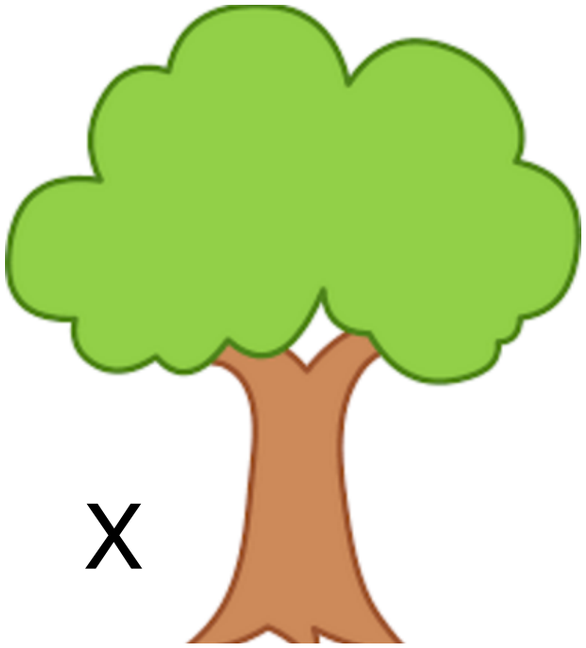
Vision as Inverse Graphics

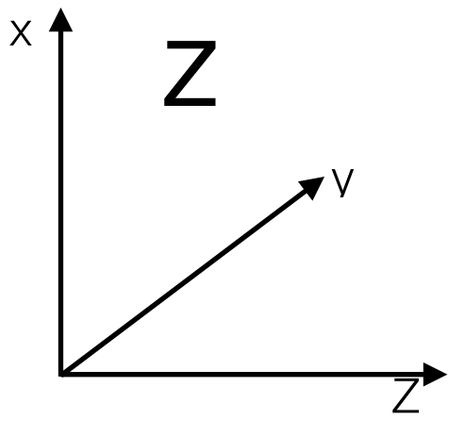
Katerina Fragkiadaki

Machine Learning Department
CMU

MACHINE LEARNING
