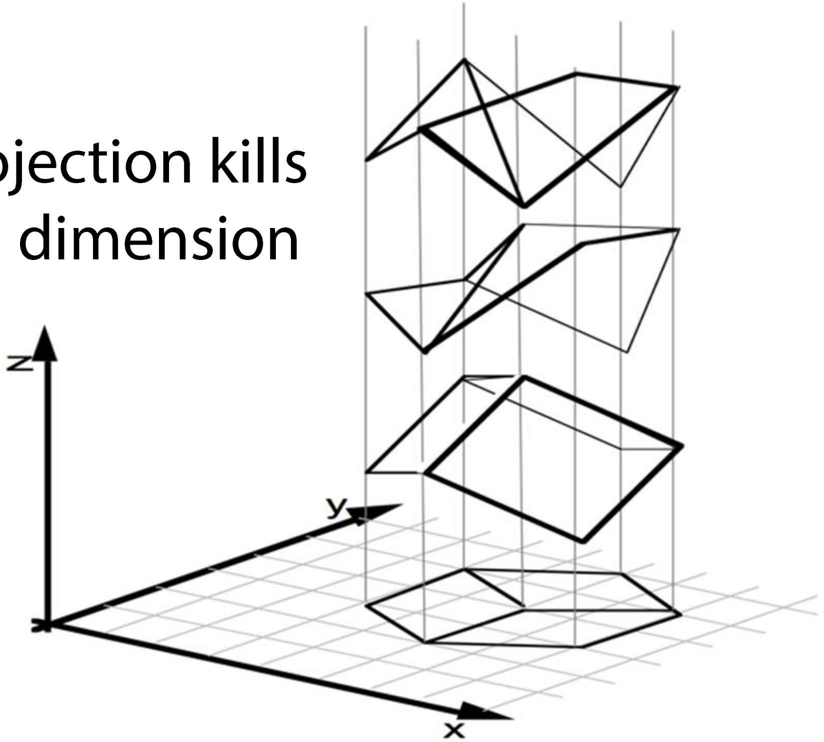
Multi-view supervision
(2016)

[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 RGB view to depth



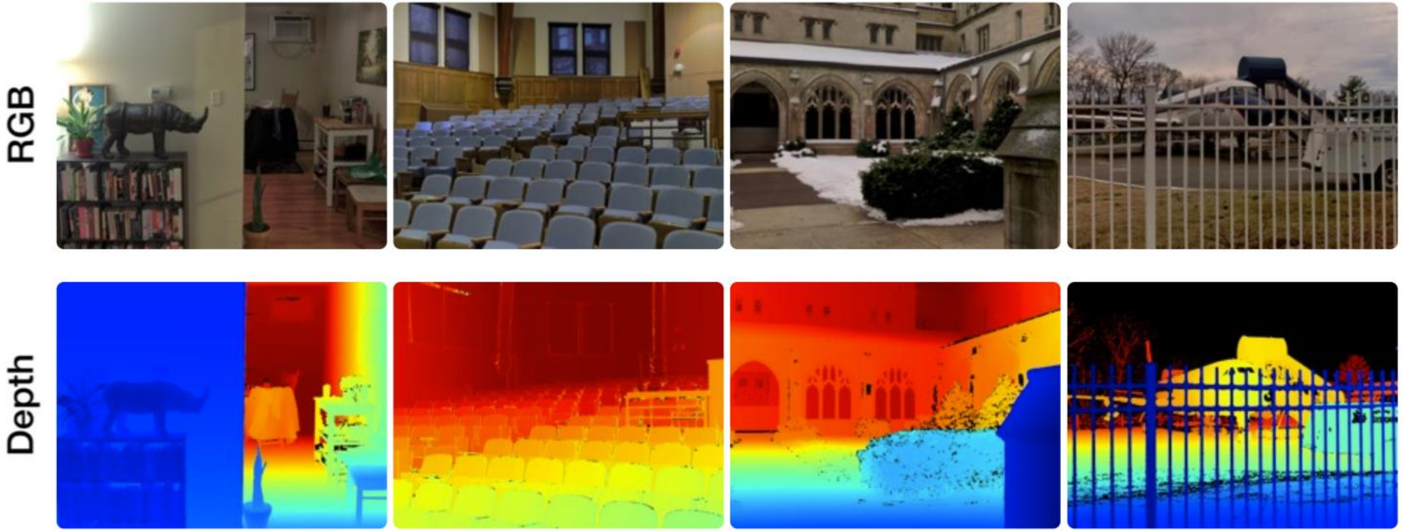
but projection kills
the 3rd dimension



[Sinha & Adelson, 1993]

Learning single-view depth prediction

- To apply deep learning to this problem we need lots of training data in the form of RGB images and corresponding depth maps

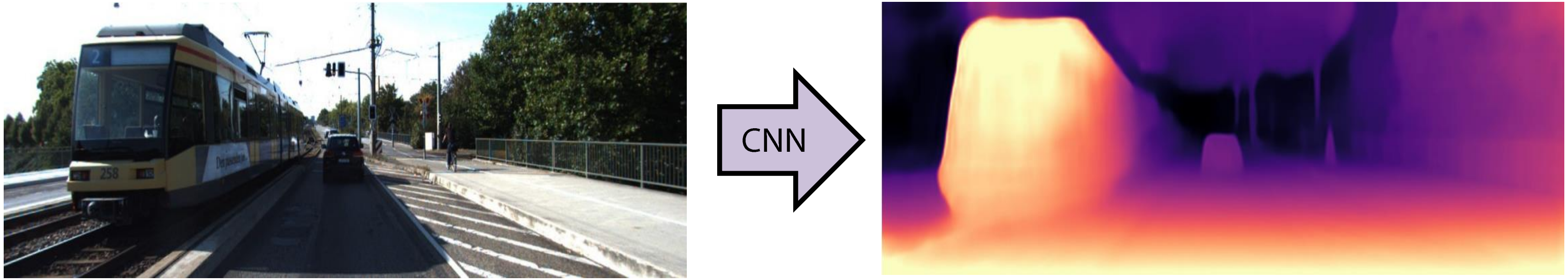


Source: https://diode-dataset.org/

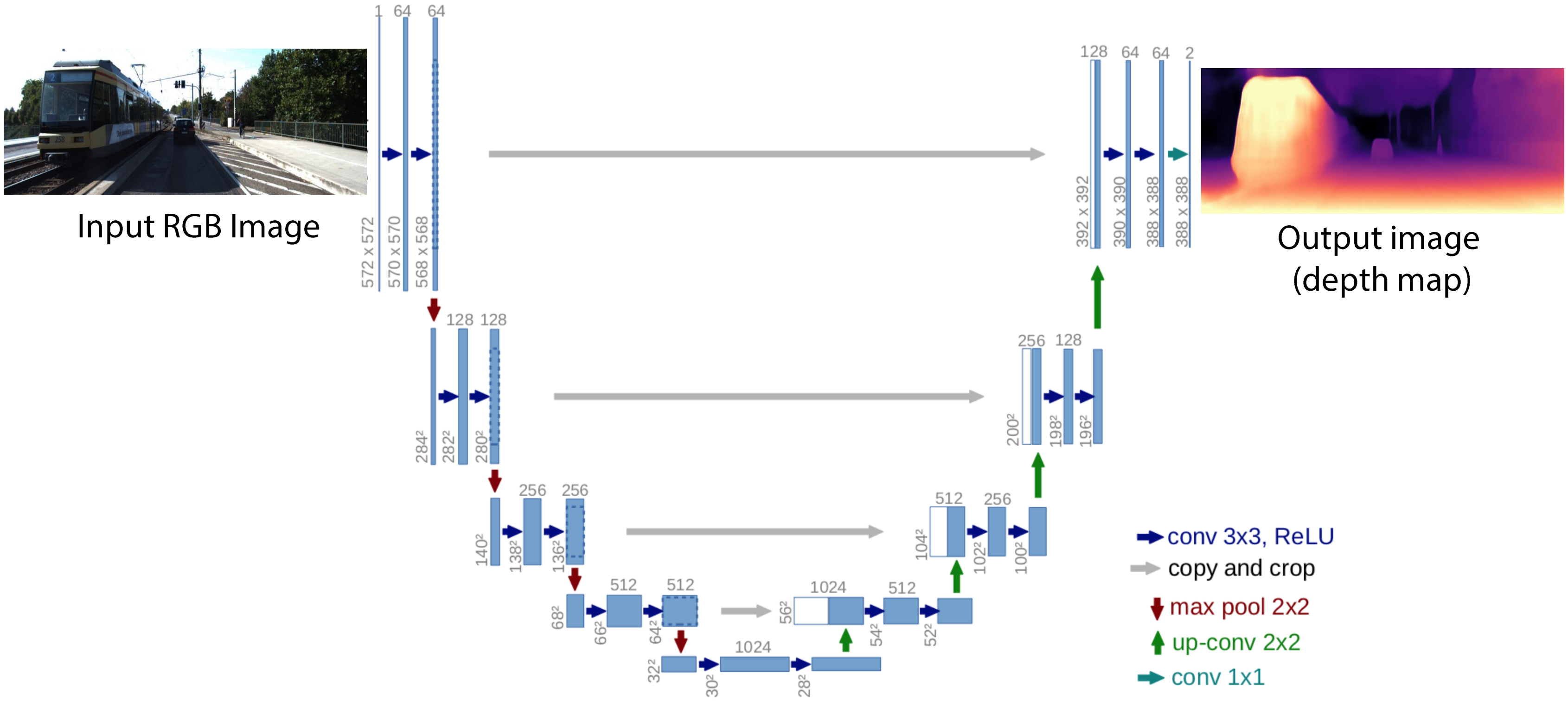
CNN architectures for single-view depth

- Need an architecture that takes in an image (an RGB image) and produces another image (a depth map)
- Similar to other problems where images are the outputs (e.g., semantic segmentation, colorization, object boundary detection)
- In contrast to image classification, where outputs are probabilities for a set of object categories (e.g., vector of length 1000)

CNN architectures for single-view depth

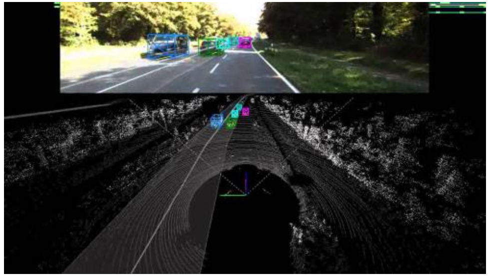


Common choice: UNet architecture



How to get training data?

