slido

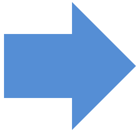
Which month is the best month?

ⓘ Start presenting to display the poll results on this slide.

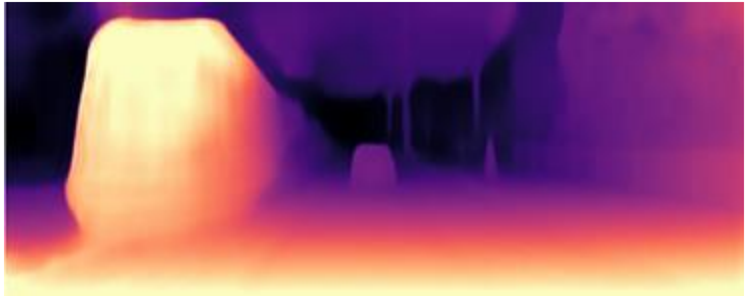
CS5670: Computer Vision

Deep Learning and 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 tomorrow, Tuesday, May 11 at 7pm
- Take-home final exam to be released May 12, due May 17 by 7pm
- Wednesday: course review / wrap up (last lecture of class)
- No quiz on Wednesday

Third Place

Luca Spinazzola and Nicolas Carchio

Second Place

Xindong Chen and Yehao Zhang

First Place

Kaiyuan Deng and Scarlett Zhang

3D Computer Vision

80s - 90s: RANSAC, Fundamental matrix

2000s: stereo, multiview stereo, Internet-scale 3D reconstruction

2015+: **learning-based 3D vision**

Single-view modeling (late 1990s)

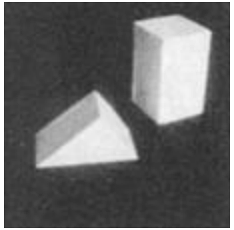
Vermeer's *Music Lesson*

Reconstructions by Criminisi et al.

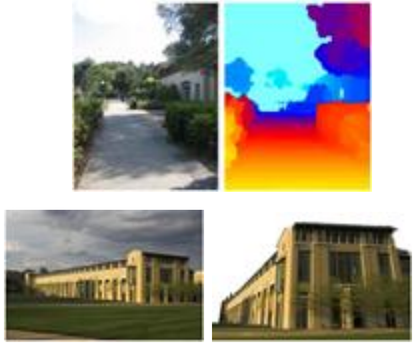
Can we use deep learning to predict geometry from a single image?

Astonishing recent progress in learning 3D perception

"Blocks world"
Larry Roberts
(1963)

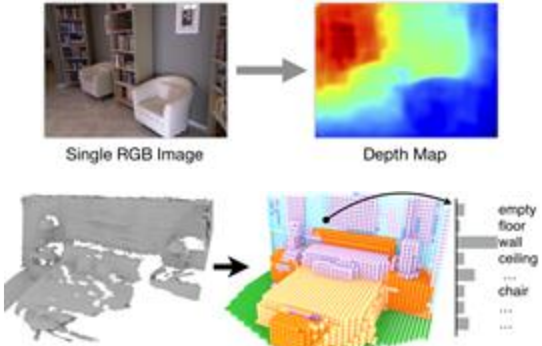


Pre-deep era
(2005)



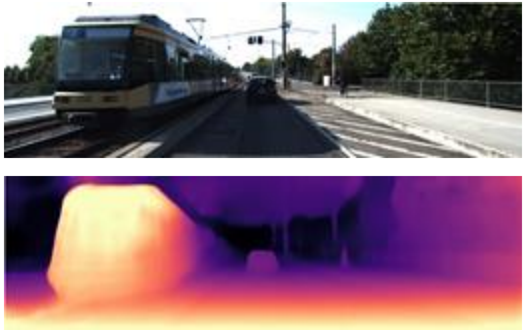
[Saxena, Chung, Ng, NIPS 2005]
[Hoiem, Efros, Hebert, SIGGRAPH 2005]

Supervised deep learning
(2014-)



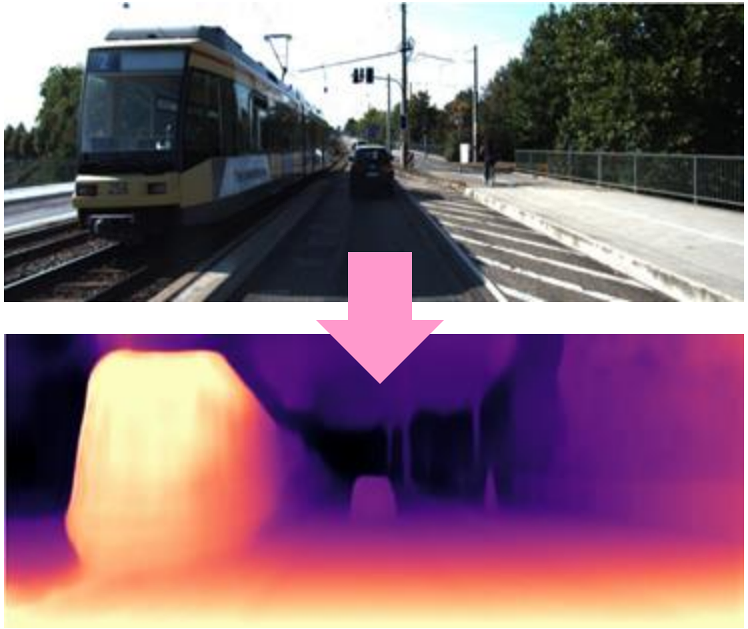
[Eigen, Puhrsch, Fergus, NIPS 2014]
[Song et al, CVPR 2017]
...
go/im2depth

Multi-view supervision
(2016-)

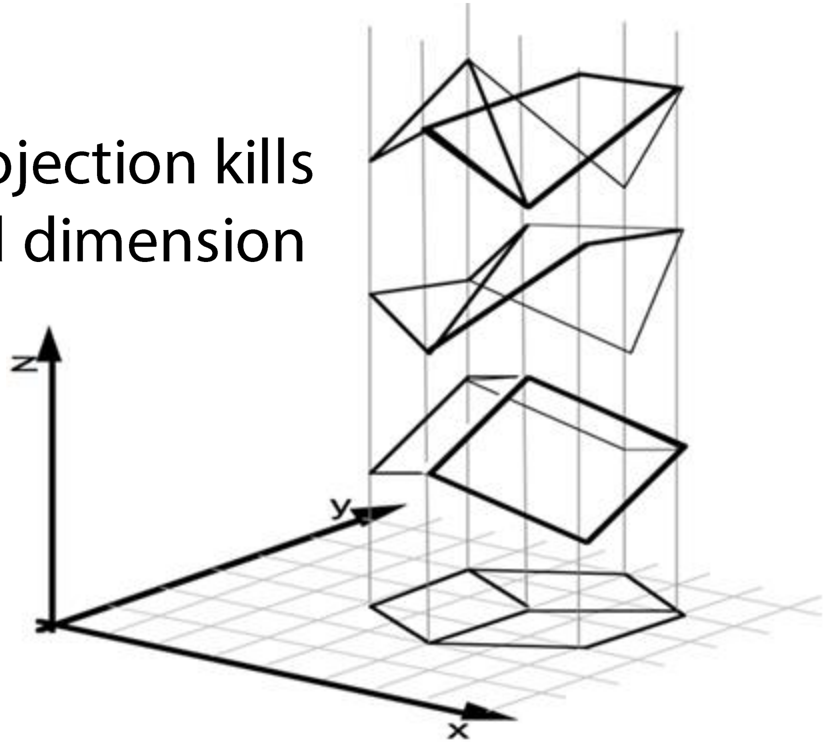


[Garg, Kumar BG, Carneiro, Reid, ECCV 2016]
[Xie, Girshick, Farhadi, ECCV 2016]
[Zhou, Brown, Snavely, Lowe, CVPR 2017]
[Vijayanarasimhan, et al., 2017]
[Godard, Mac Aodha & Brostow, CVPR 2017]
[Mahjourian, Wicke & Angelova, CVPR 2018]
...