DEPARTMENT

Understanding the world from images and videos

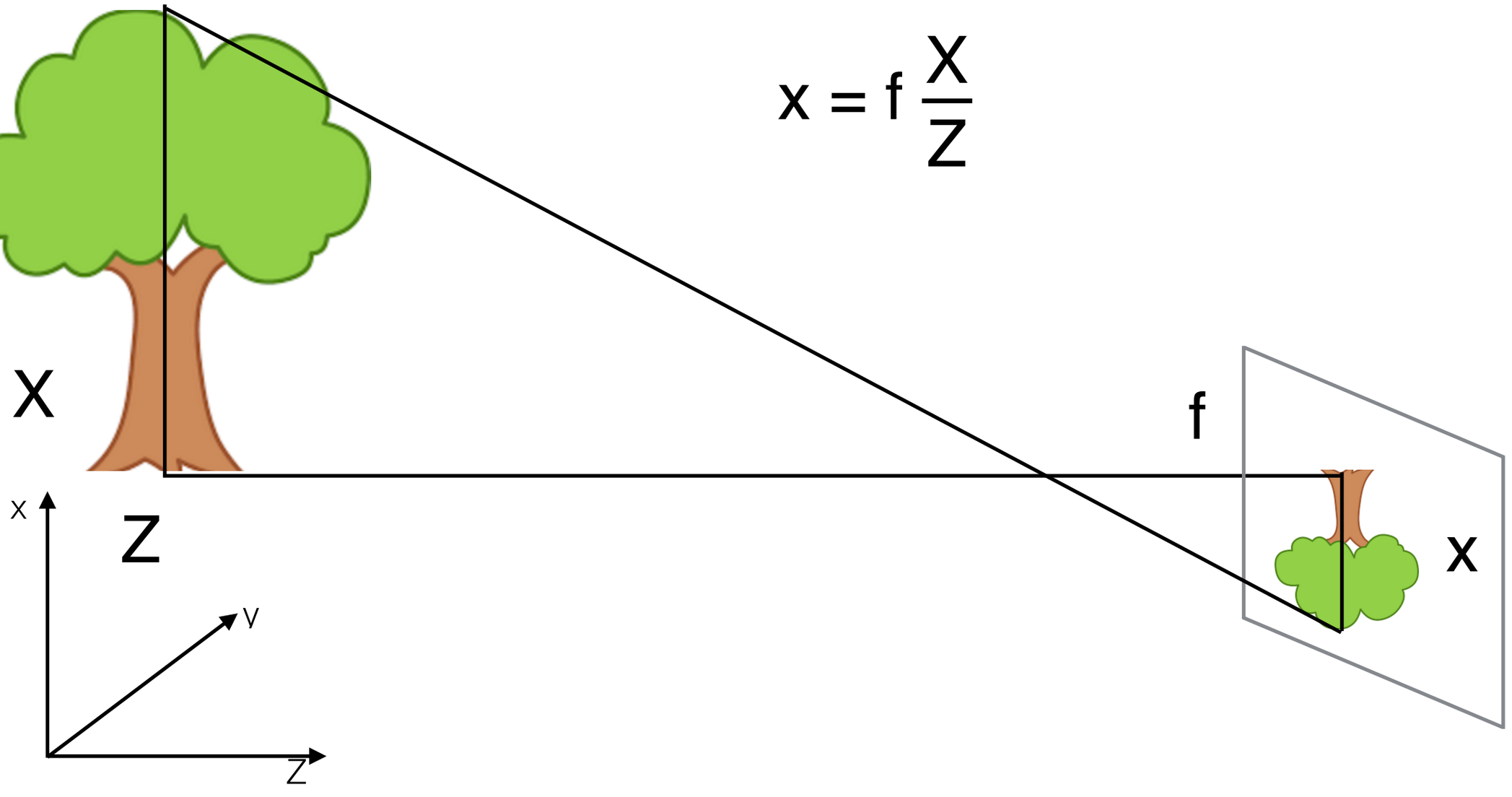




3D

Understanding the world from images and videos

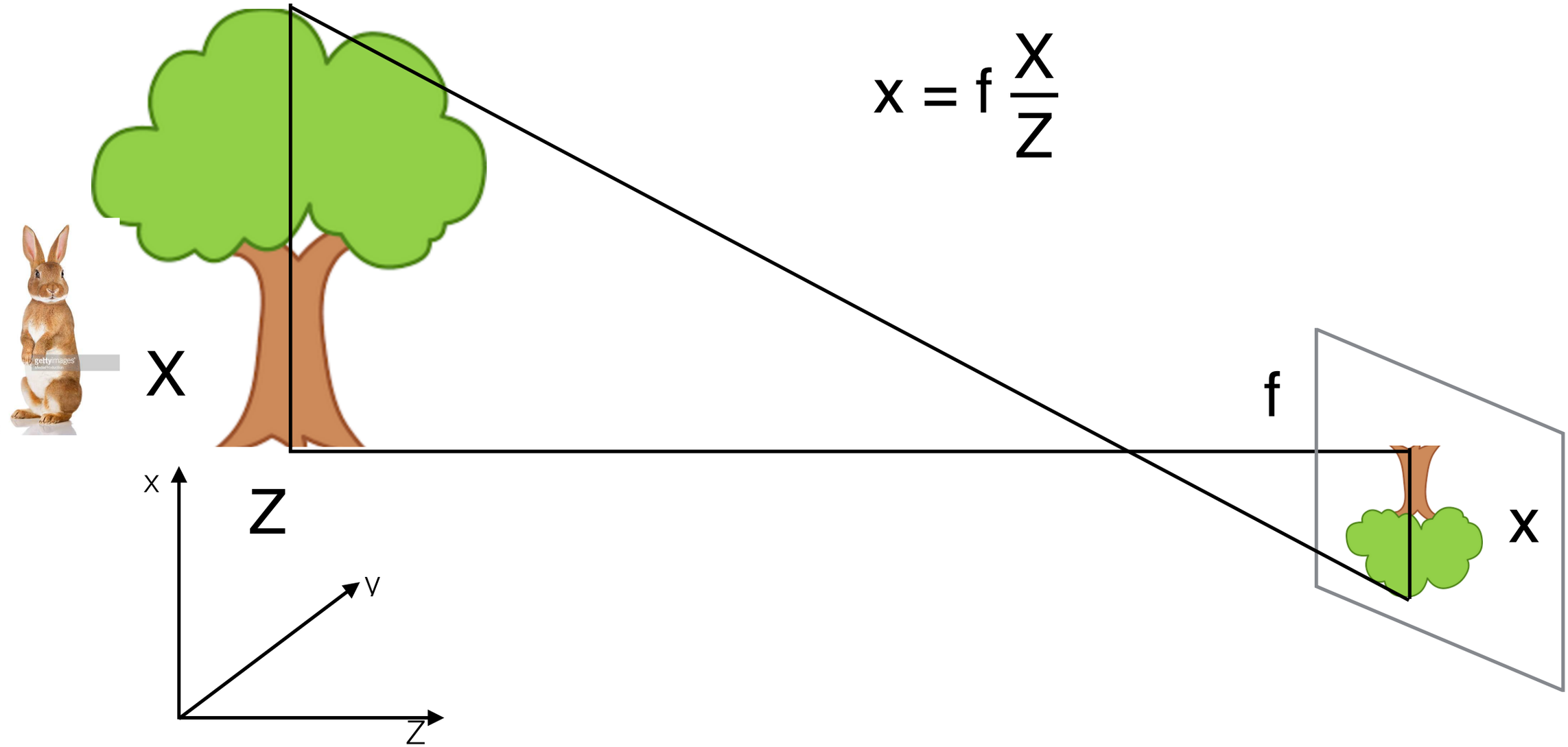