LIDAR



KITTI [Geiger et al. 2012]

Kinect

NYU [Eigen et al. 2014]

Manual annotation

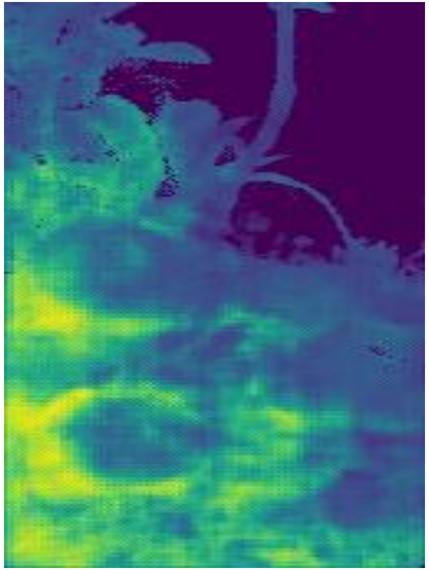
Depth in the Wild [Chen et al. 2016]

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

Problem: generalizing beyond training data

- If you train on images of streets scenes from KITTI, you won't get good results on test images like this:

Input RGB image



Predicted depth map from KITTI-trained model

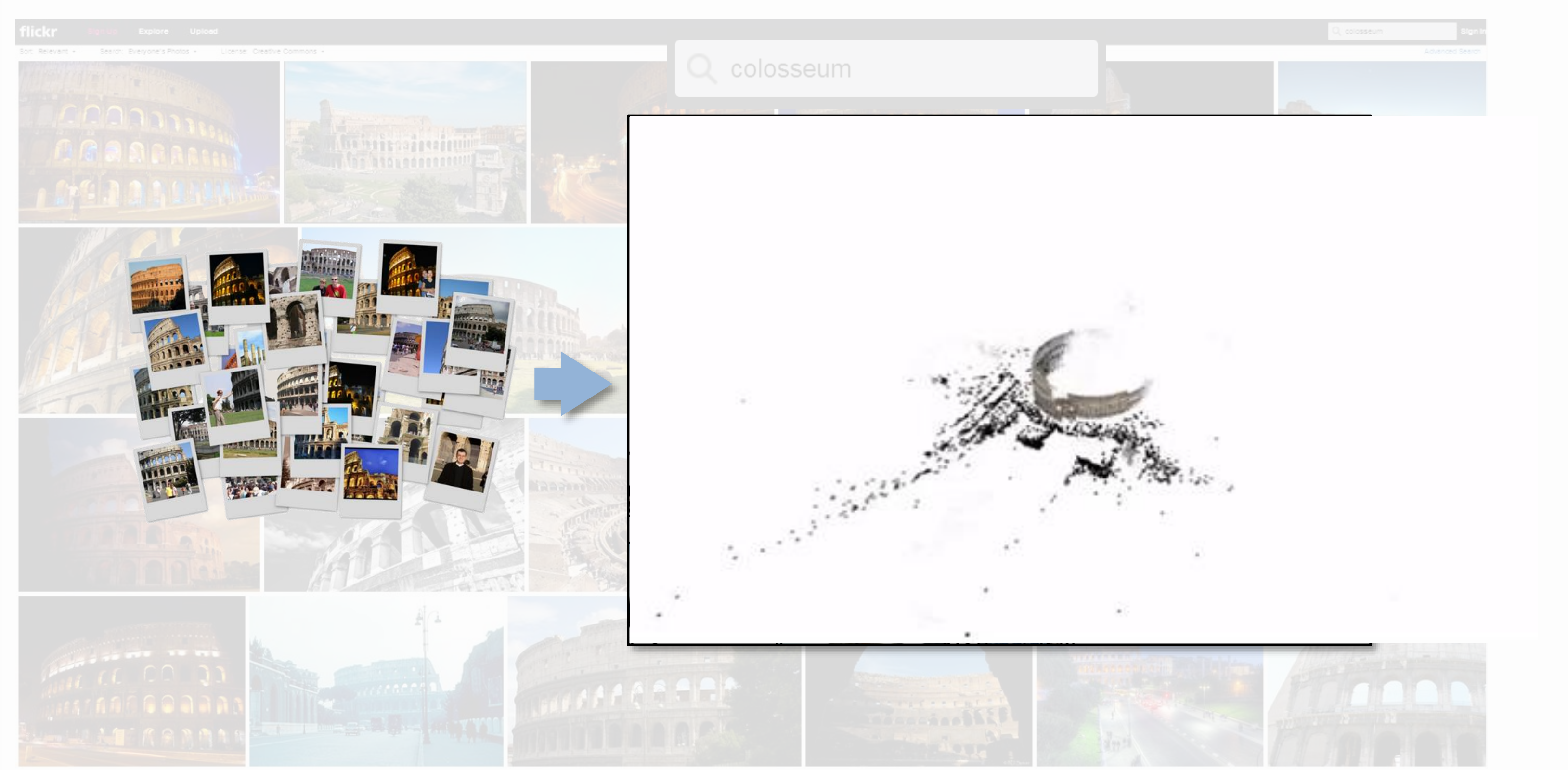
How can we gather more diverse data?

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

Training: Multiple views

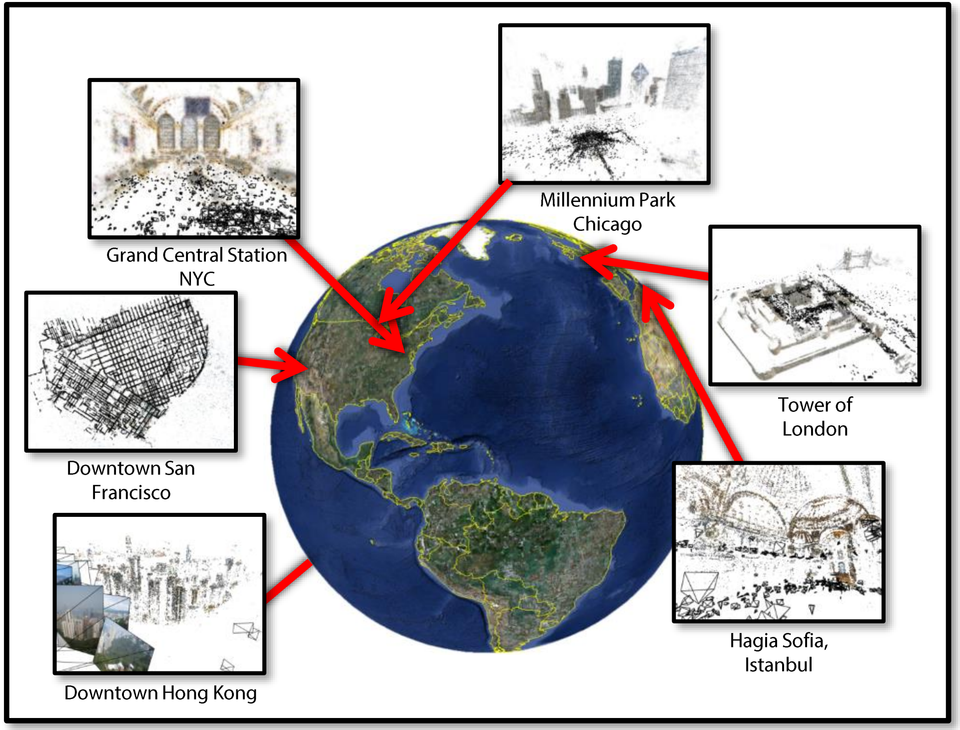
Testing: Single Image

Another source of training data: Structure from Motion reconstructions



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

Reconstructing the World's Landmarks



[Li, Snavely, Huttenlocher, Fua. ECCV 2012]

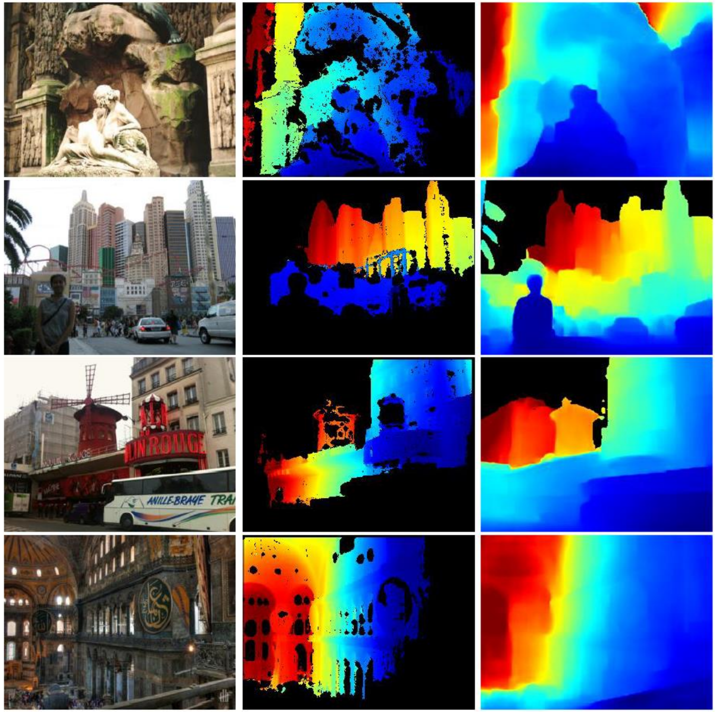
MegaDepth dataset

>130K (RGB, depth map) pairs

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

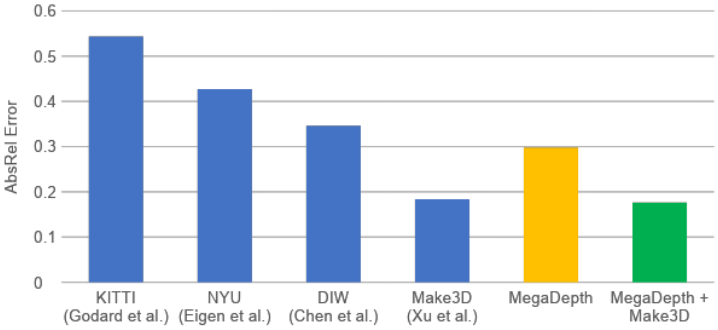
[Zhengqi Li and Noah Snavely. *MegaDepth: Learning Single-View Depth Prediction from Internet Photos*. CVPR 2018]