Canonical problem: single RGB view to depth



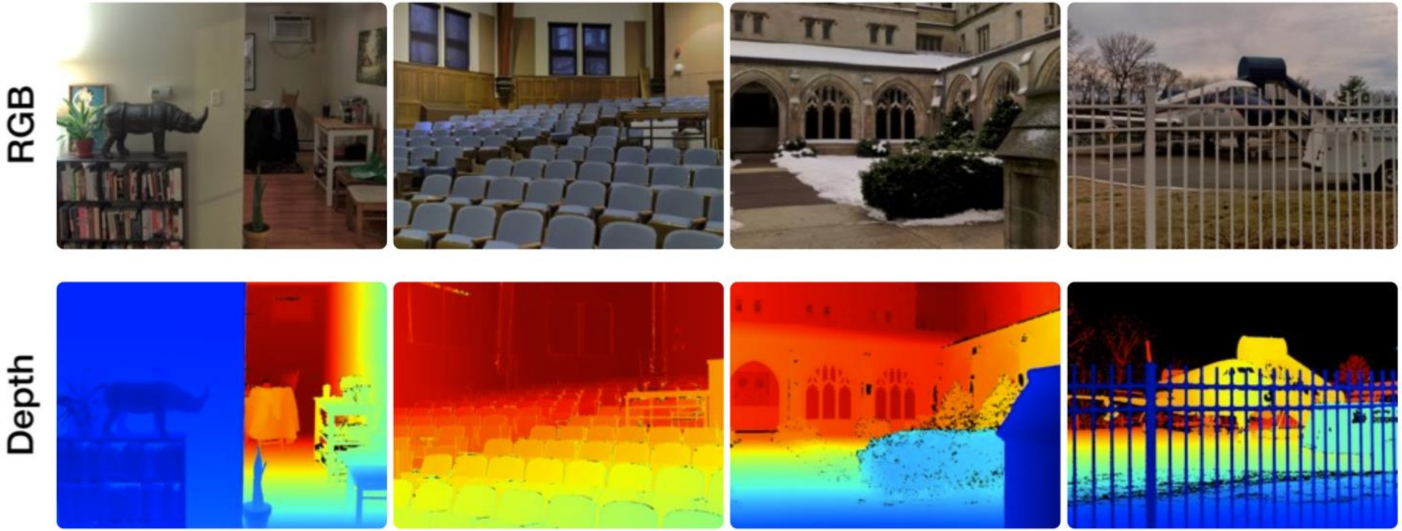
but projection kills
the 3rd dimension



[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

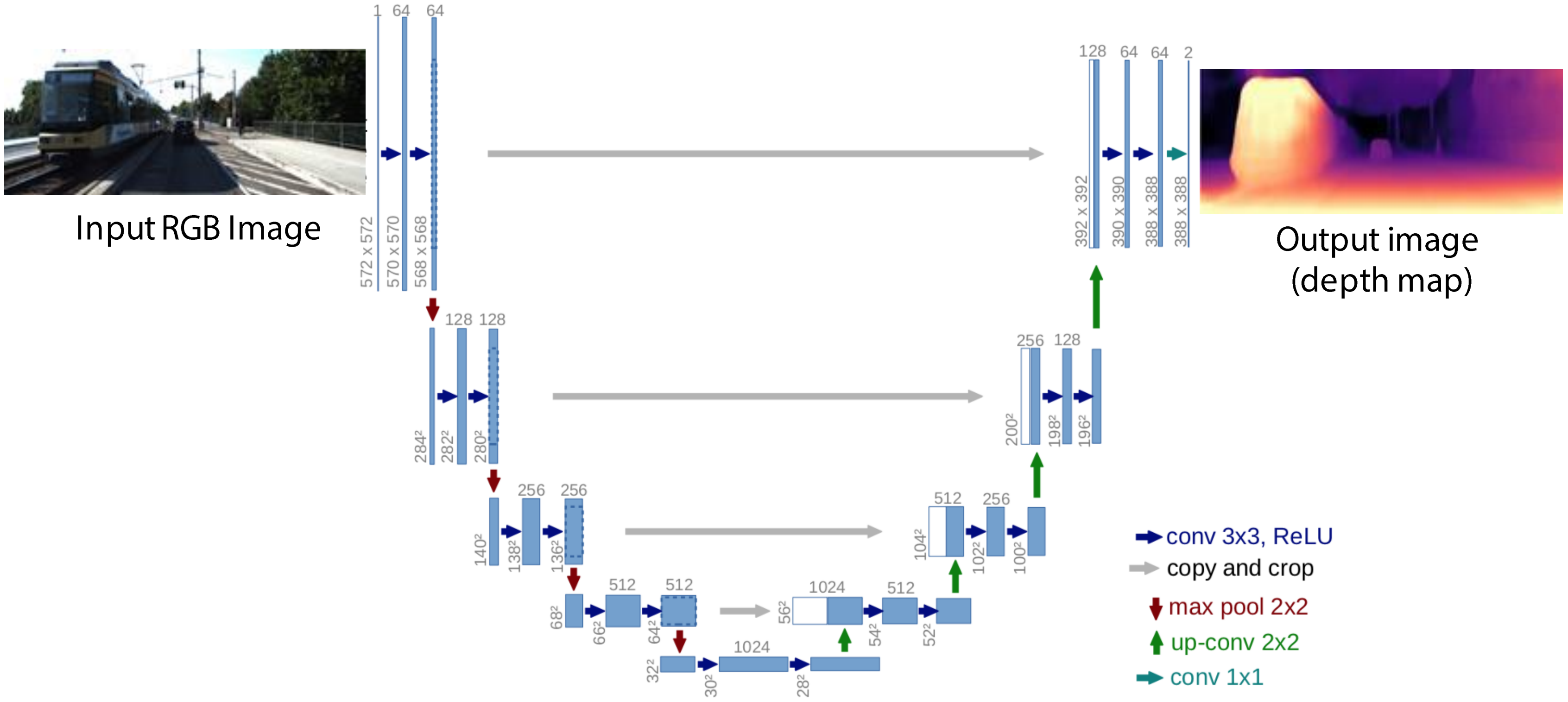


Source: https://diode-dataset.org/

CNN architectures for single-view depth

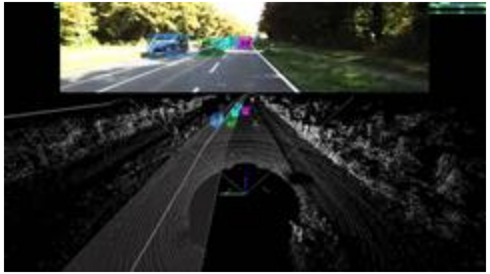
- Need an architecture that takes in an image (an RGB image) and produces another image (a depth map)
- Similar to other problems where images are the outputs (e.g., semantic segmentation, colorization, object boundary detection)
- In contrast to image classification, where outputs are probabilities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

Common choice: UNet architecture



How to get training data?

LIDAR



KITTI [Geiger et al. 2012]

Kinect

NYU [Eigen et al. 2014]

Manual annotation

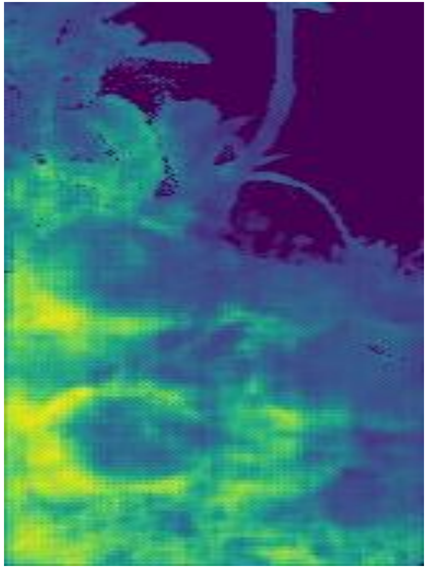
Depth in the Wild [Chen et al. 2016]

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

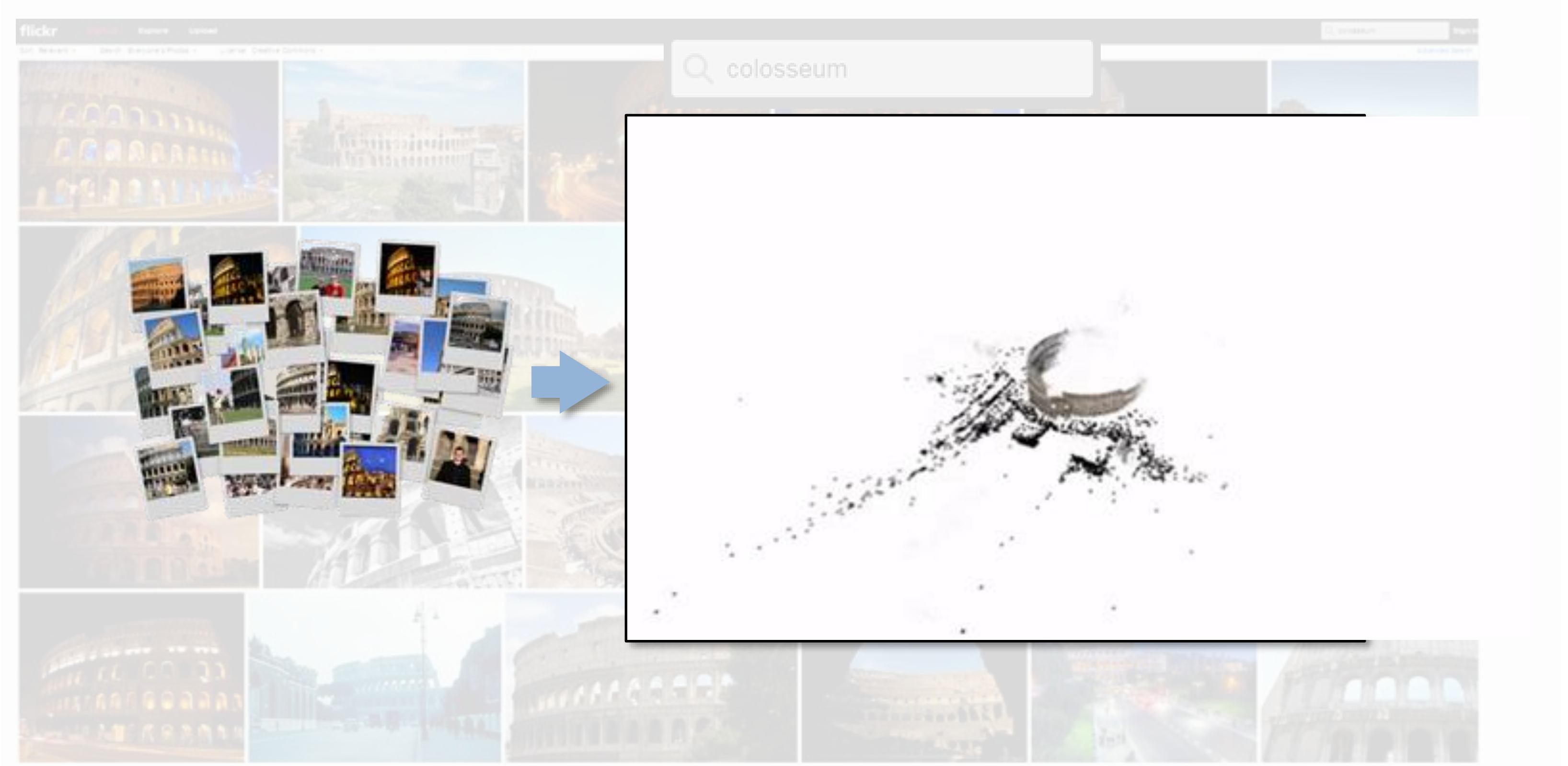
How can we gather more diverse data?