$$x = f \frac{X}{Z}$$



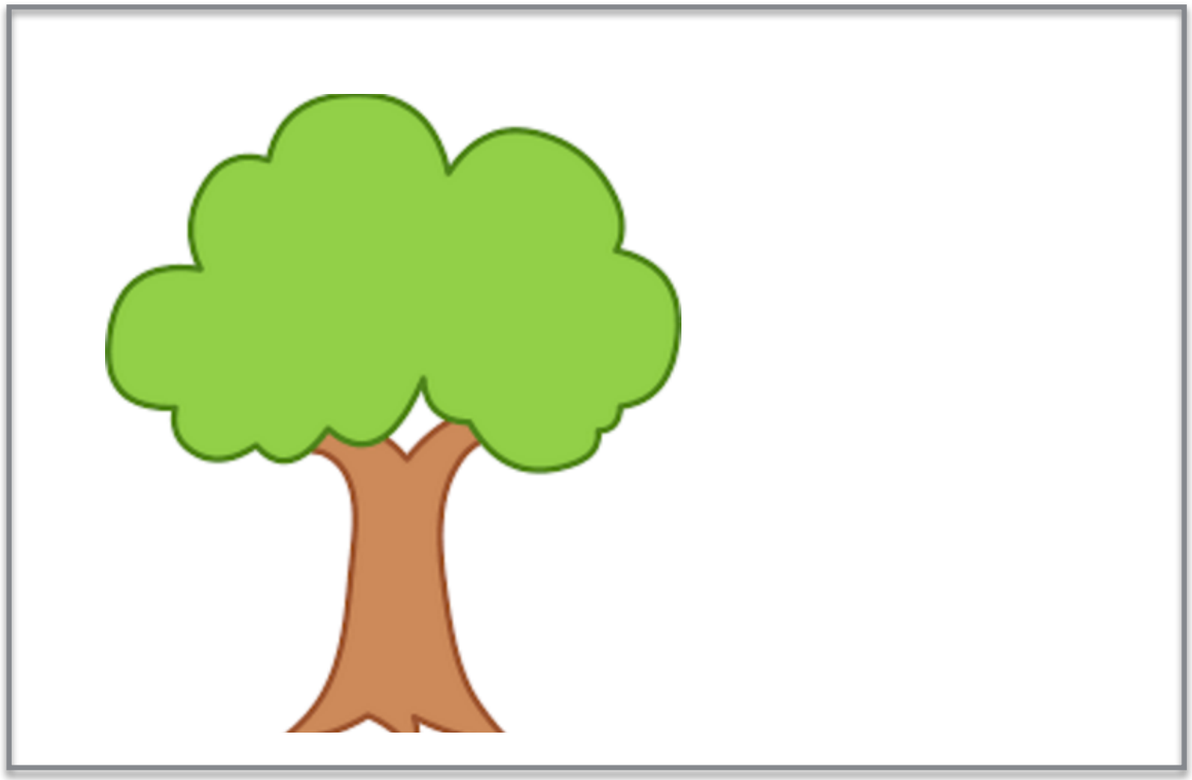
3D

2D

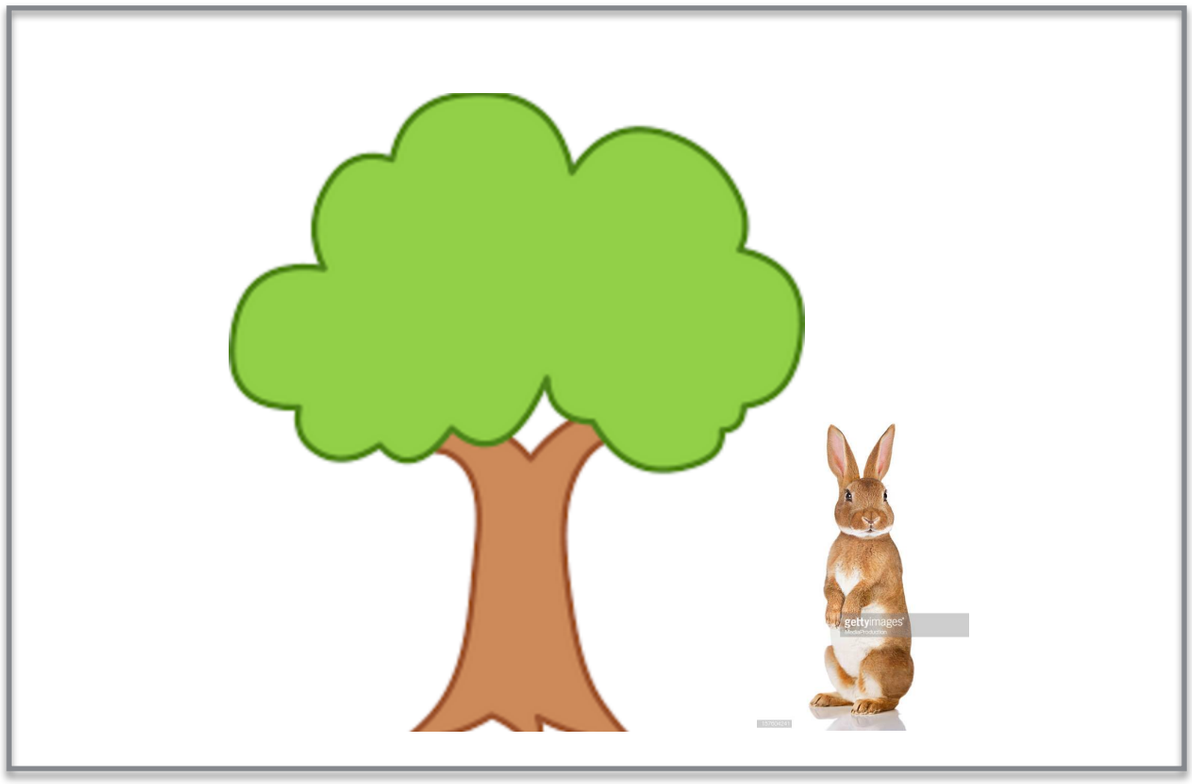
Understanding the world from images and videos