MegaDepth-trained prediction results

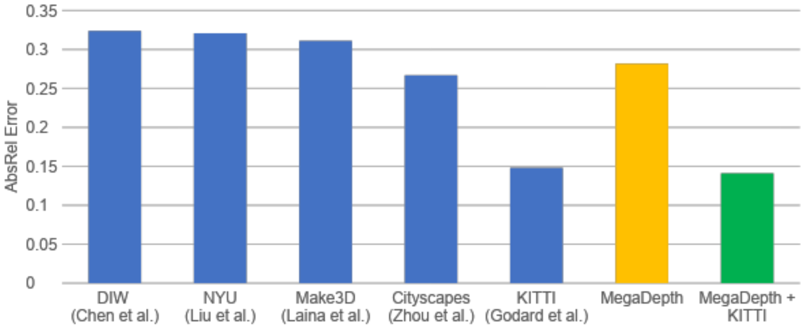


Input Ground truth Predicted depth

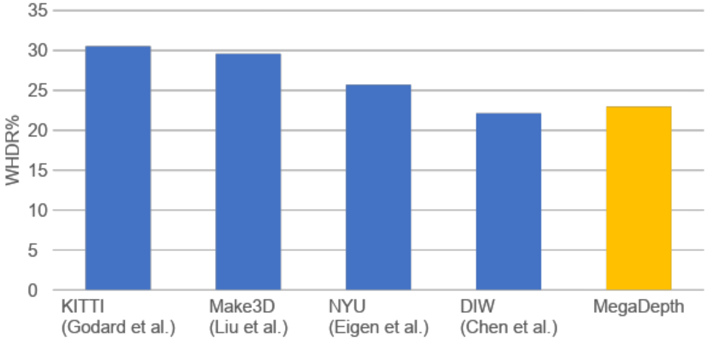
Internet data generalizes well



Train on X, test on Make3D

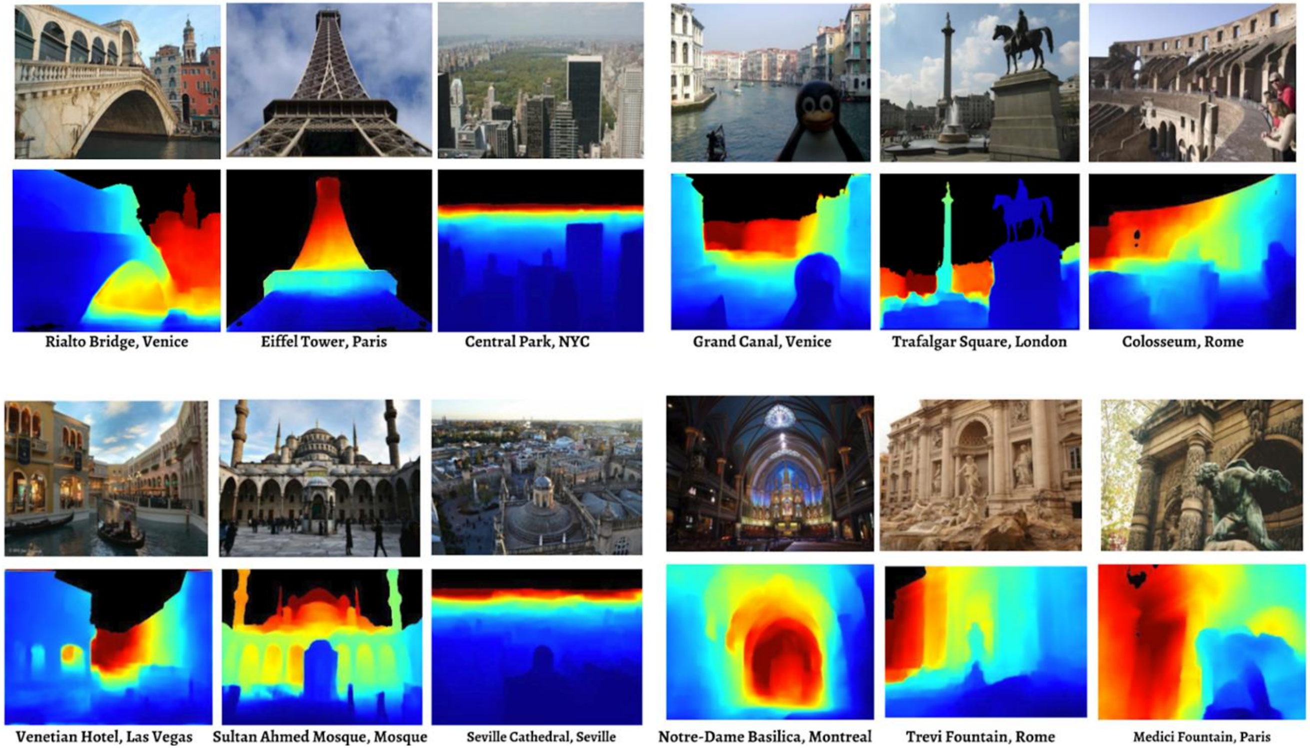


Train on X, test on KITTI



Train on X, test on DIW

More depth prediction results



Single-view depth from Megadepth model

Input RGB image

Predicted depth map

Questions?

A related task: view synthesis

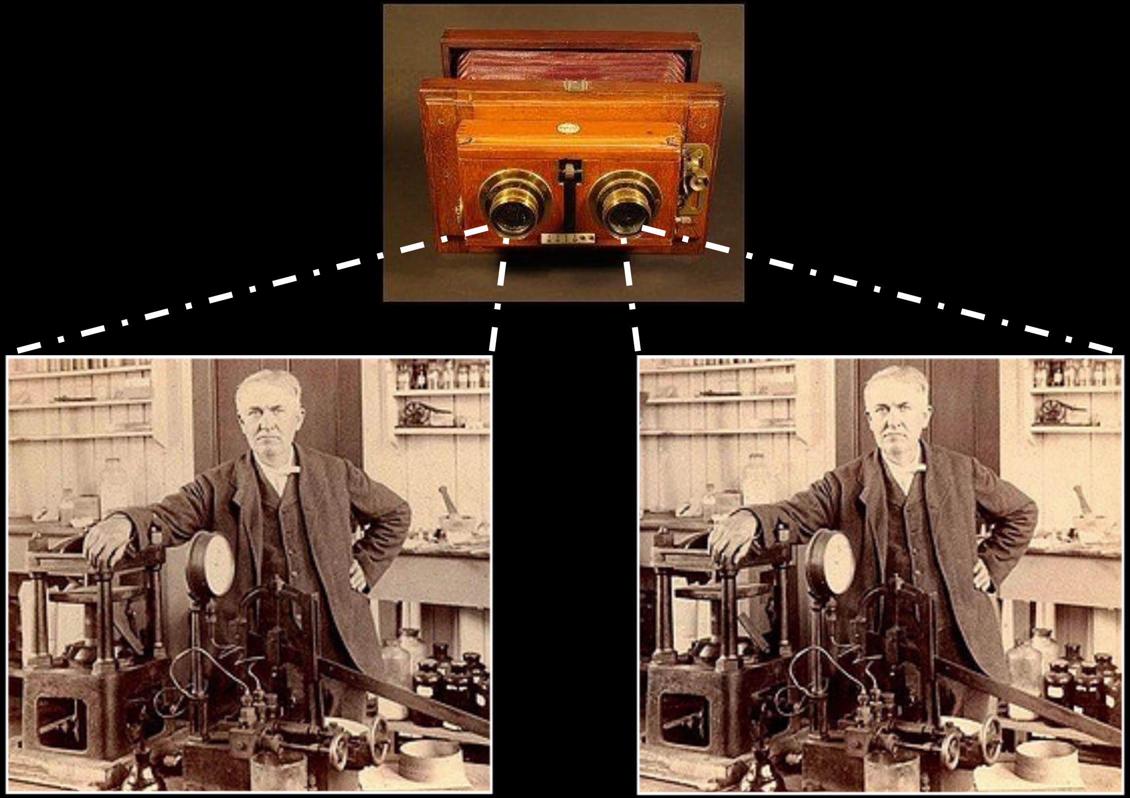
- So much for single-view depth
- Another thing we might want to do is *render new views of the captured scene* (i.e., view synthesis)
  - Related to light fields lecture from a few weeks back
- Involves more than just depth, but also filling in missing content behind the foreground

Cool recent work on view synthesis

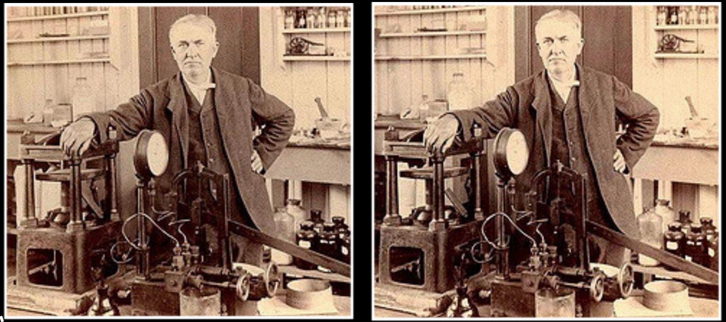
- Meng-Li Shih, Shih-Yang Su, Johannes Kopf, Jia-Bin Huang
  3D Photography using Context-aware Layered Depth Inpainting
- https://shihmengli.github.io/3D-Photo-Inpainting/