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

Training: Multiple views

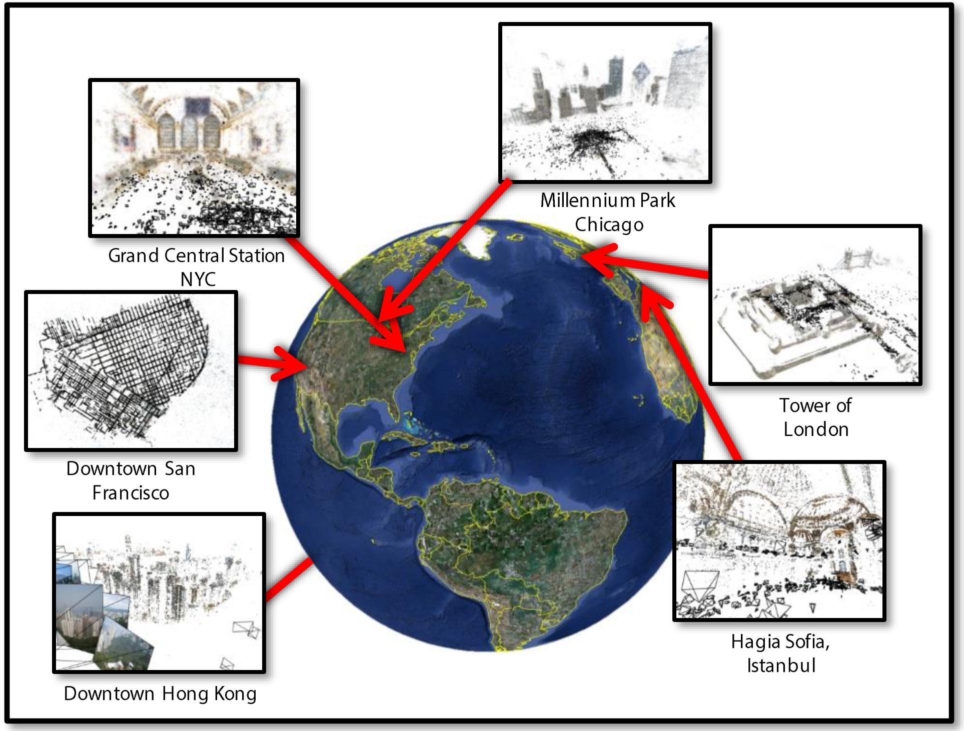
Testing: Single Image

Idea 1: Structure from Motion reconstructions



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

Reconstructing the World's Landmarks



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

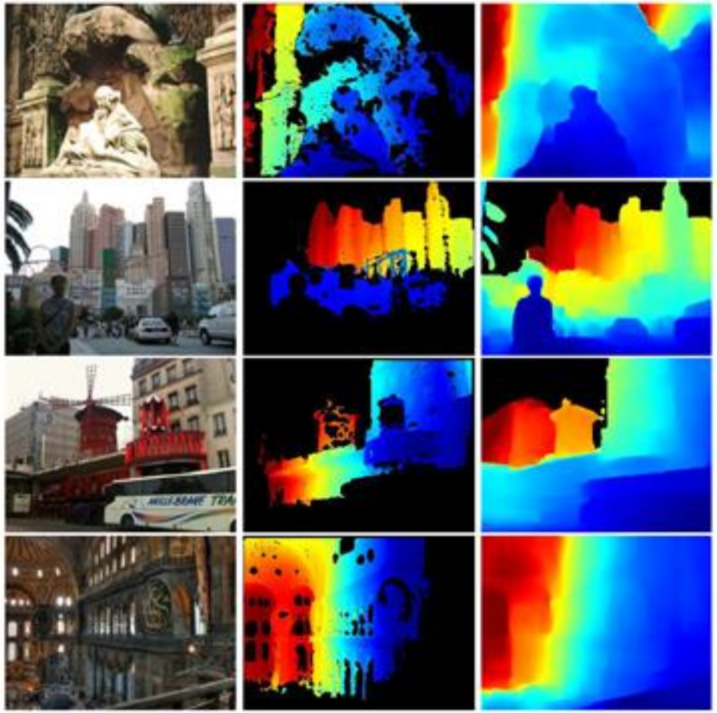
MegaDepth dataset

>130K (RGB, depth map) pairs

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

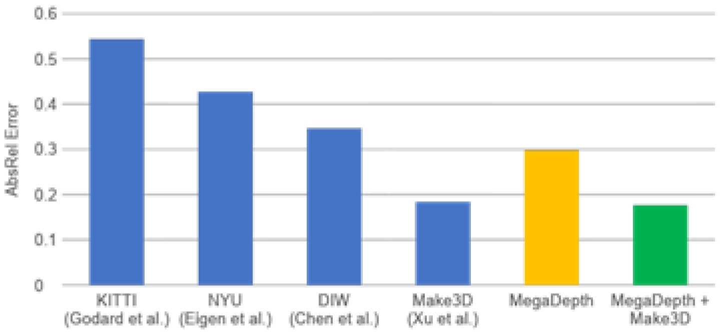
[Zhengqi Li and Noah Snavely. *MegaDepth: Learning Single-View Depth Prediction from Internet Photos*. CVPR 2018]

MegaDepth-trained prediction results

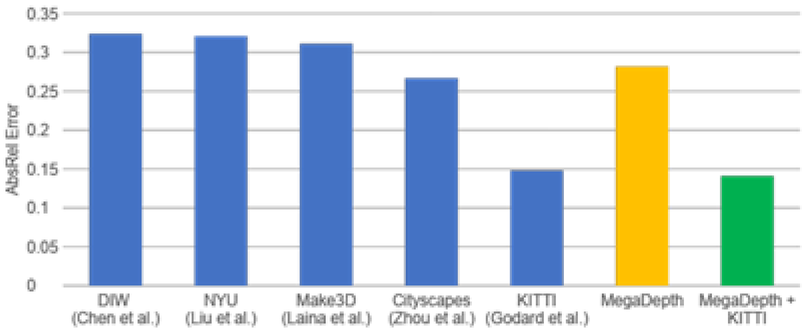


Input Ground truth Predicted depth

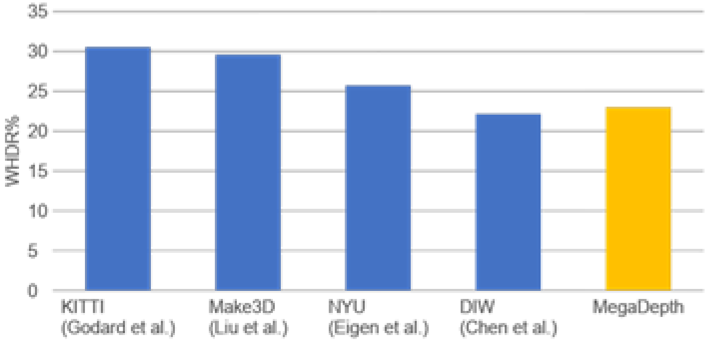
Internet data generalizes well



Train on X, test on Make3D

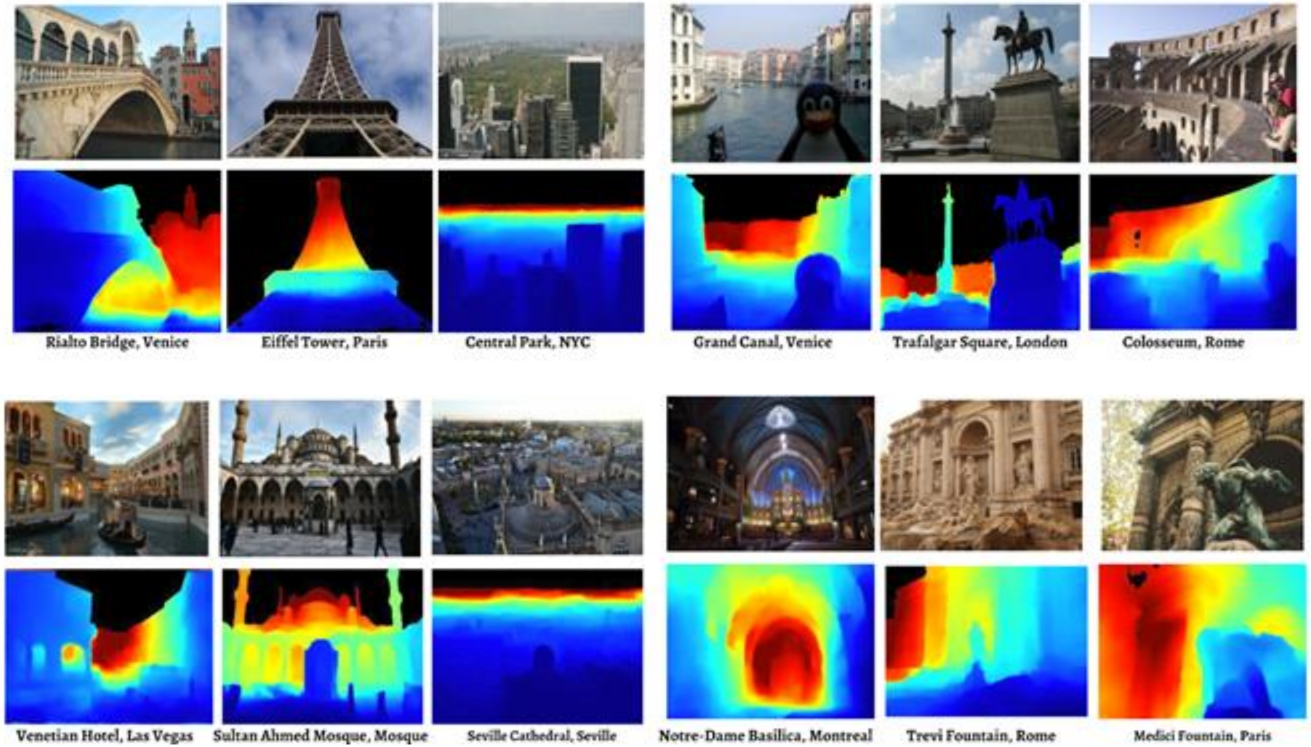


Train on X, test on KITTI



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

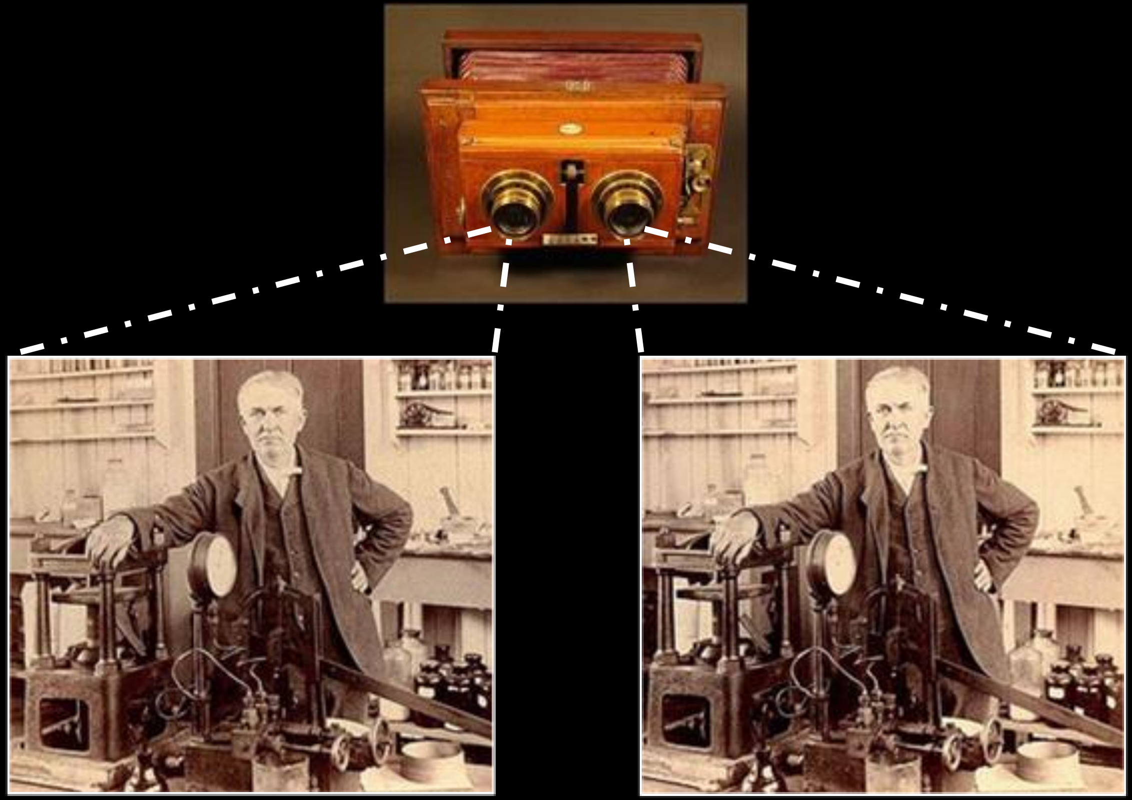
- So much for single-view depth
- Another thing we might want to do is *render new views of the captured scene* (i.e., view synthesis)
- Involves more than just depth, but also filling in missing content behind the foreground

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/

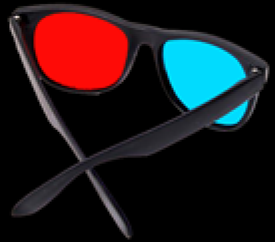
3D Photography using Context-aware Layered Depth Inpainting