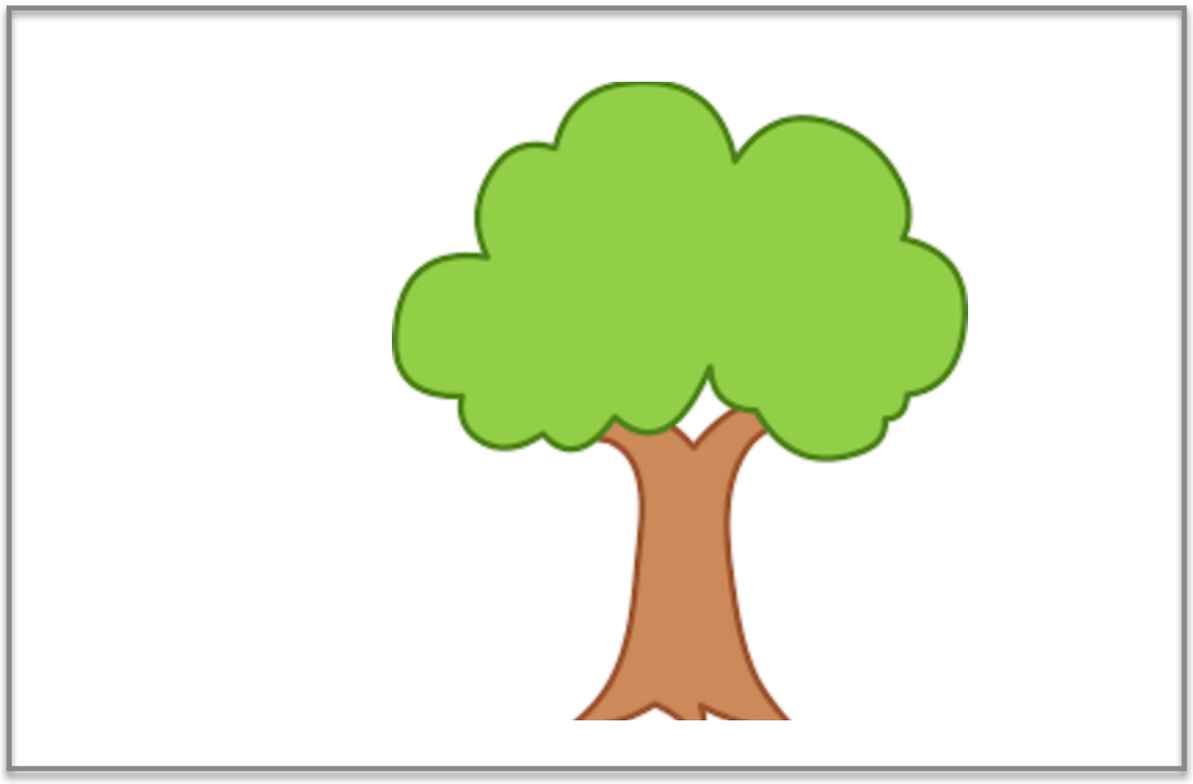
Visuomotor Coordination



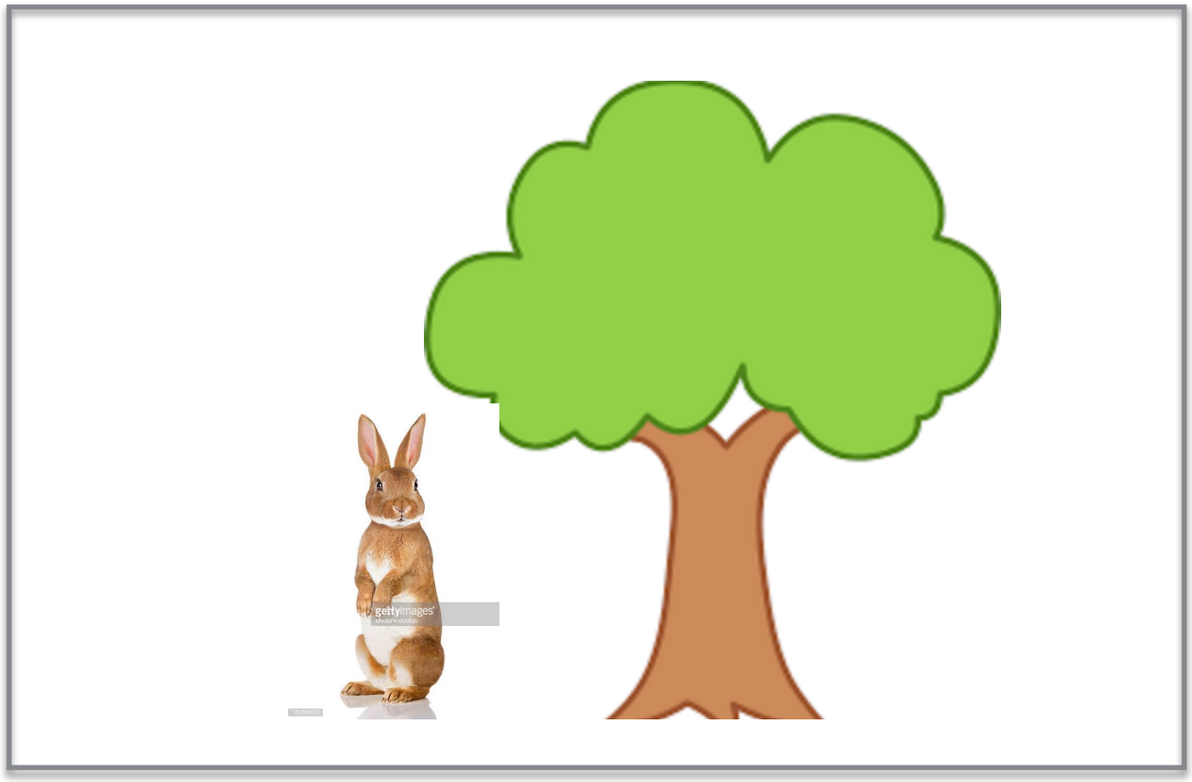
Visuomotor Coordination



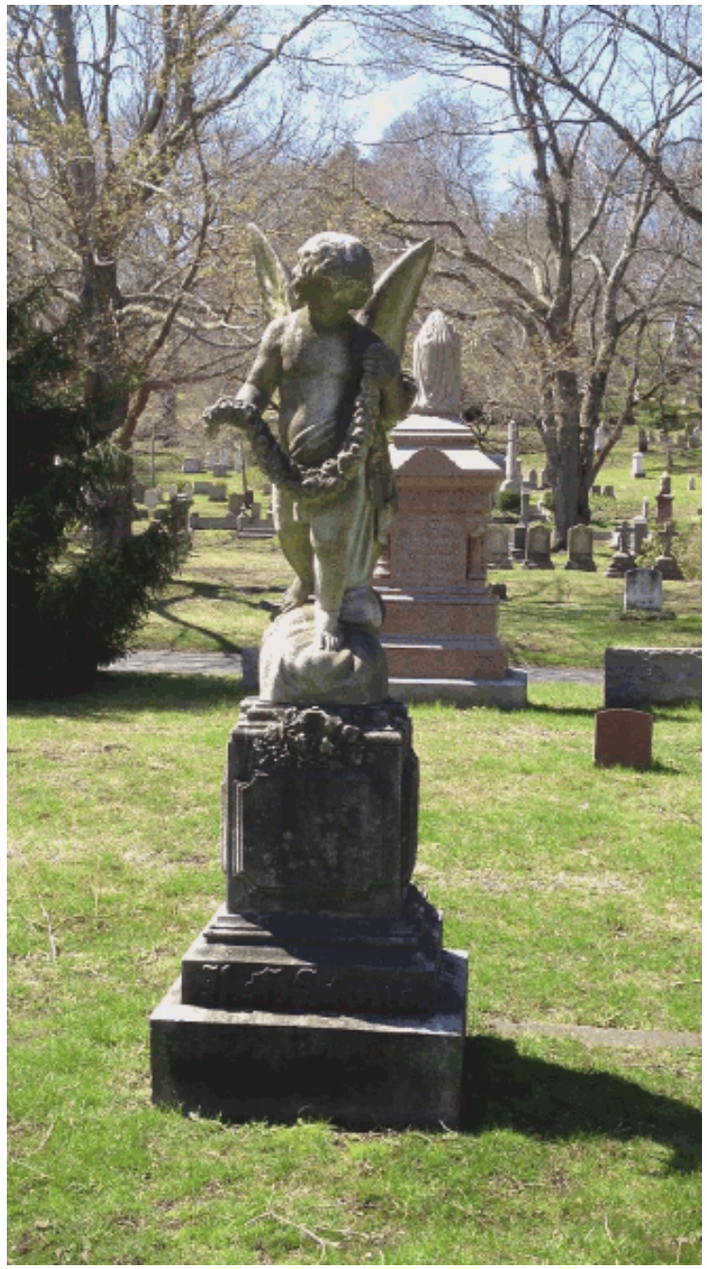
Visuomotor Coordination