3D Photography using Context-aware Layered Depth Inpainting

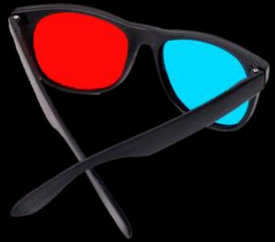
Stereo Photography



Stereo Photography



Viewing Devices



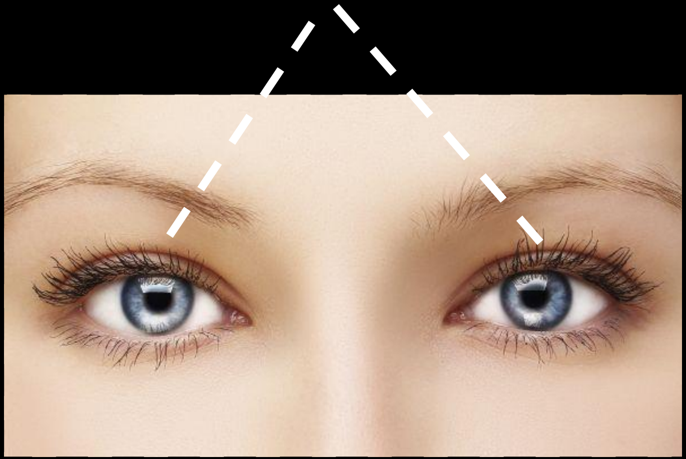
Stereo Photography

Queen Victoria at World Fair, 1851

Stereo Photography

Issue: Narrow Baseline

~6.5 cm



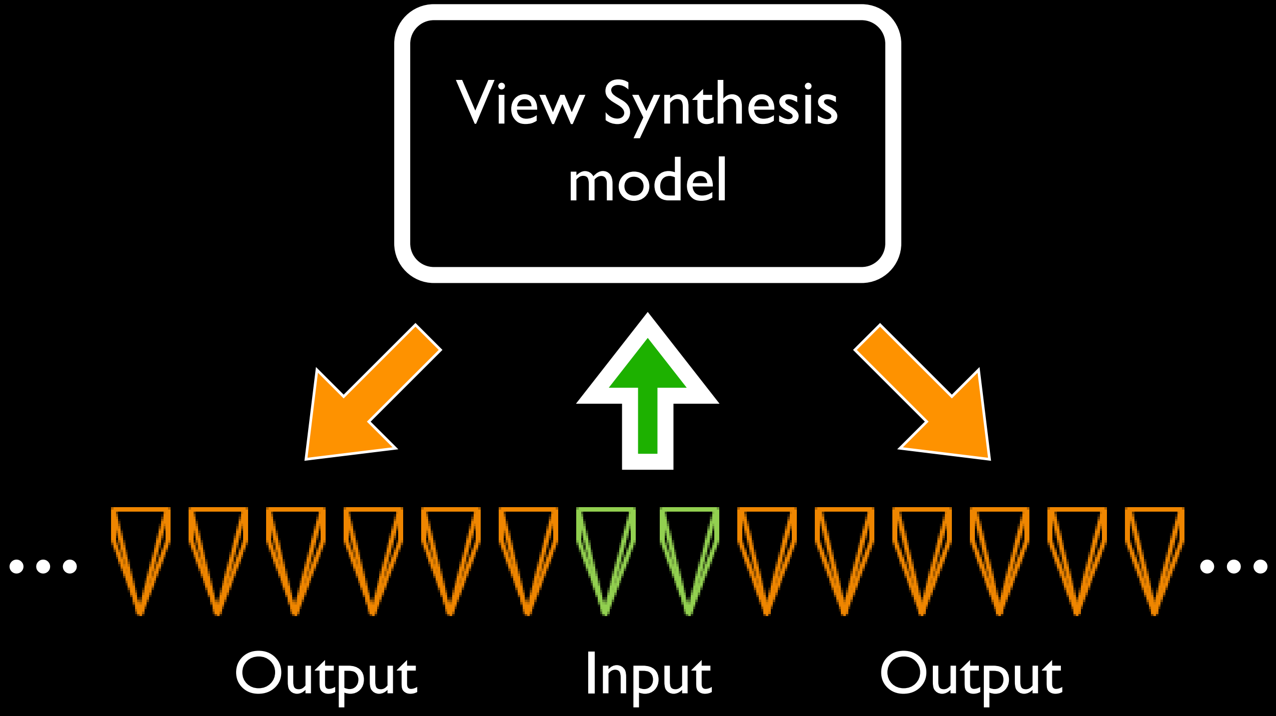
~1.5 cm


Left
MAZZER


Right
MAZZER


MAZZER

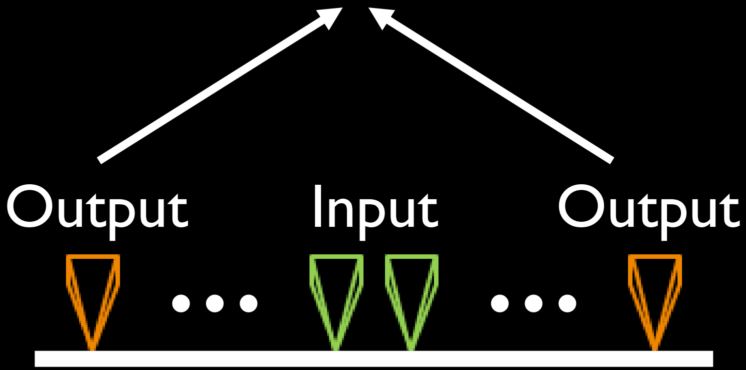
Problem Statement



Challenges

Extrapolation

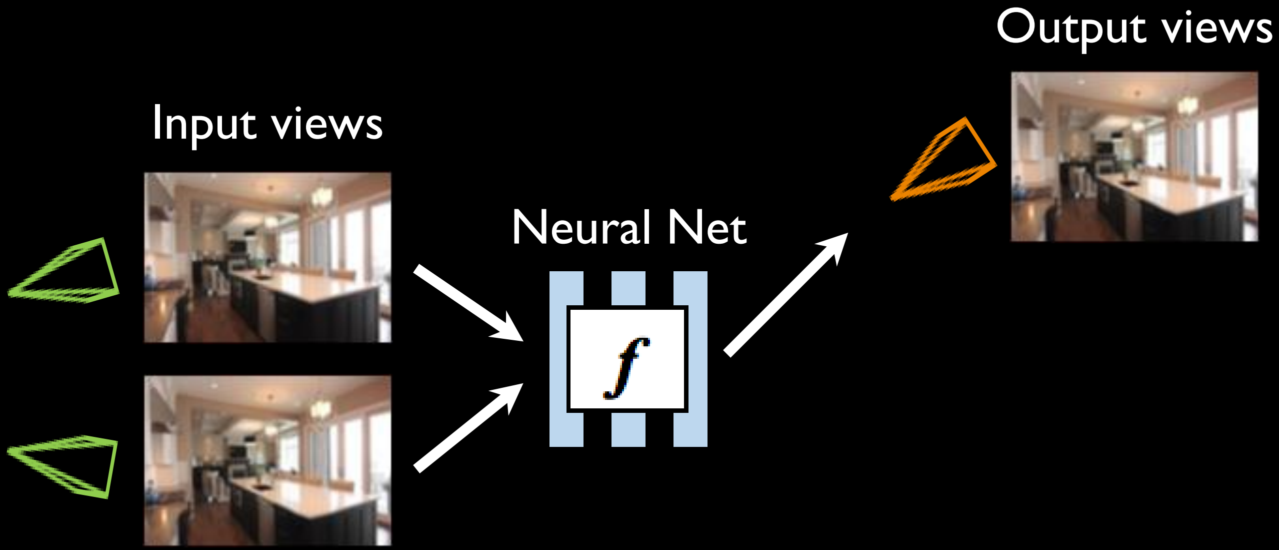
Large disocclusion



Non-Lambertian Effects

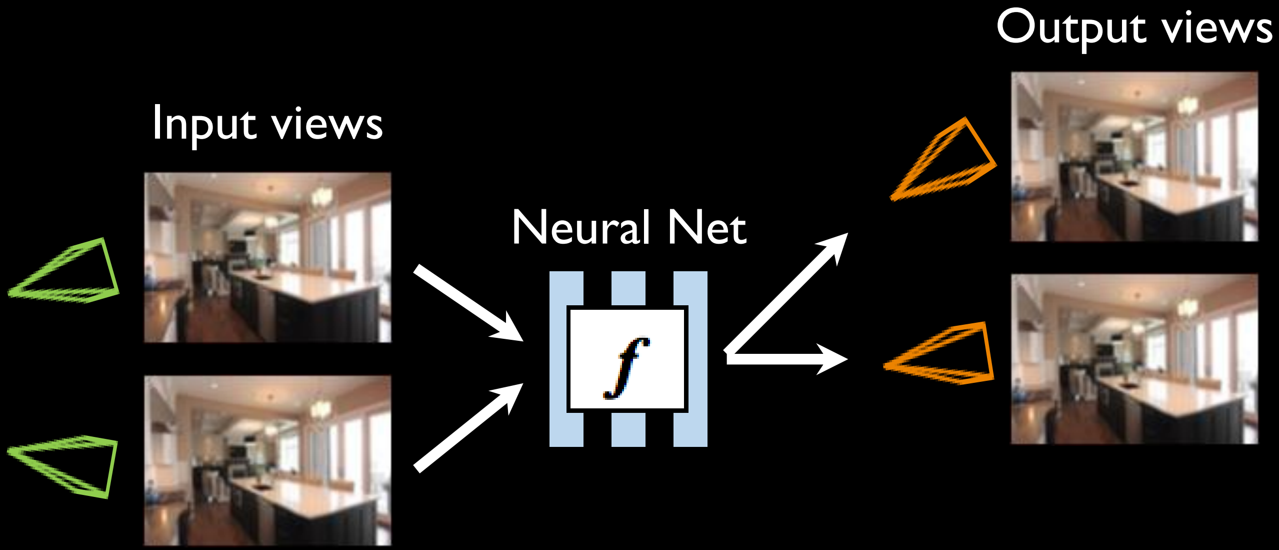
Reflections, transparencies, etc.