Stereo Photography



Stereo Photography

Viewing Devices



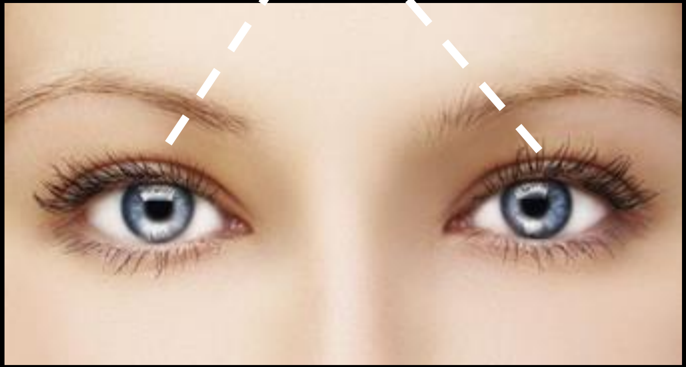
Stereo Photography

Queen Victoria at World Fair, 1851

Stereo Photography

Issue: Narrow Baseline

~6.5 cm



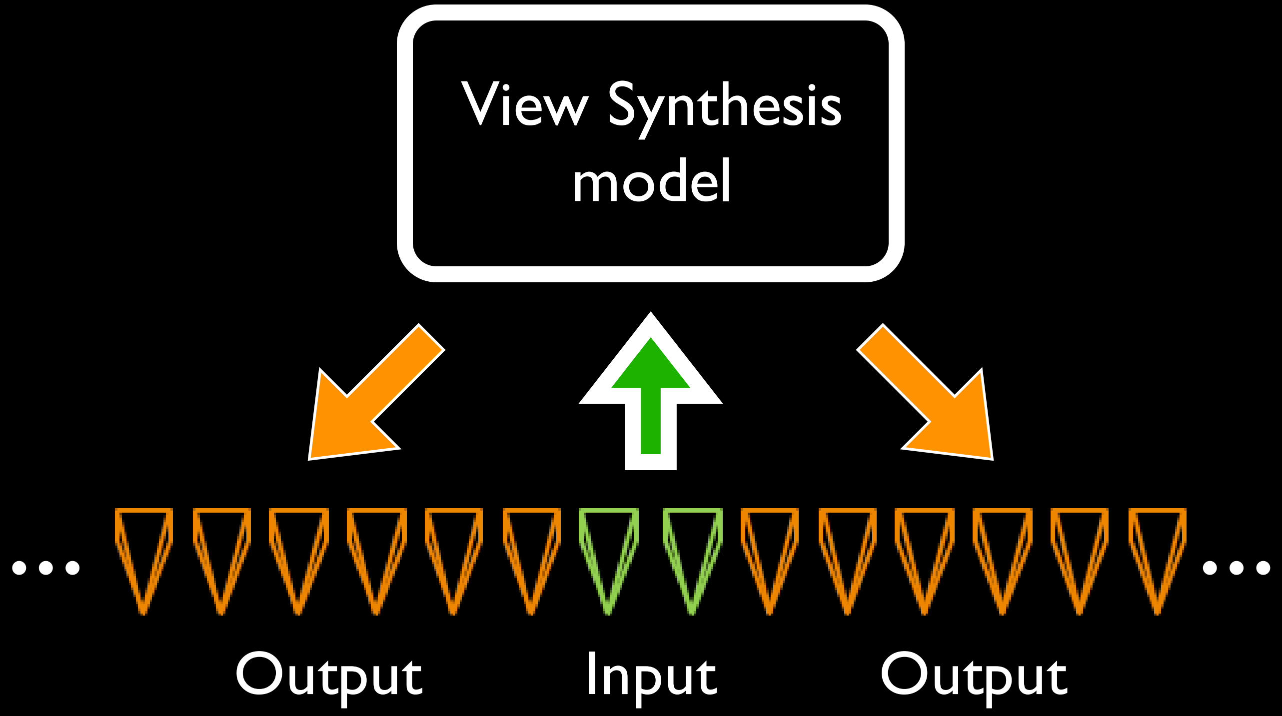
~1.5 cm


Left
MAZZER

Right

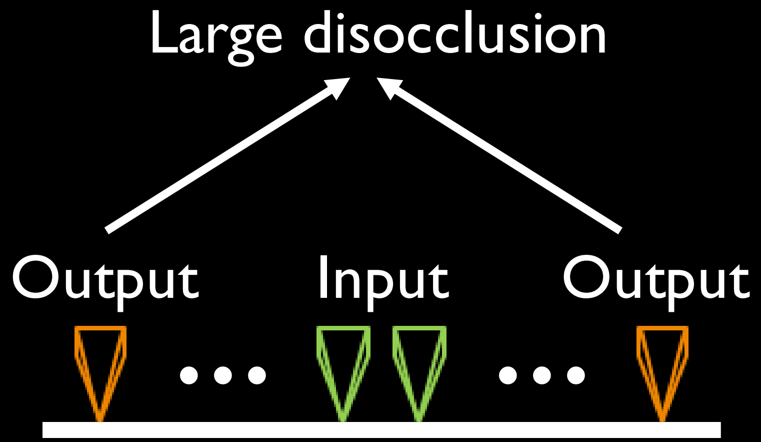

MAZZER

Problem Statement



Challenges

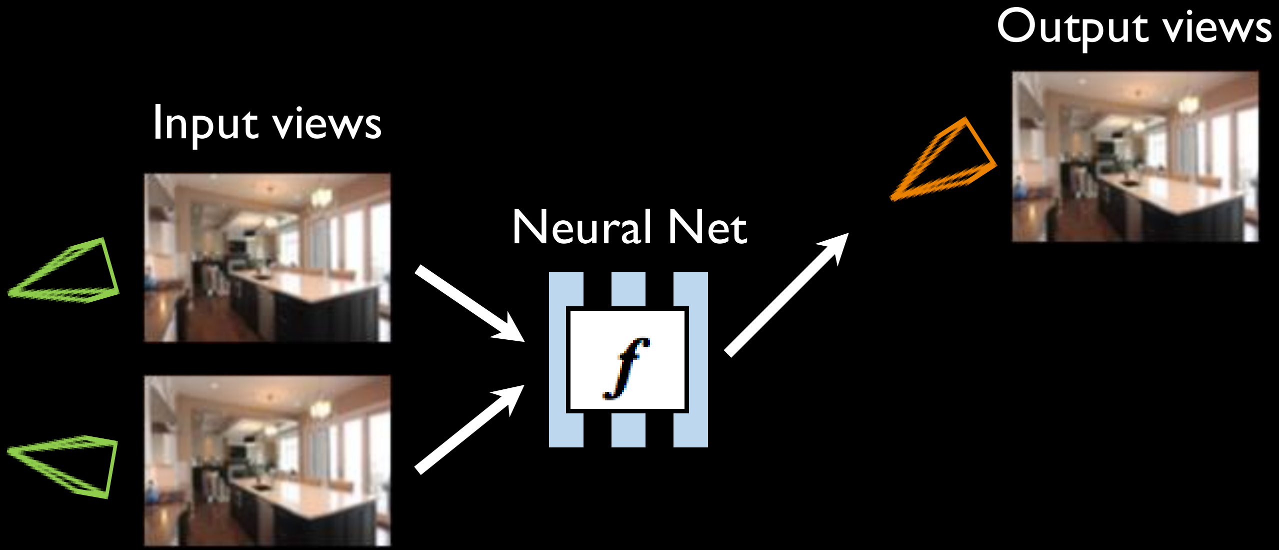
Extrapolation



Non-Lambertian Effects

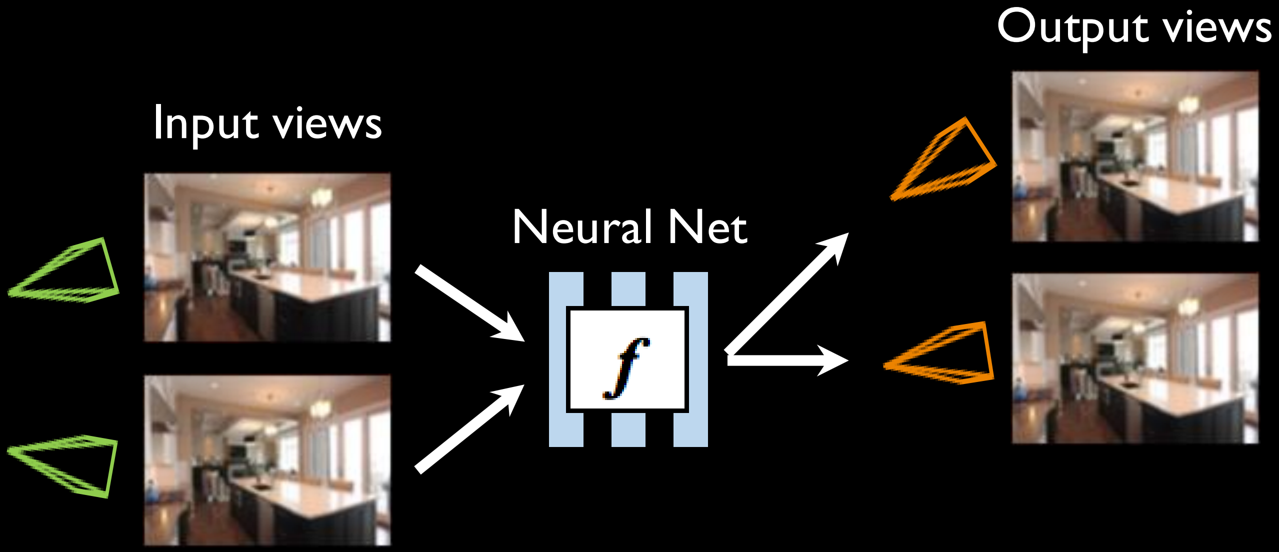
Reflections, transparencies, etc.

Prior Methods: No Shared Scene Representation



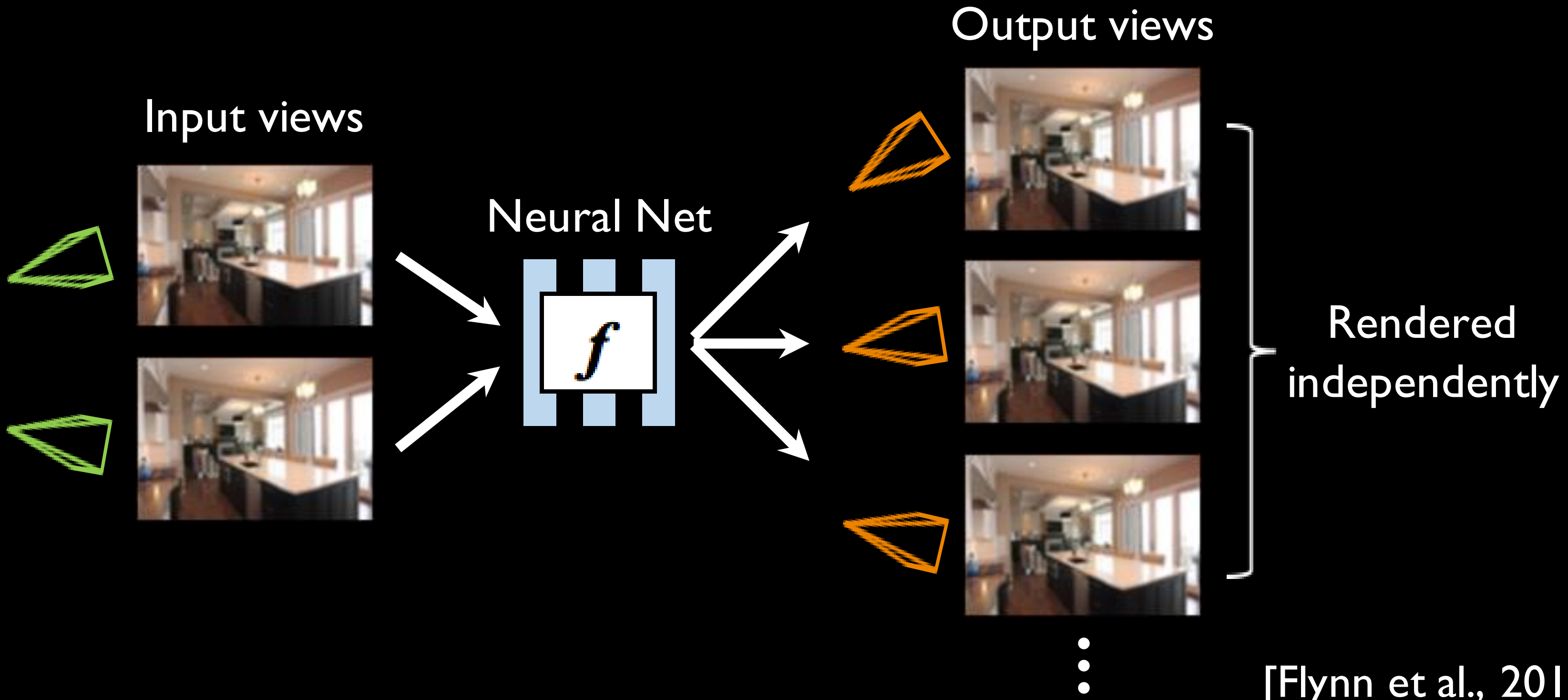
[Flynn et al., 2015]
[Kalantari et al. 2016]