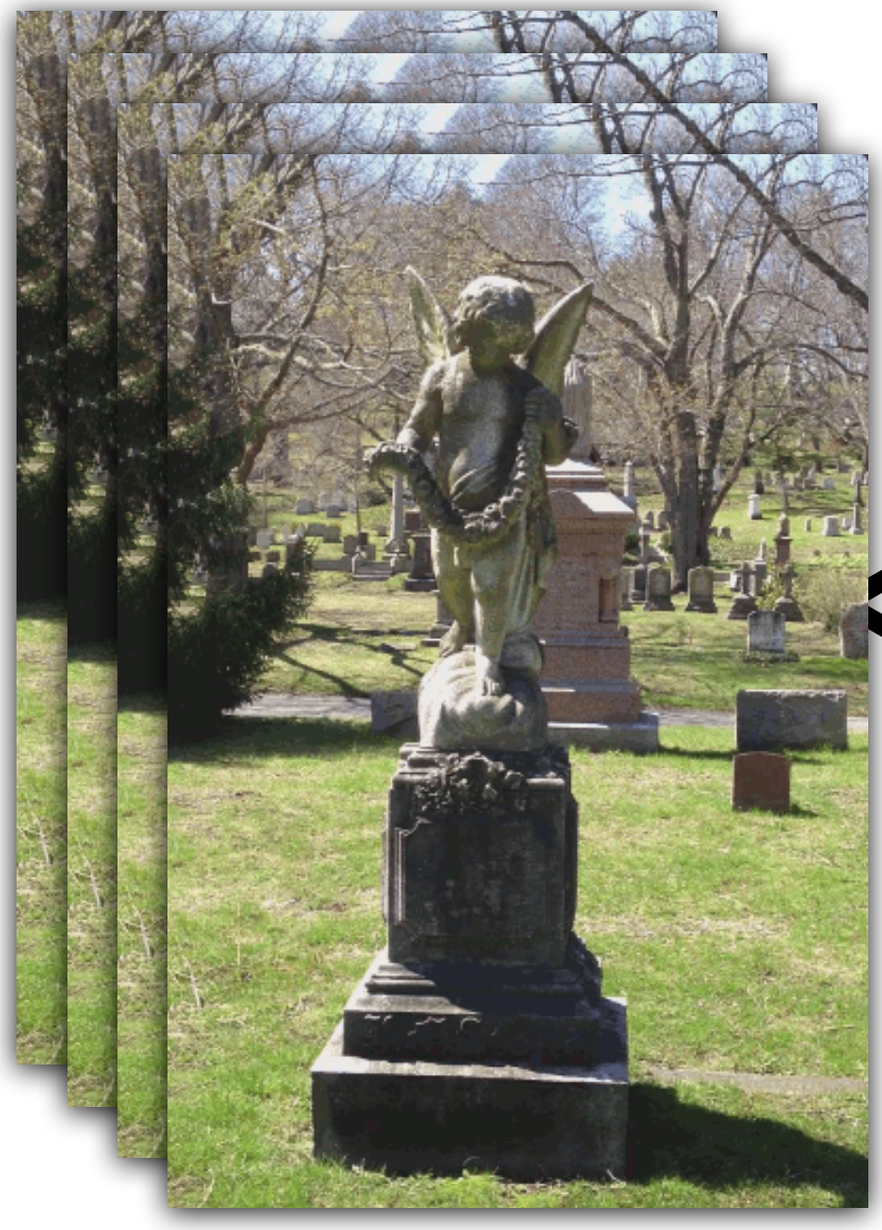
Visuomotor Coordination



Simultaneous Localization and Mapping



Simultaneous Localization and Mapping



3D point cloud

Camera motion