Prior Methods: No Shared Scene Representation



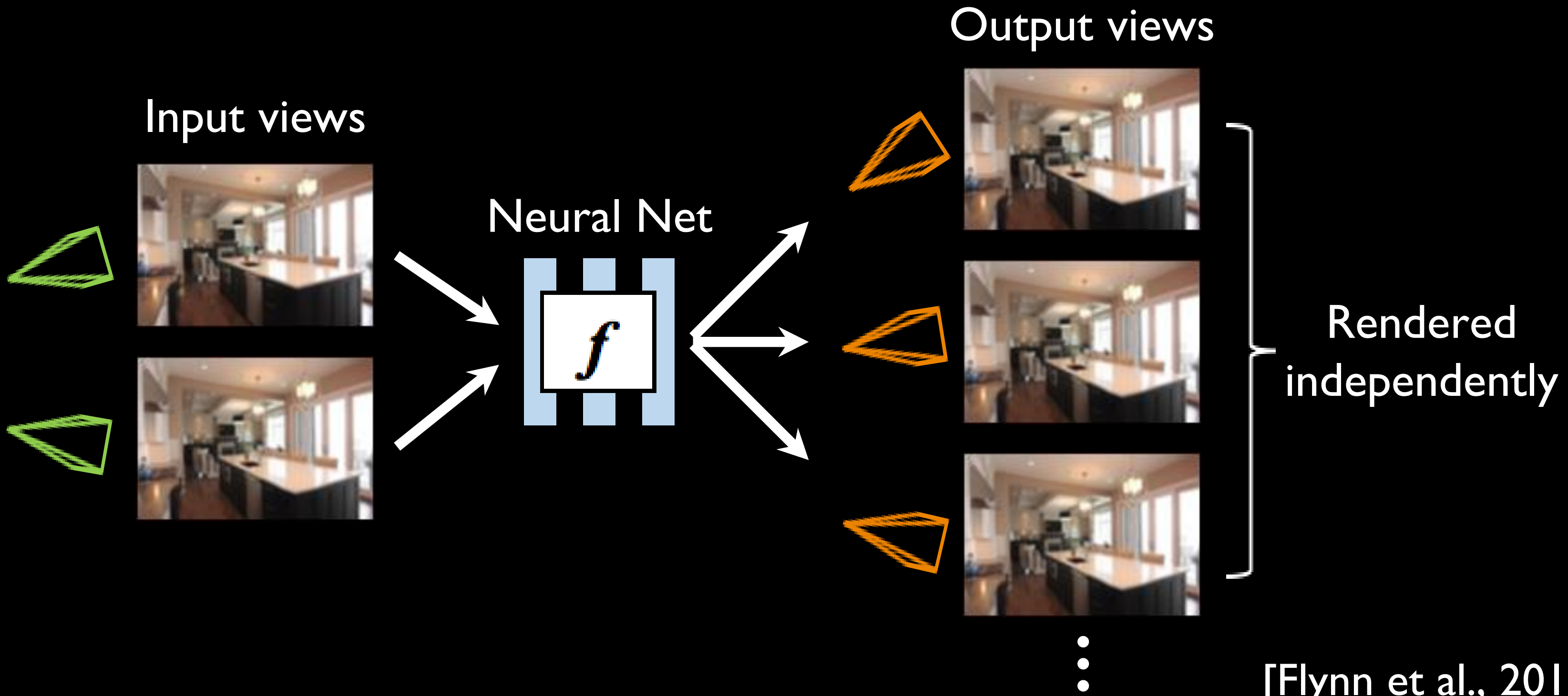
[Flynn et al., 2015]
[Kalantari et al. 2016]

Prior Methods: No Shared Scene Representation



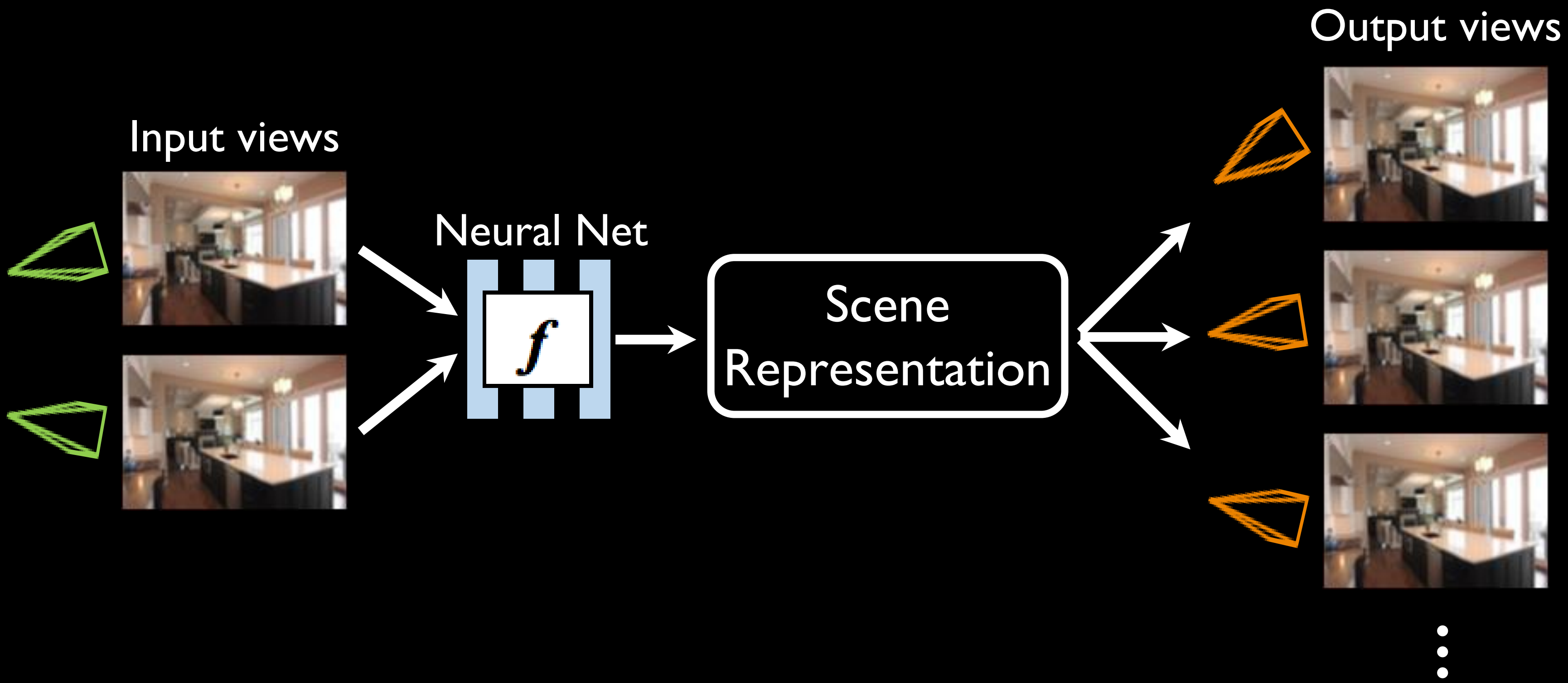
[Flynn et al., 2015]
[Kalantari et al. 2016]

Prior Methods: No Shared Scene Representation



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

Ours: Shared Scene Representation

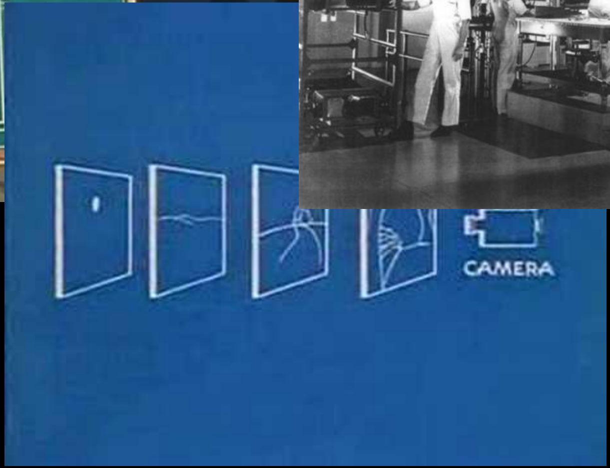


Stereo Magnification: Learning View Synthesis using Multiplane Images

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

SIGGRAPH 2018

Multiplane Camera (1937)



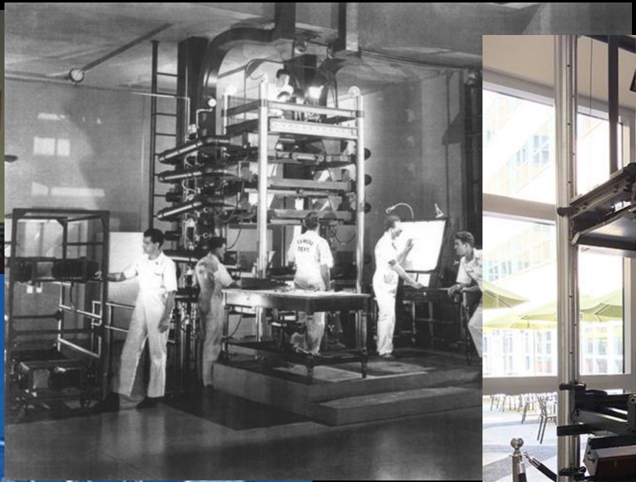
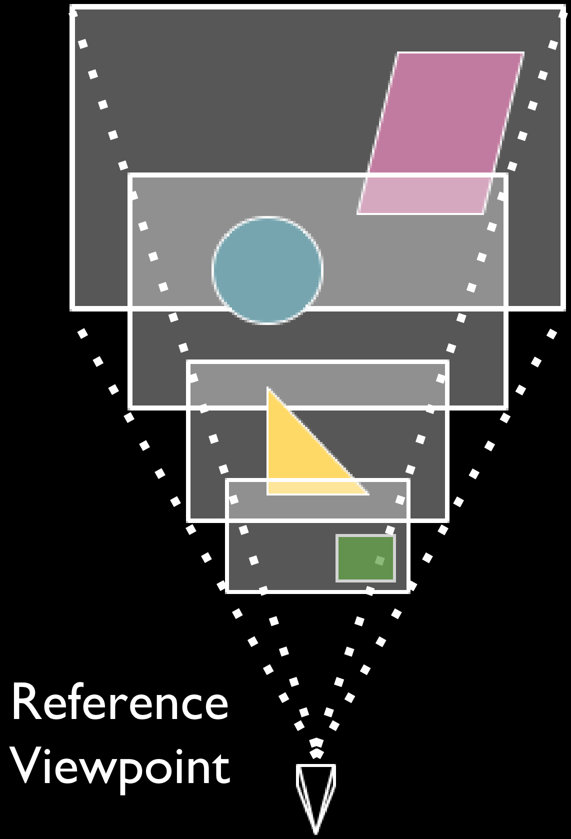