Prior Methods: No Shared Scene Representation



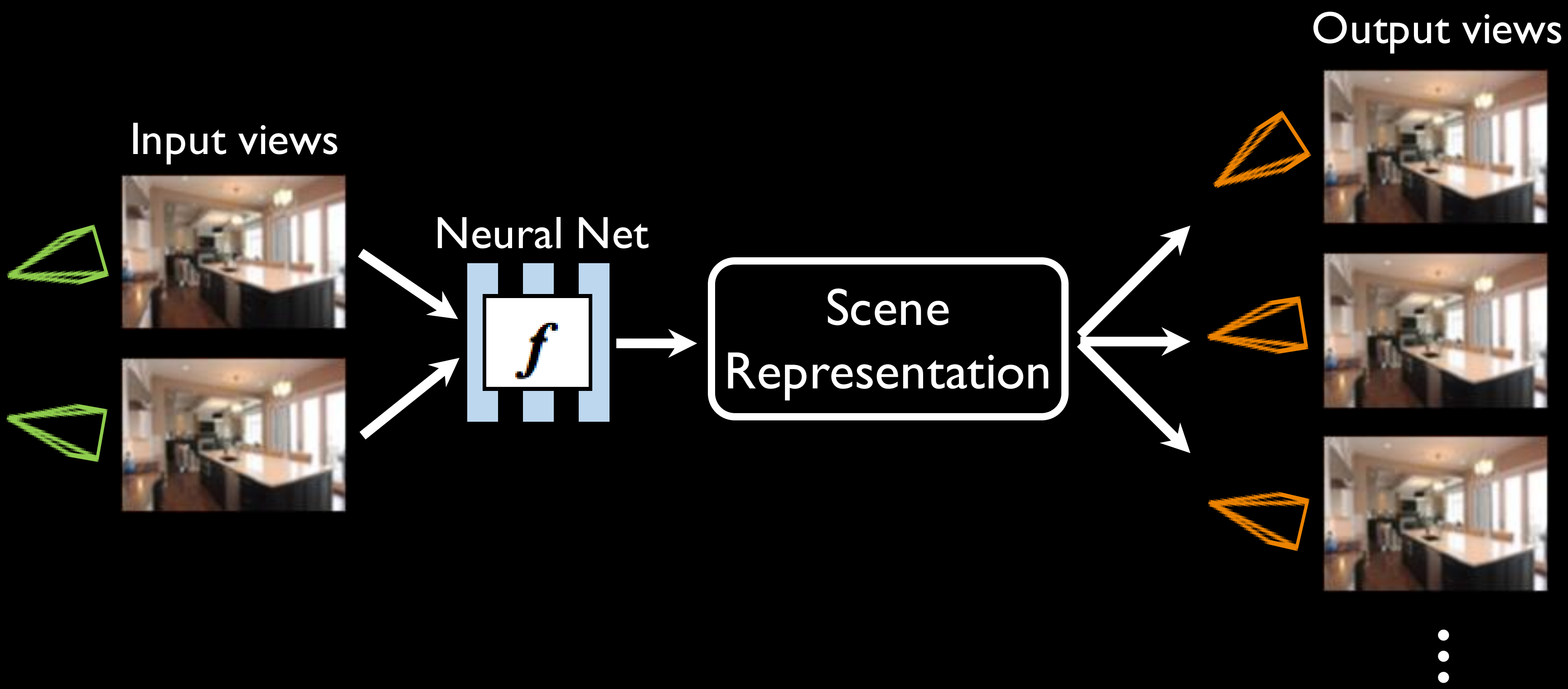
[Flynn et al., 2015]
[Kalantari et al. 2016]

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

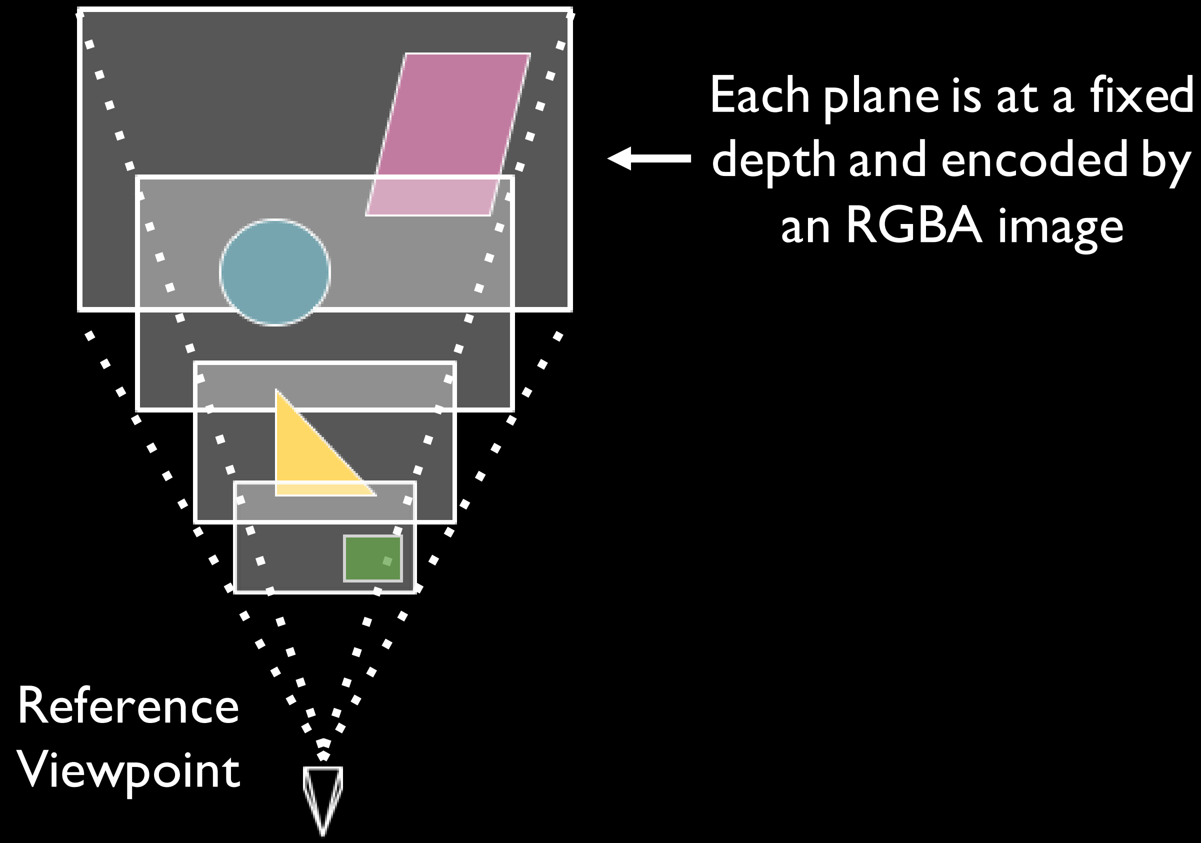
SIGGRAPH 2018

Multiplane Camera (1937)