Why Learning in SLAM

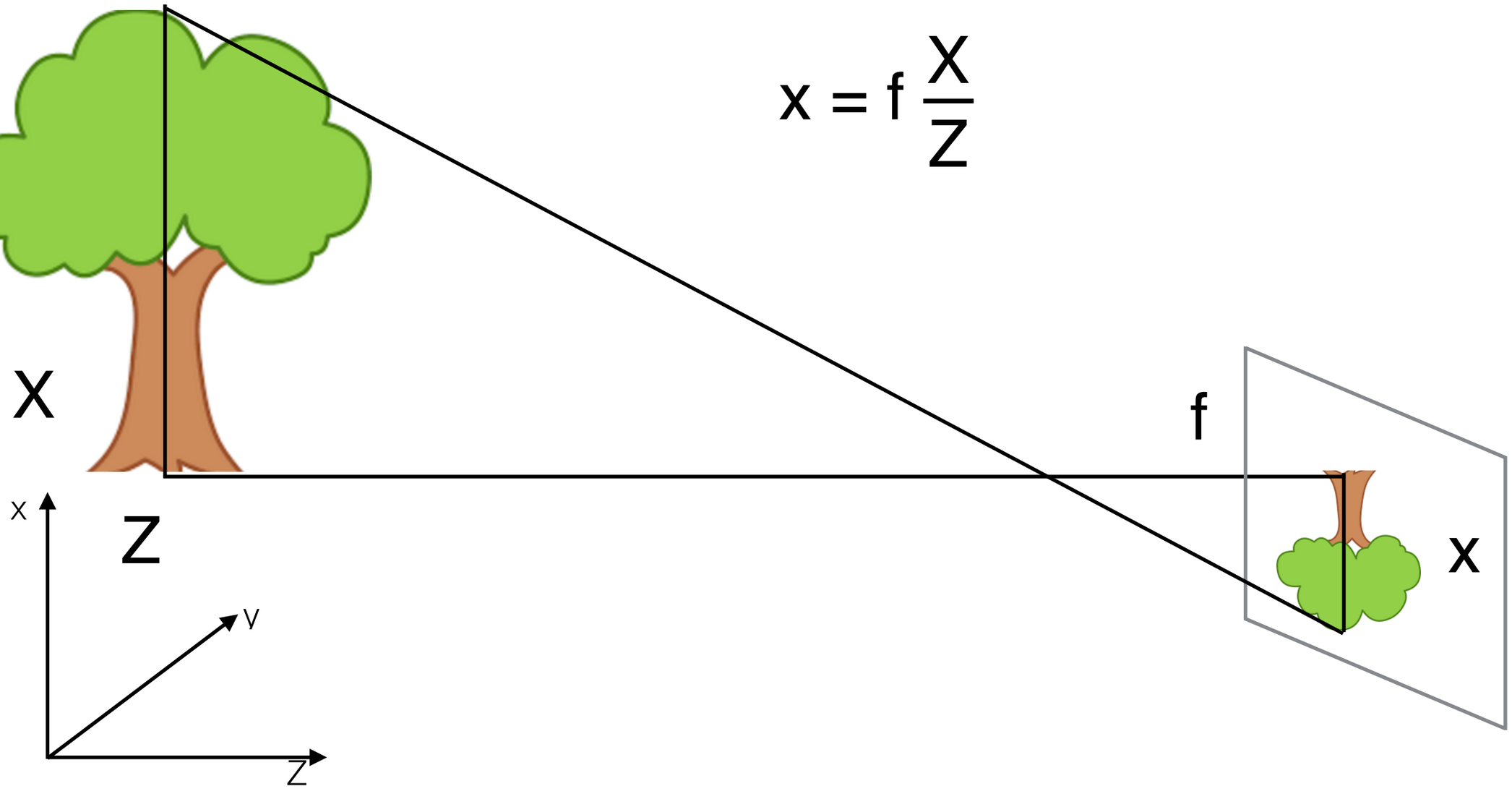
- Scale Ambiguity
- Moving Objects
- Mapping the Invisible
- Geometrically-consistent deep memories for recognition in videos