Image credits: Disney

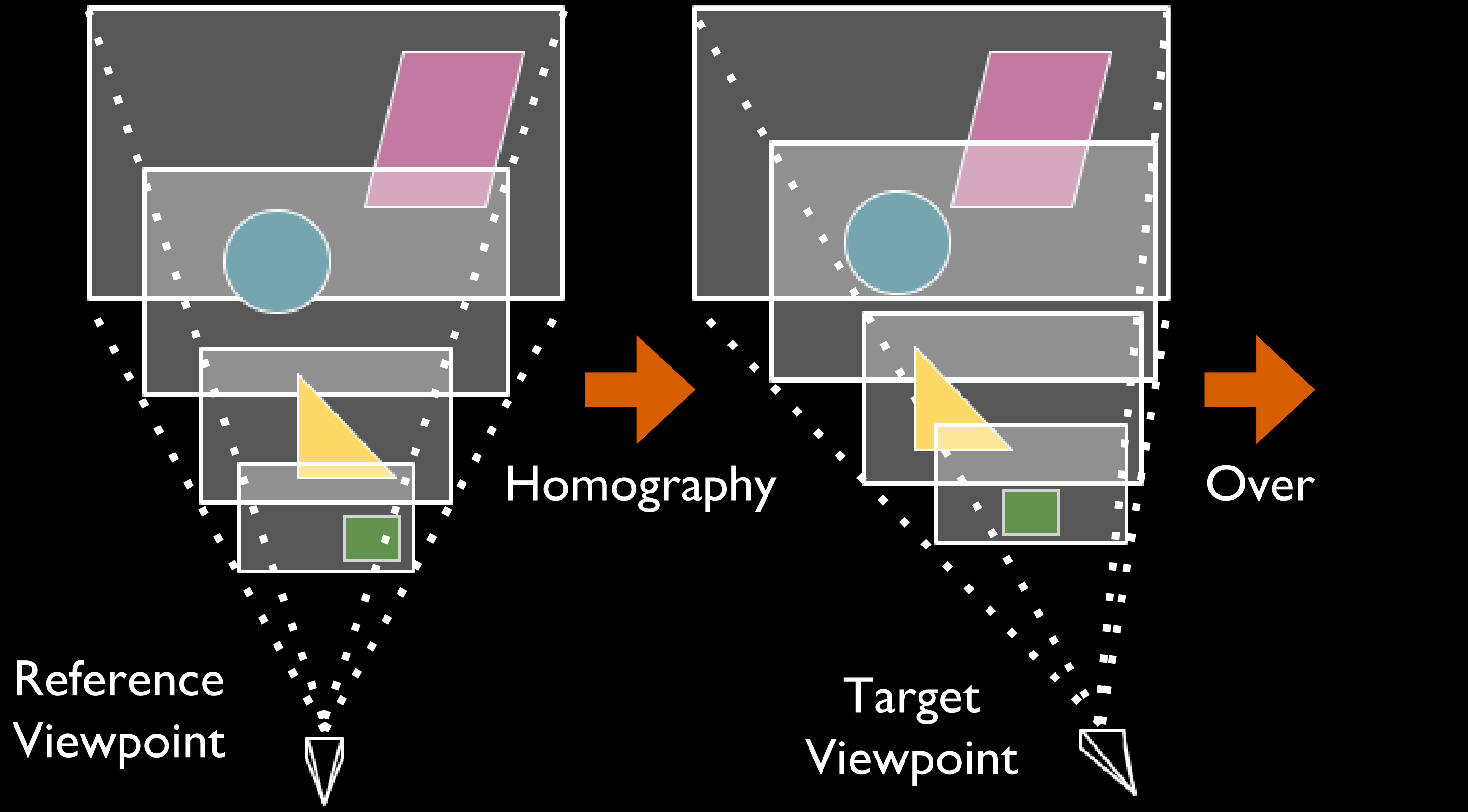
https://www.youtube.com/watch?v=kN-eCBAOw60 (from 1957)

Multiplane Images (MPIs)