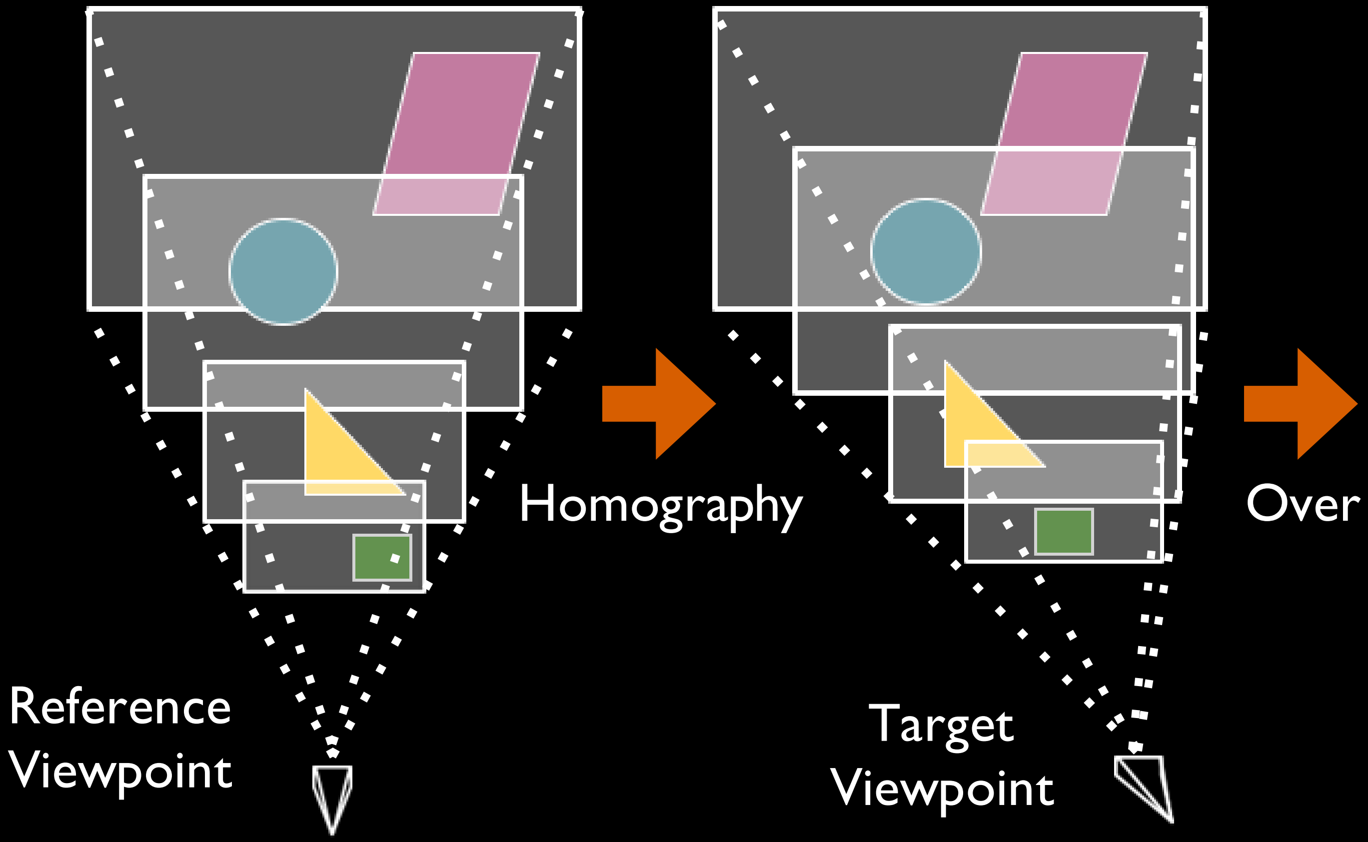
Image credits: Disney

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

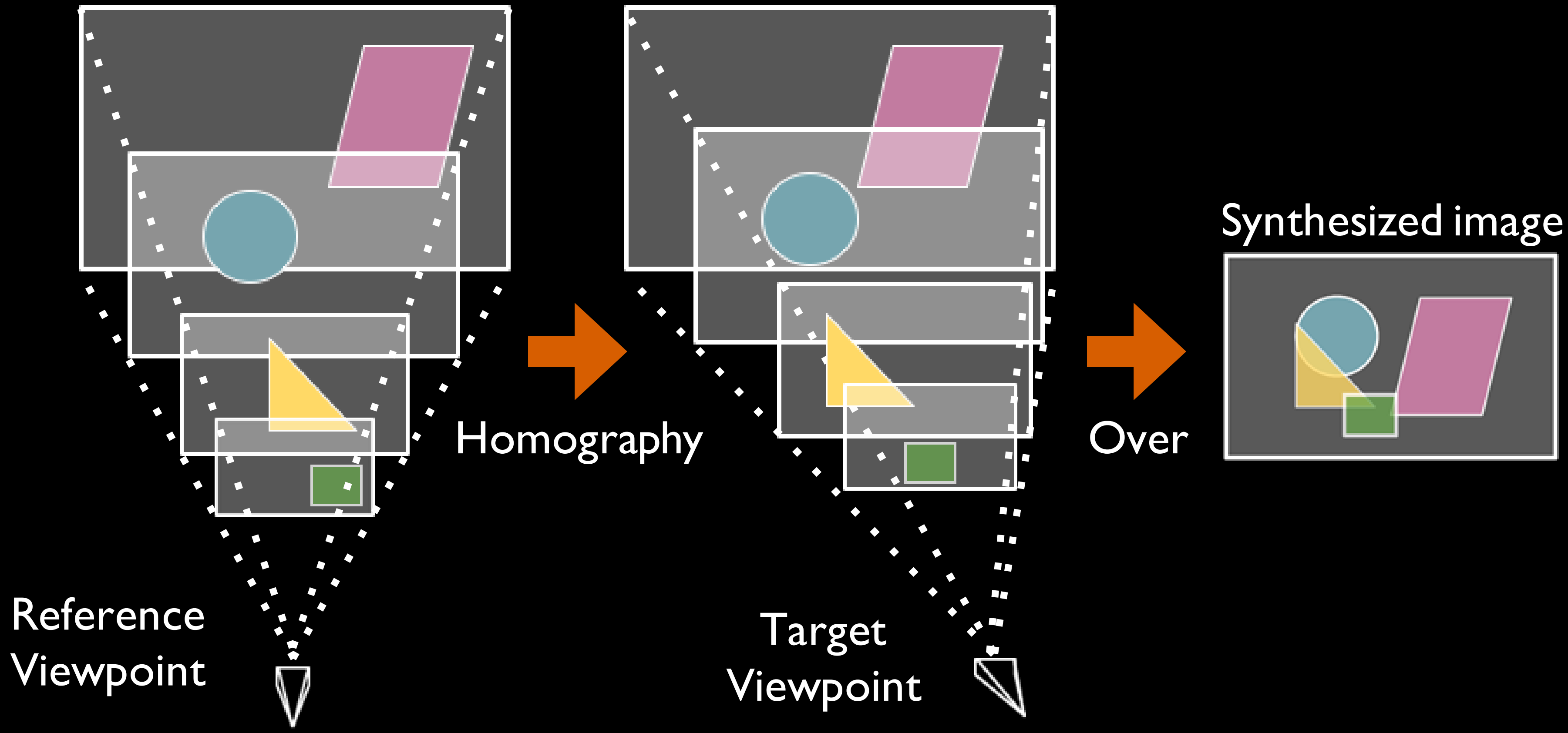
Multiplane Images (MPIs)