Why Learning in SLAM

- Scale Ambiguity
- Moving Objects
- Mapping the Invisible
- Geometrically-consistent deep memories for recognition in videos

Scale Ambiguity

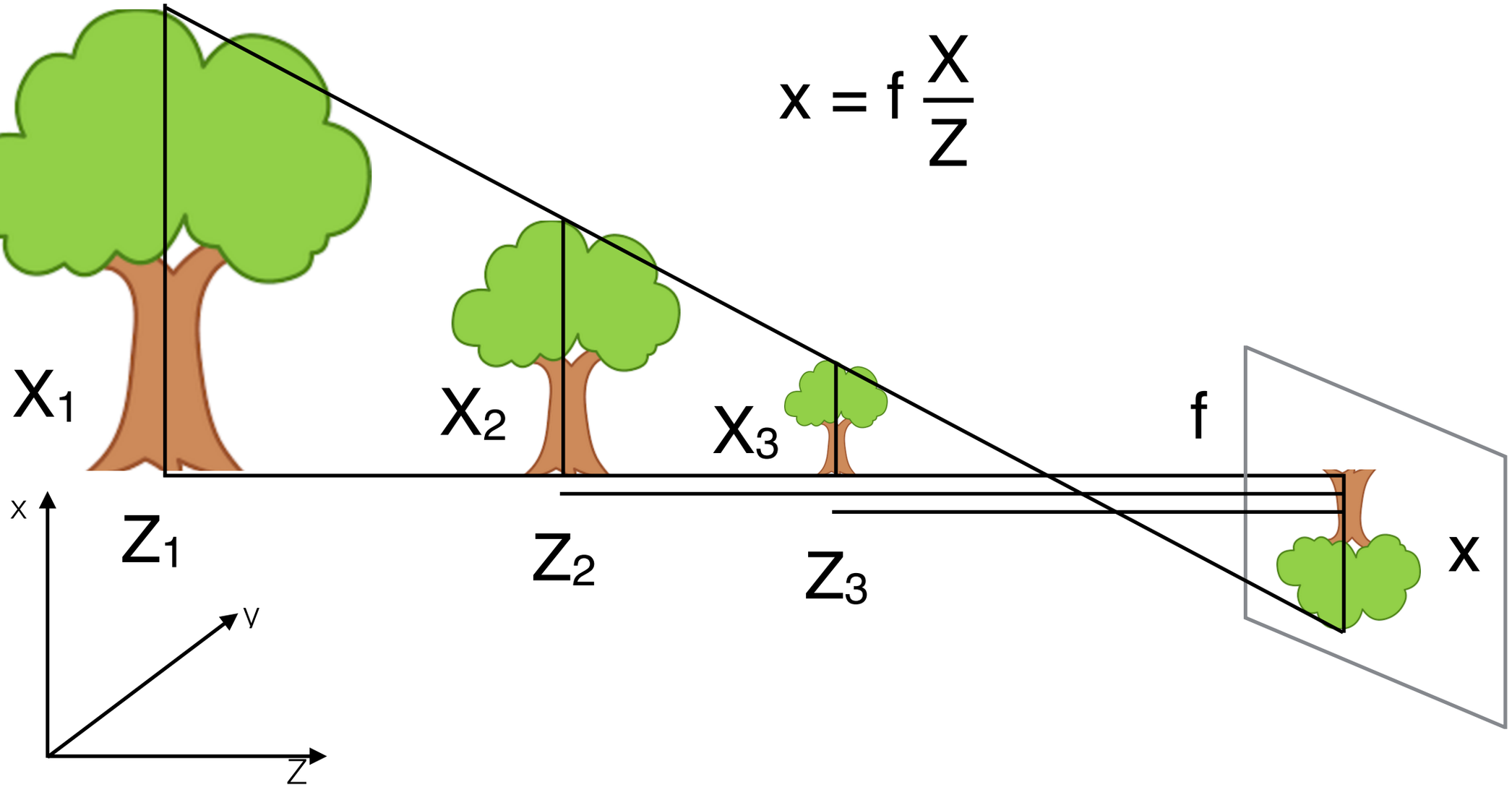
$$x = f \frac{X}{Z}$$



3D **2D**

Scale ambiguity

$$x = f \frac{X}{Z}$$



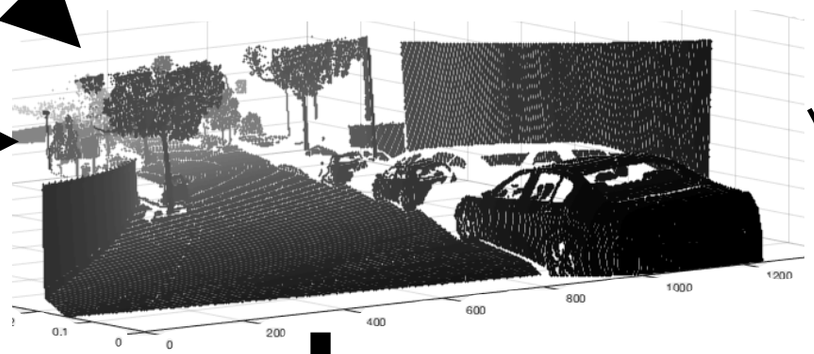
3D **2D**

Depth Net

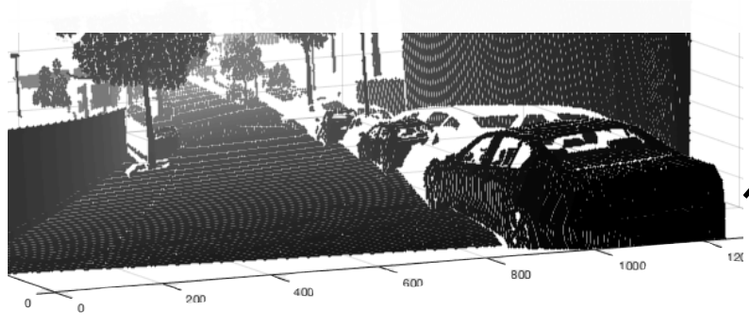
Depth map

3D point cloud

focal f

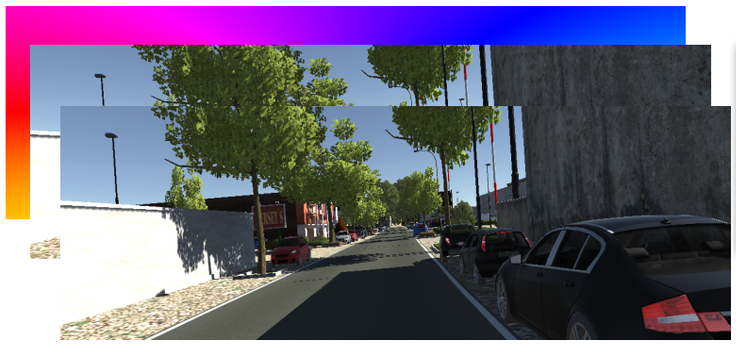


Transformed 3D point cloud



camera projection

flow field