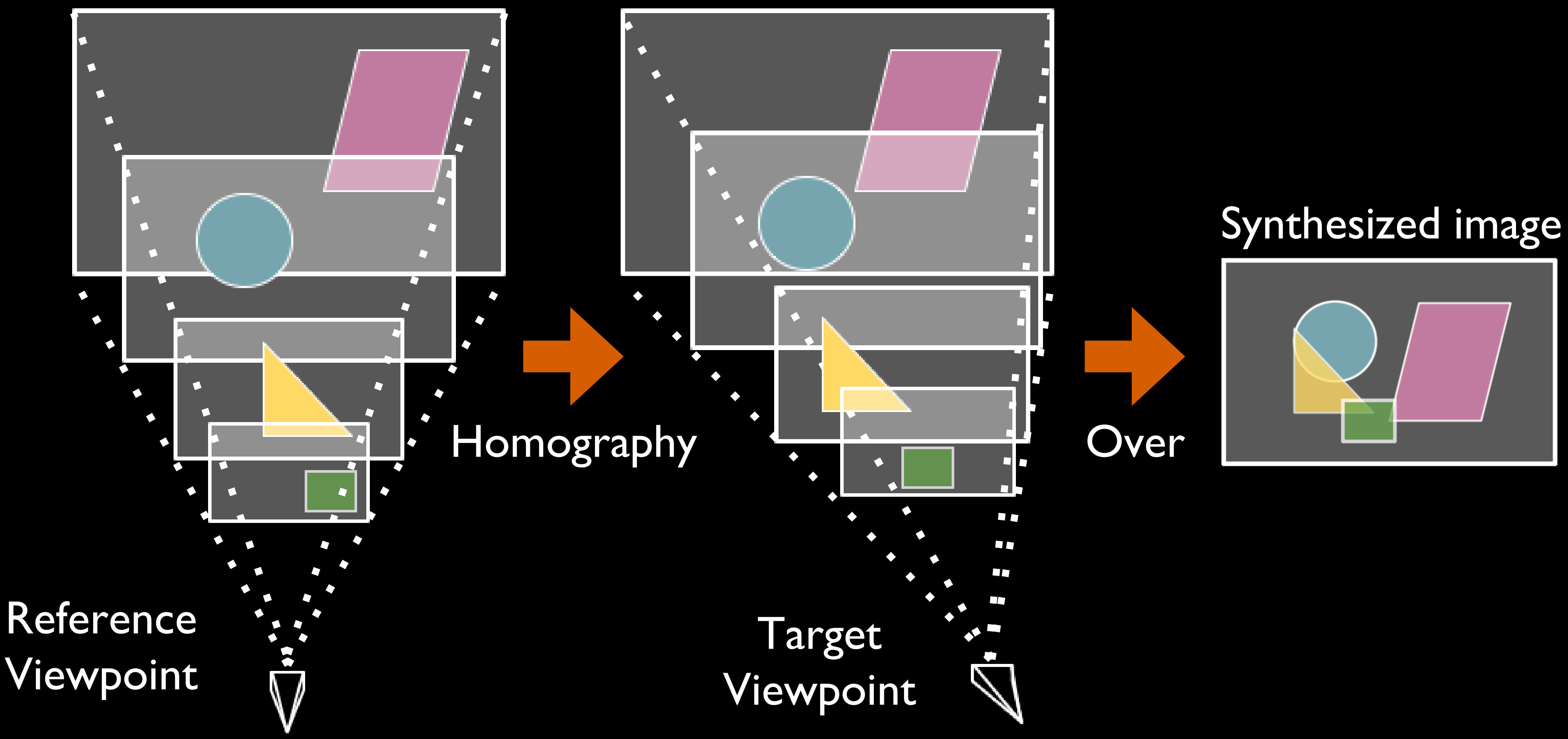


Each plane is at a fixed depth and encoded by an RGBA image

View Synthesis using Multiplane Images

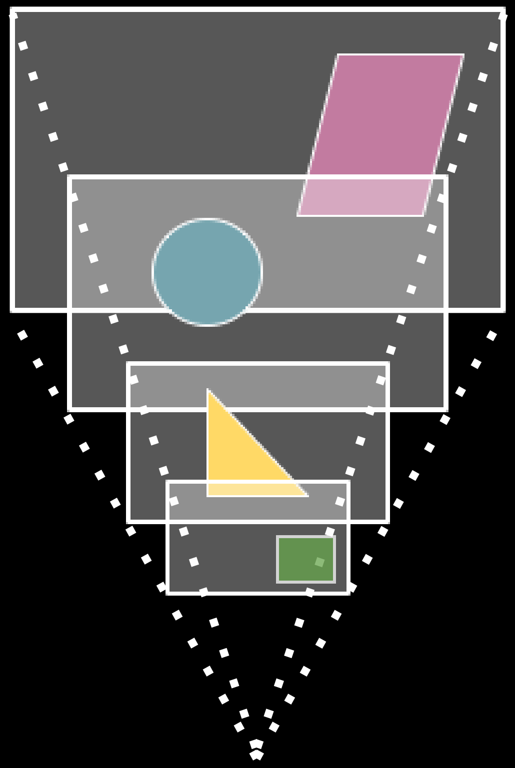


View Synthesis using Multiplane Images



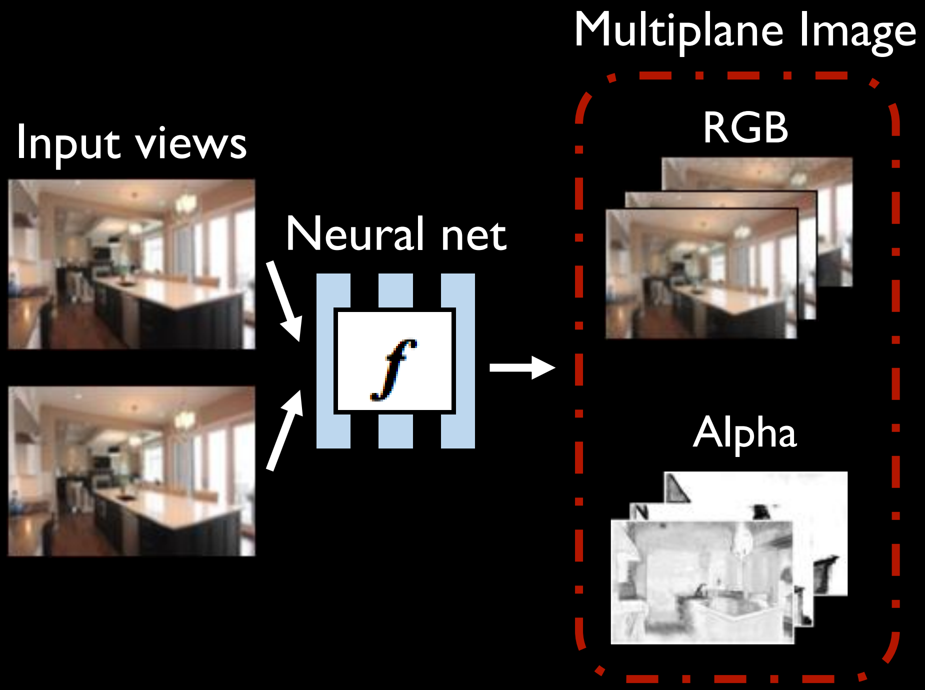

MAZZER

Properties of Multiplane Images

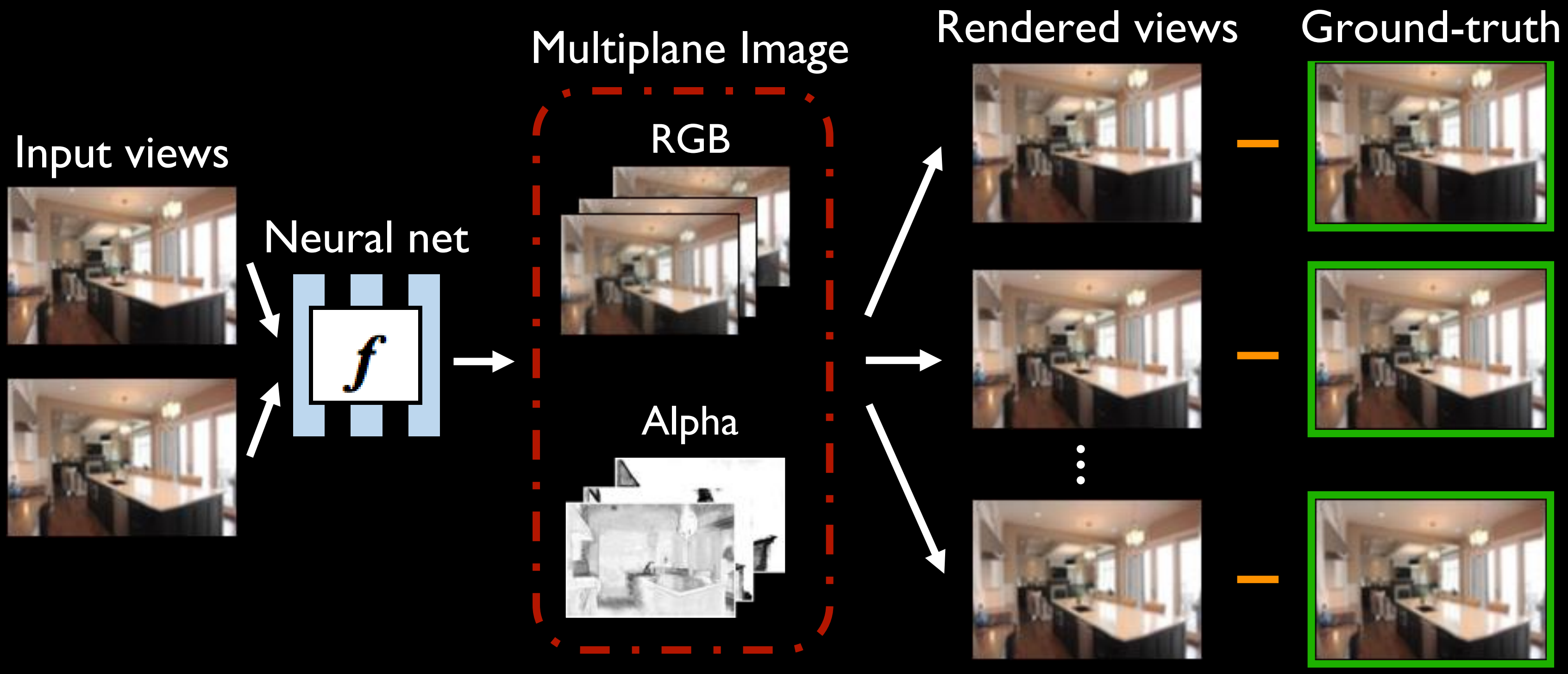


- Models disocclusion
- Models soft edges and non-Lambertian effects
- Efficient for view synthesis
- Differentiable rendering