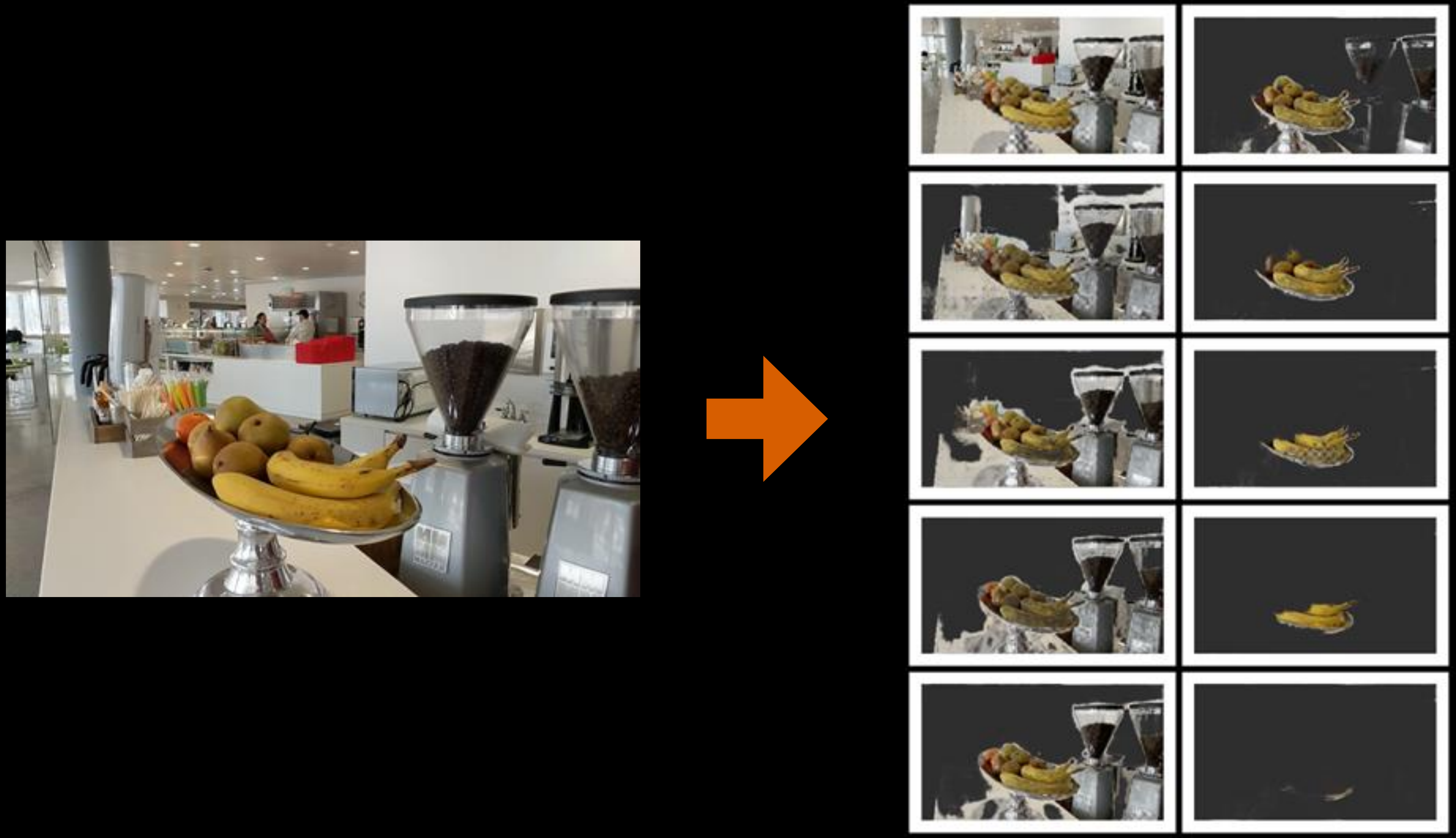


View Synthesis using Multiplane Images

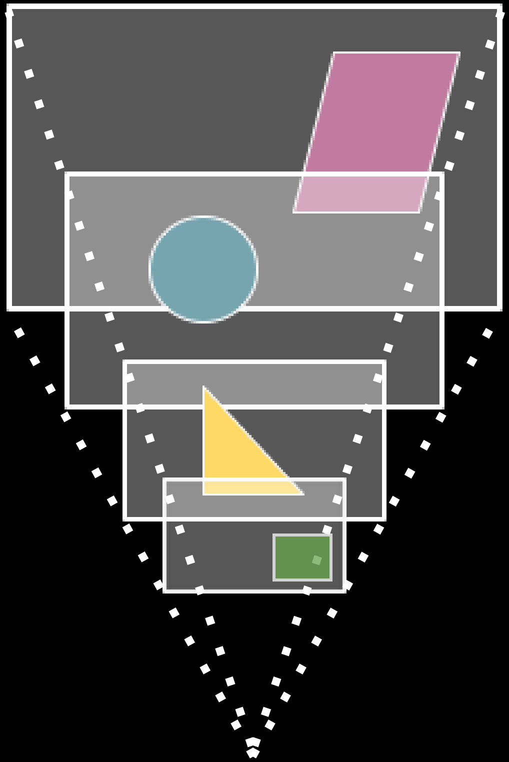


View Synthesis using Multiplane Images



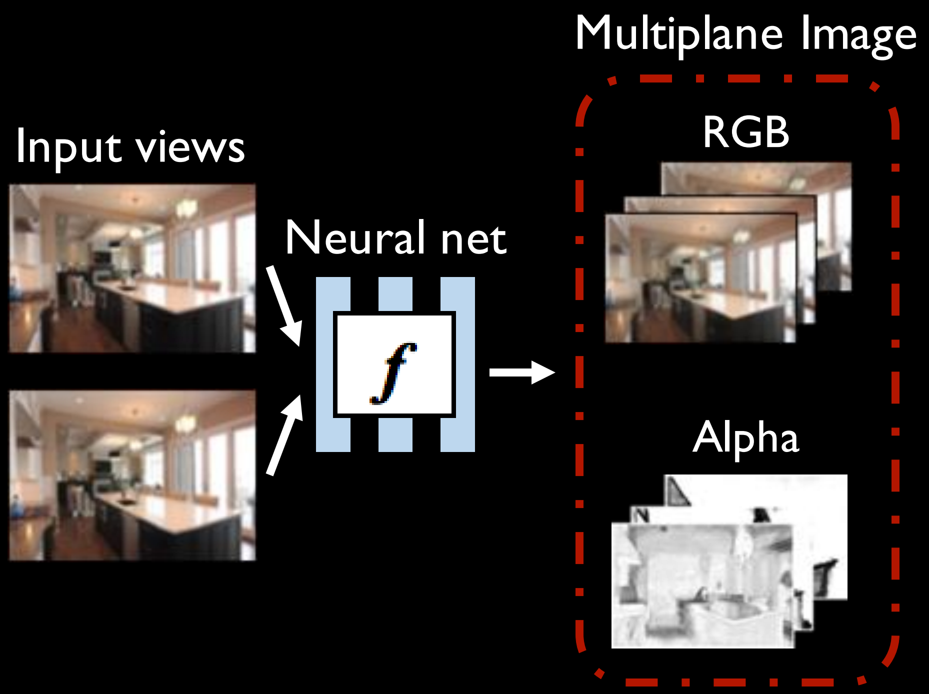


Properties of Multiplane Images

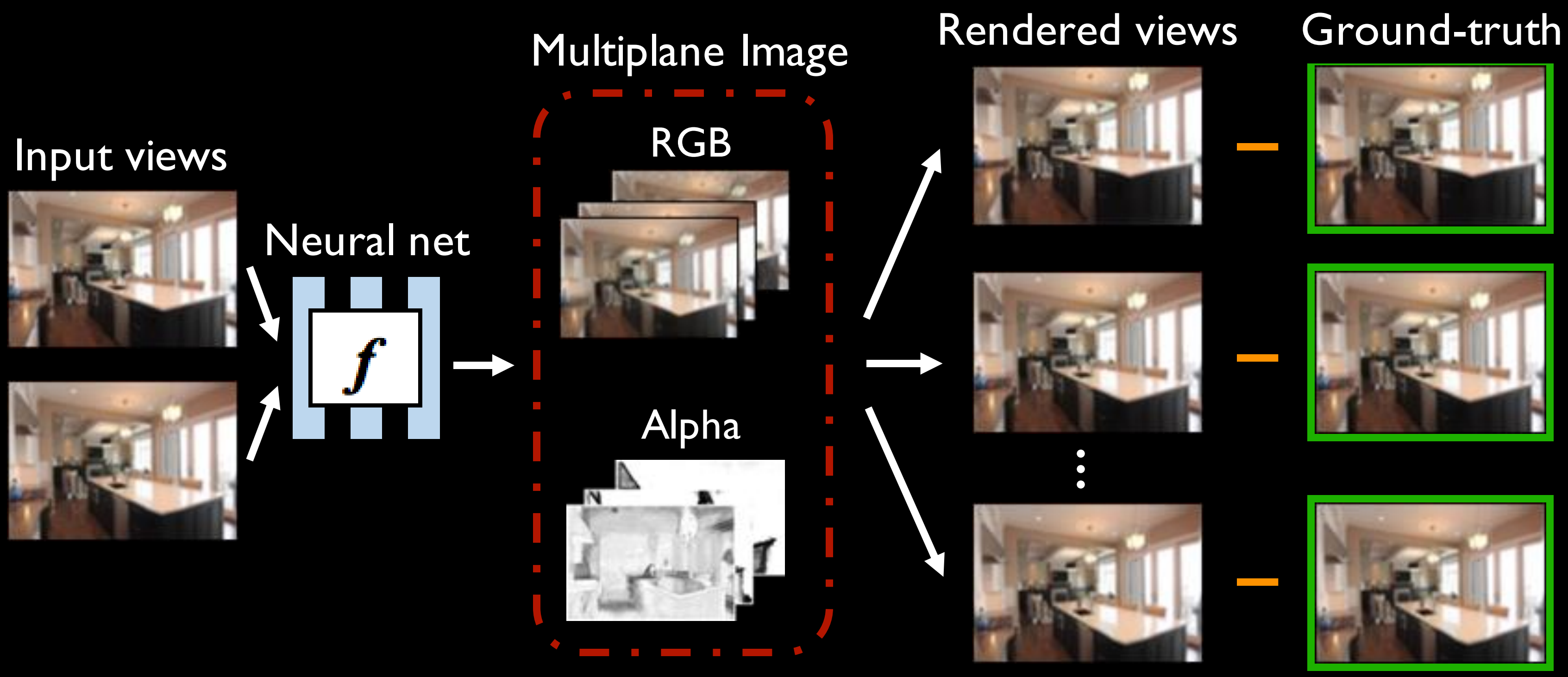


- Models disocclusion
- Models soft edges and non-Lambertian effects
- Efficient for view synthesis
- Differentiable rendering