Learning Multiplane Images



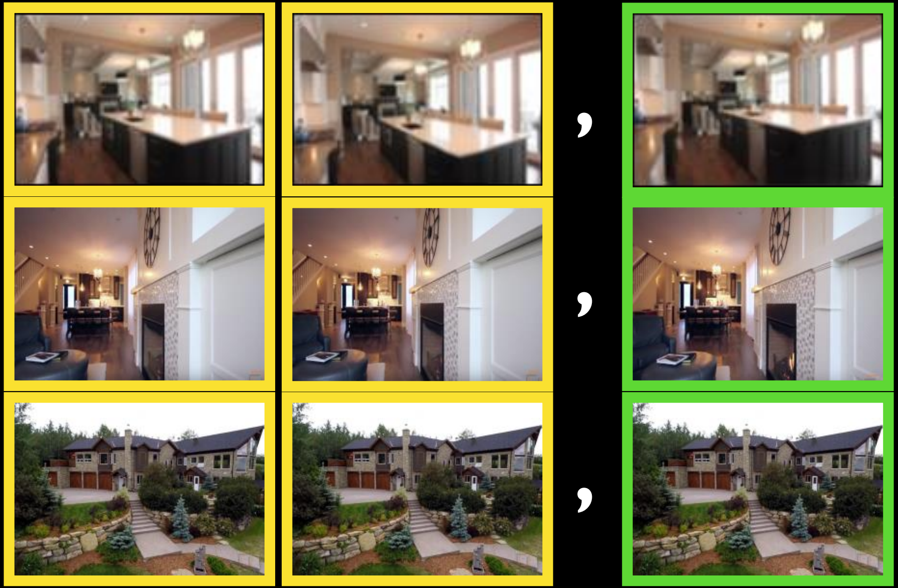
Learning Multiplane Images



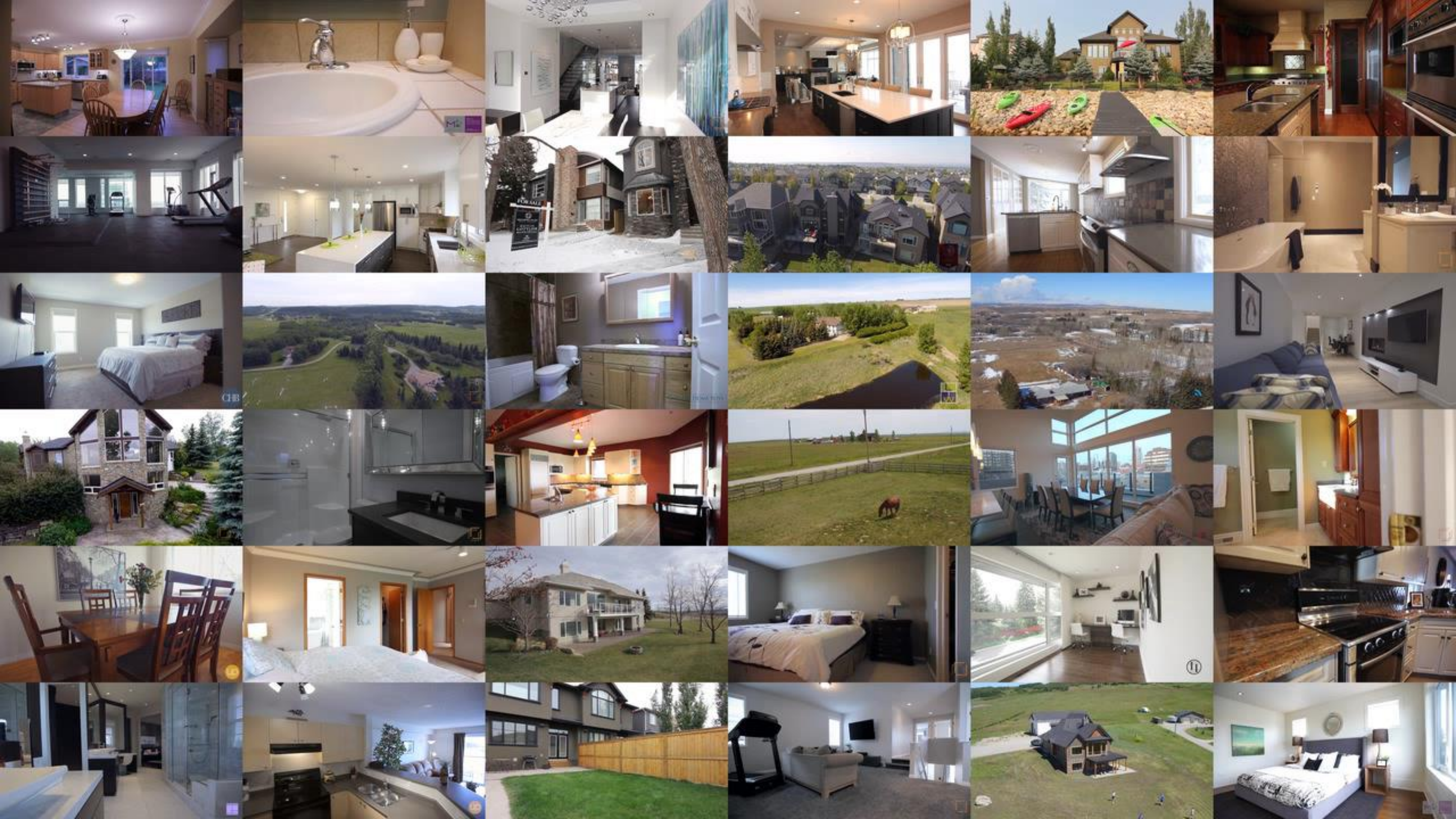
Training Data

Input views

Target view

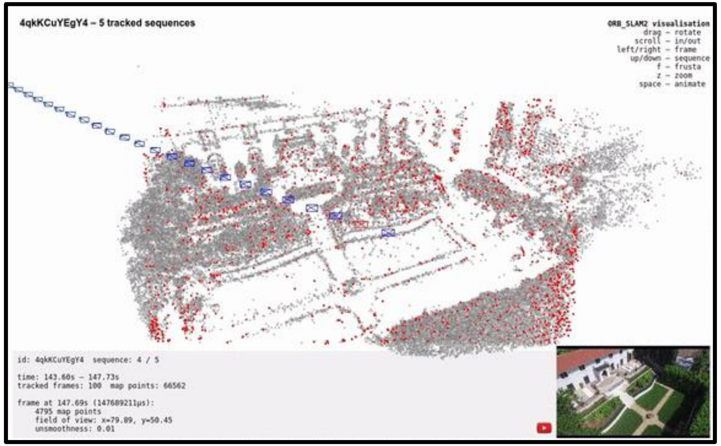


Need massive set of triplets with known camera poses


FOR SALE

RealEstate10K

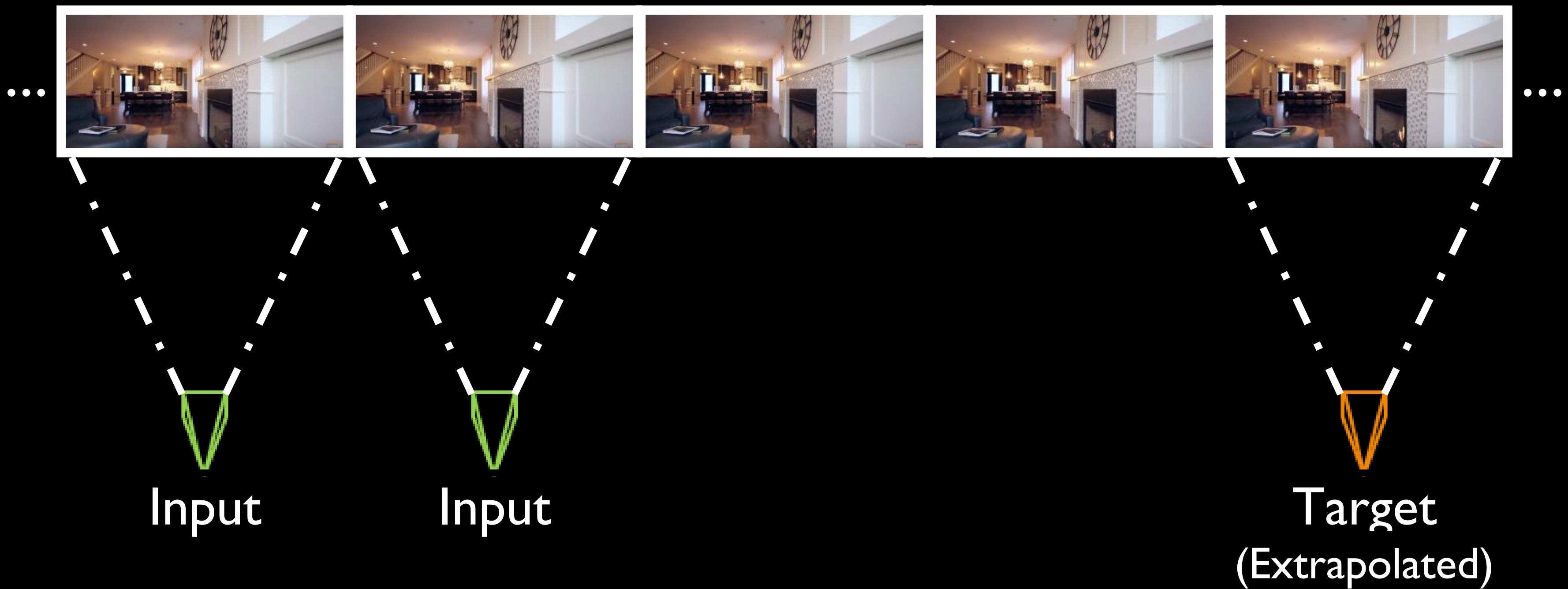
SLAM



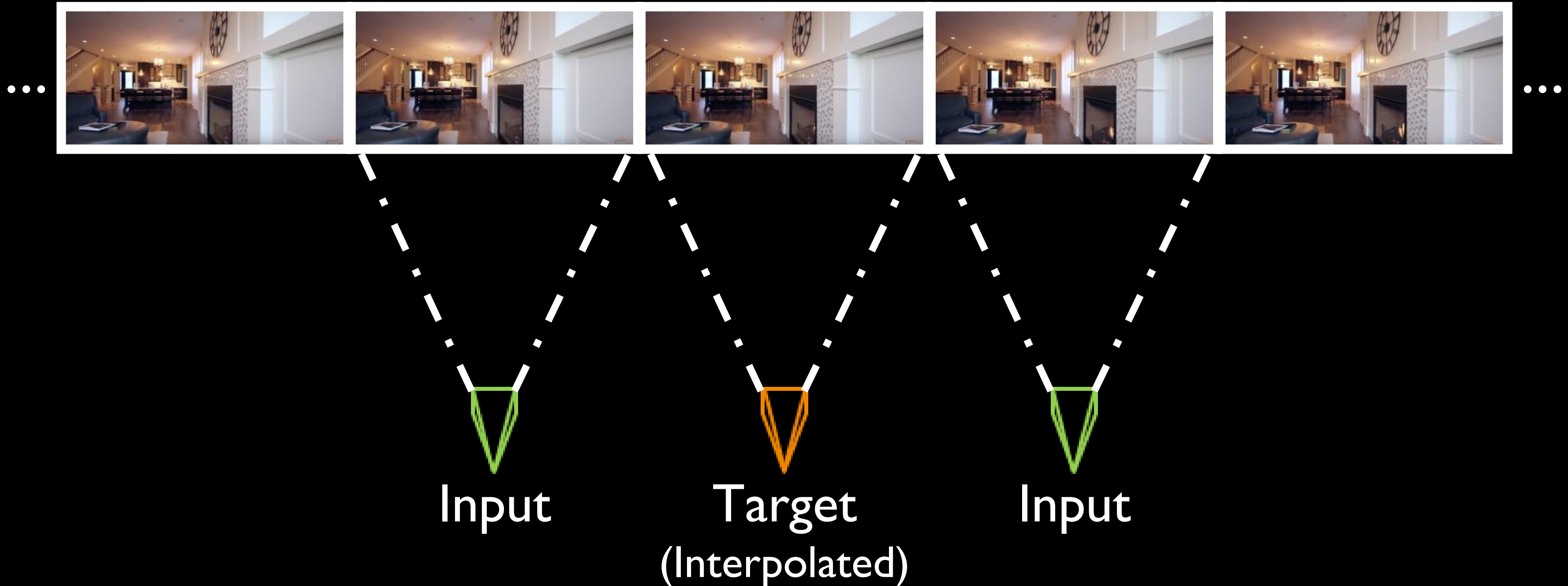
10 million frames from **80,000 video clips** from **10,000 videos**

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**

Questions?

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, *especially people*

Statues vs. people

https://www.balletforadults.com/back-to-basics-the-five-positions-of-the-arms/

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 (best paper runner up)

MannequinChallenge Dataset

- 2000 YouTube videos
- Frozen people, moving camera
- Diverse scenes, natural poses


SEMINÁRIO

MannequinChallenge Training Data

"Ground truth" depth from SfM + Multi View Stereo (MVS)

Input video

Estimated depth

Removing Humans for View Synthesis

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...

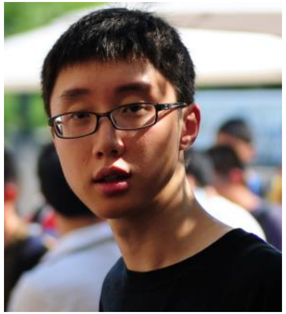
Future directions

- 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

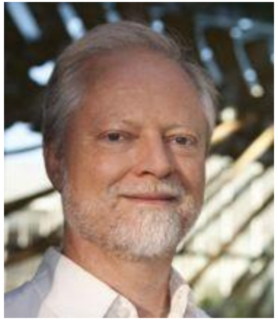

Zhengqi Li


Tinghui Zhou


John Flynn


Graham Fyffe


Shubham Tulsiani


David Lowe


Matt Brown

Questions?