Learning Multiplane Images



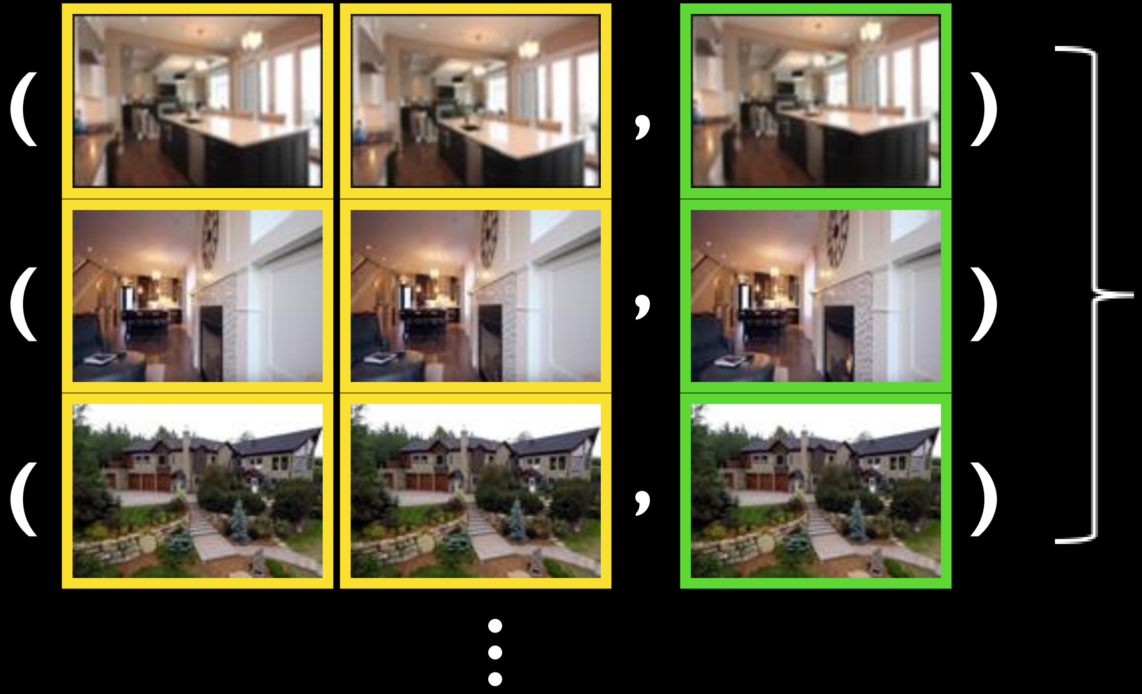
Learning Multiplane Images



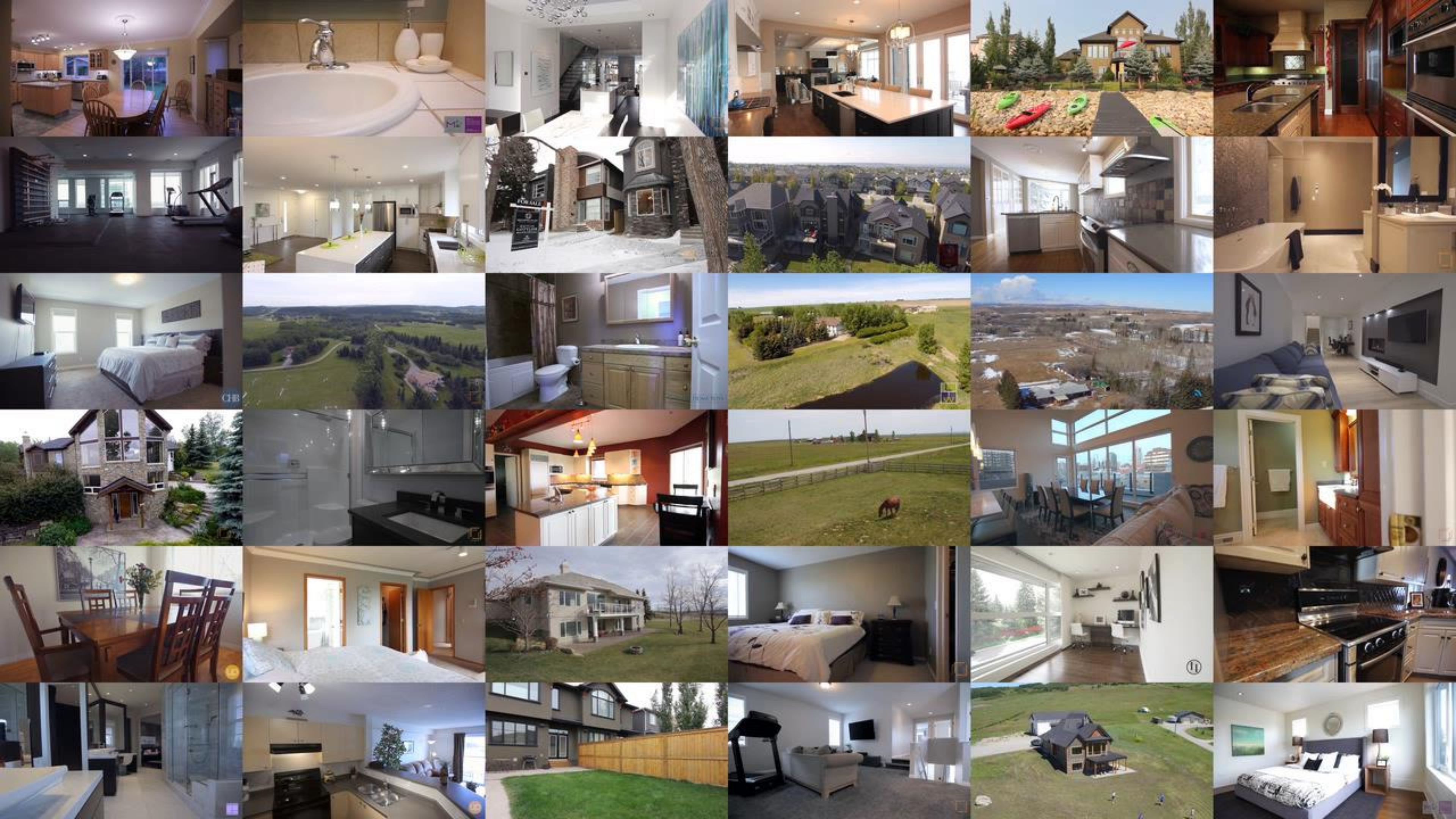
Training Data

Input views

Target view

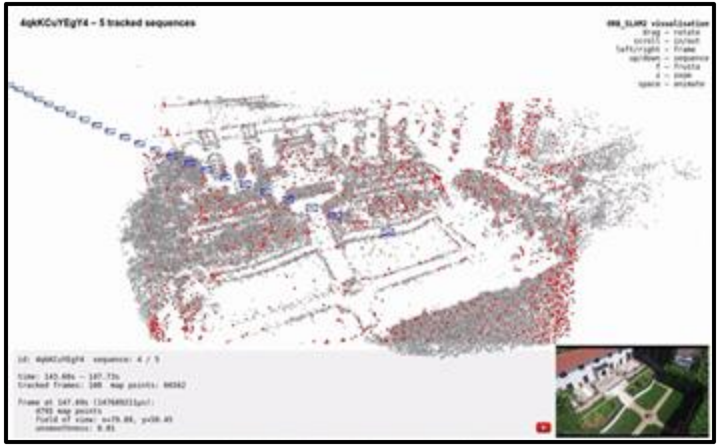


Need massive set of triplets with known camera poses


FOR SALE

RealEstate10K

SLAM

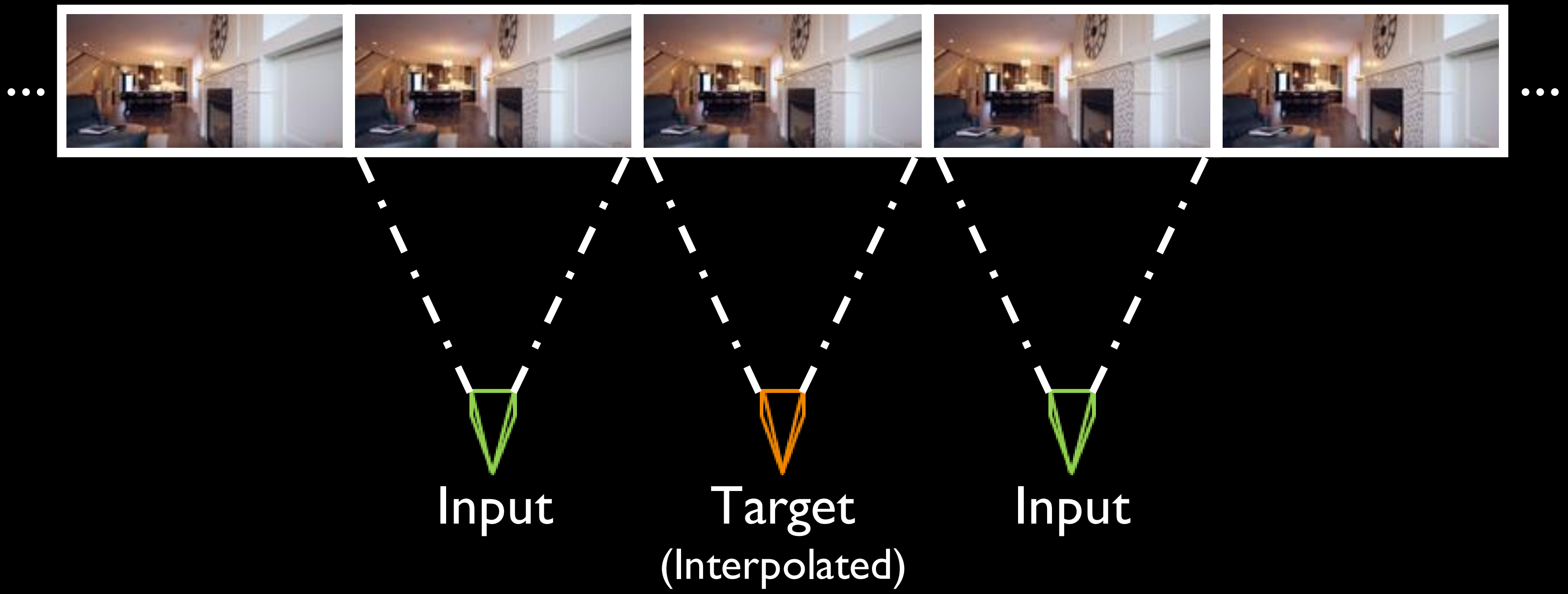


10 million frames from **80,000 video clips** from **10,000 videos**

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

Sampling Training Examples

Sampling Training Examples



Results

Left

Right

Output

Image 1

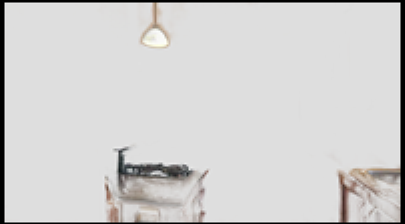

Image 2

Output

Reference input view

Plane 0

Plane 9



Plane 13

Plane 16

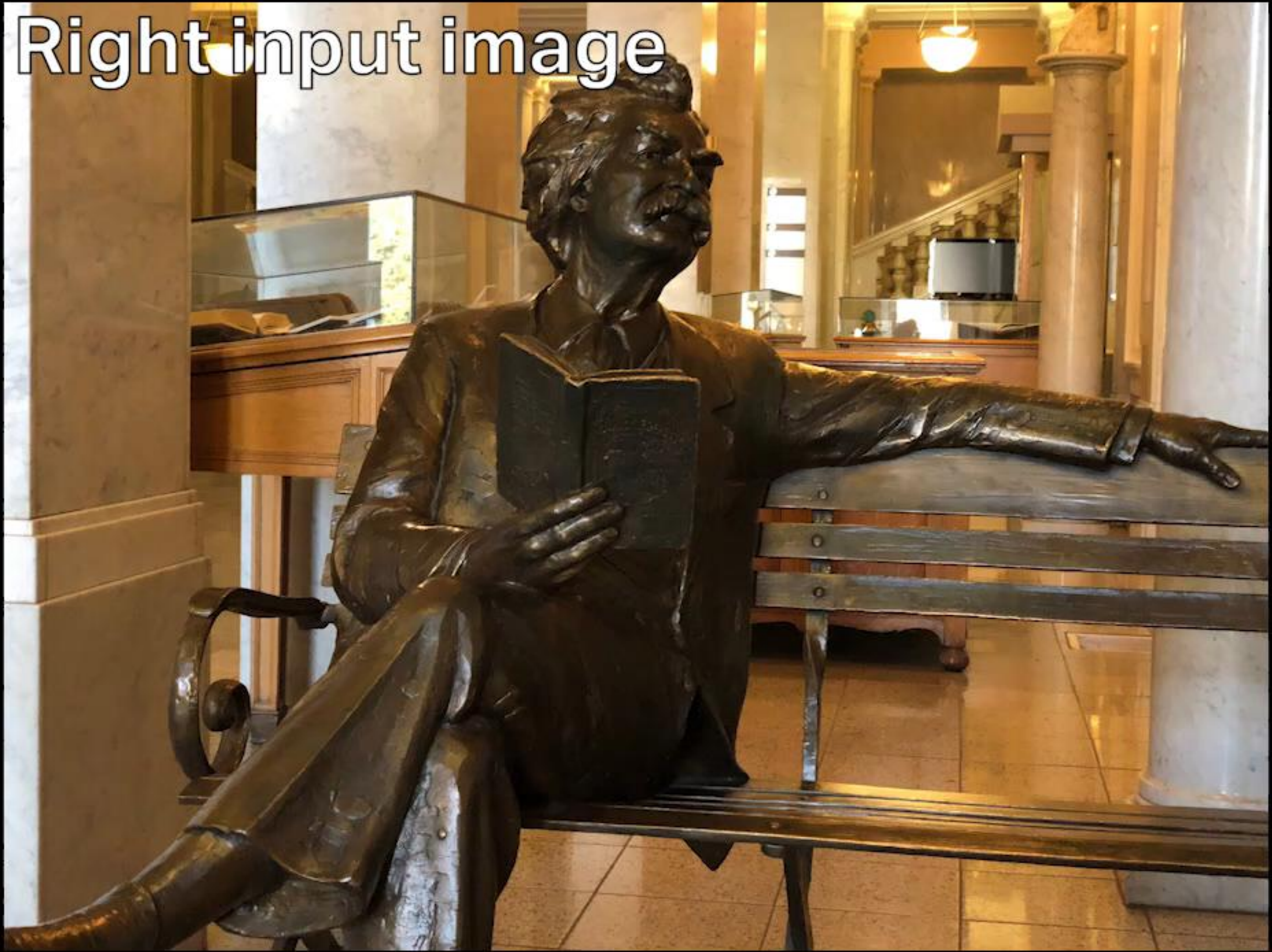
Plane 24

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**

Other recent MPI-based methods

Single-view MPI Prediction

Input images

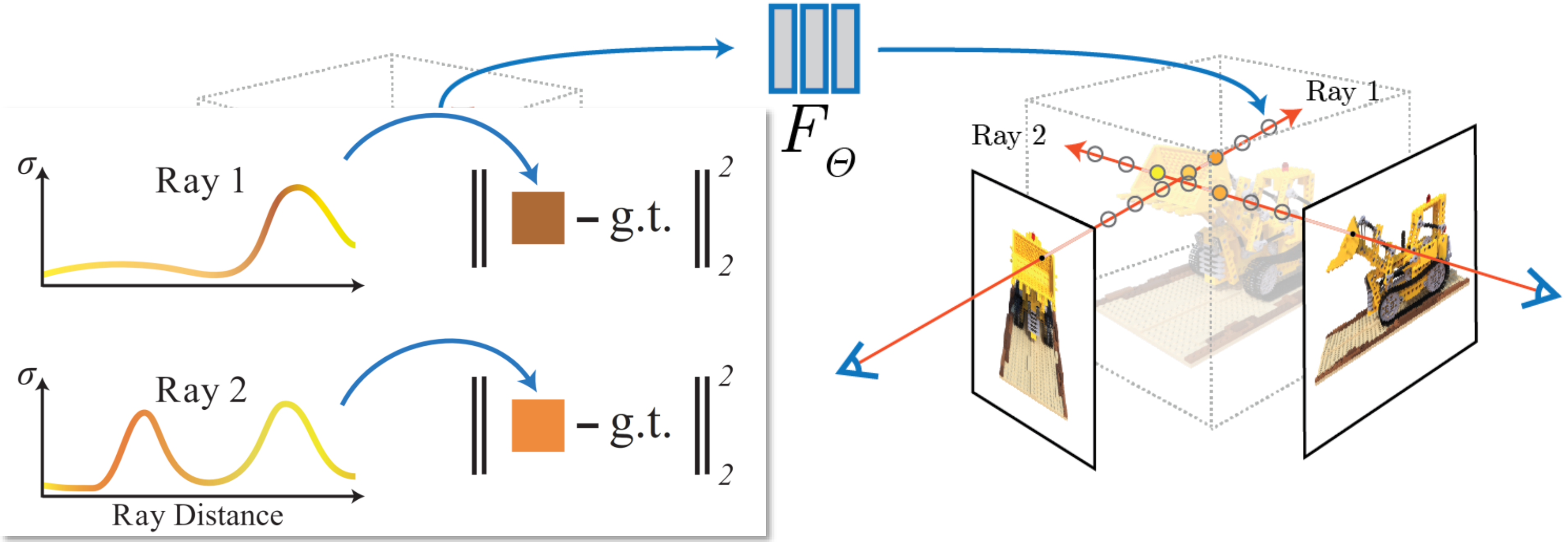
Richard Tucker & Noah Snavely, *Single-View View Synthesis with Multiplane Images*, CVPR 2020

Capturing varying appearance

Zhengqi Li, Wenqi Xian, Abe Davis, Noah Snavely. *Crowdsampling the Plenoptic Function*. ECCV 2020.

MPIs yield artifacts when moving the camera too far

NeRF: Full Neural 3D reconstruction



Ben Mildenhall*, Pratul P. Srinivasan*, Matthew Tancik*, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV 2020. https://www.matthewtancik.com/nerf

NeRF Results

More NeRF results

NeX

Suttisak Wizadwongsa*, Pakkapon Phongthawee*, Jiraphon Yenphraphai*, Supasorn Suwajanakorn. **NeX: Real-time View Synthesis with Neural Basis Expansion**. CVPR 2021.
https://nex-mpi.github.io/

NeRF in the Wild

Ricardo Martin-Brualla*, Noha Radwan*, Mehdi S. M. Sajjadi*, Jonathan T. Barron, Alexey Dosovitskiy, Daniel Duckworth. *NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections*. CVPR 2021. https://nerf-w.github.io/

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

Thanks to the folks behind this work


Richard Tucker


Zhengqi Li


Tinghui Zhou


John Flynn


Graham Fyffe


Tali Dekel


Forrester Cole


Ce Liu


Bill Freeman

Questions?