EgoNet

[R, t]

photometric loss

Depth map

Depth Net

3D point cloud

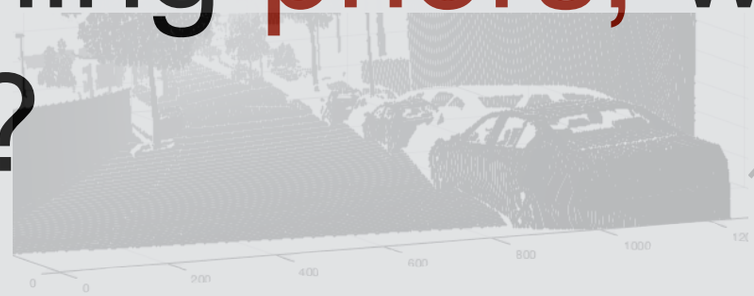
focal f

We can regularize depth and ego motion using priors, e.g., spatial/temporal smoothness

What if, instead of designing priors, we learn them?

EgoNet

[R, t]



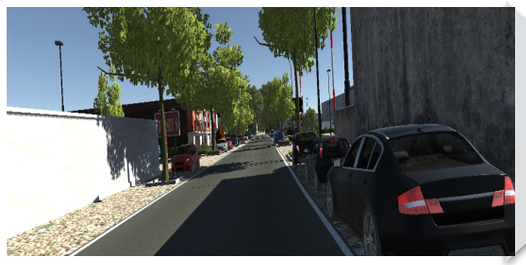
Transformed 3D point cloud

camera projection

flow field

photometric loss

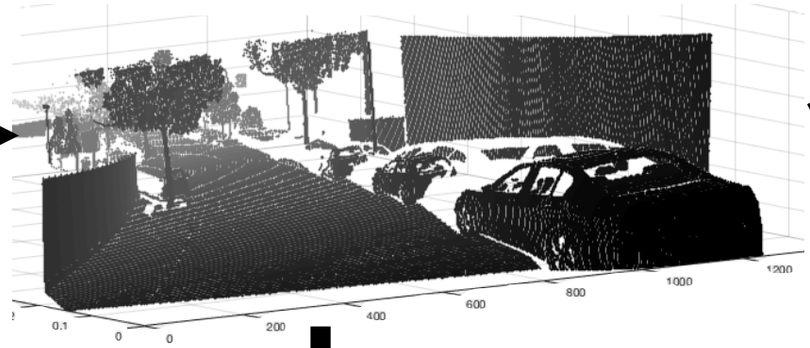
Depth map



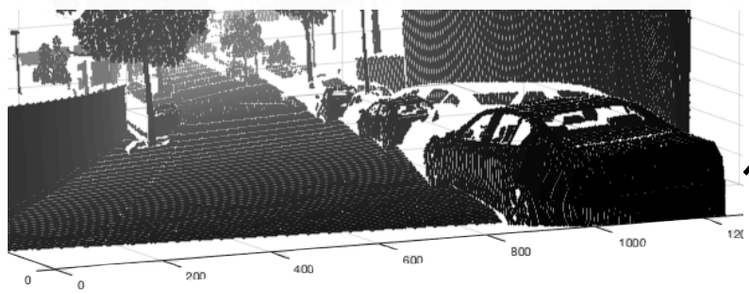
DepthNet

3D point cloud

focal f

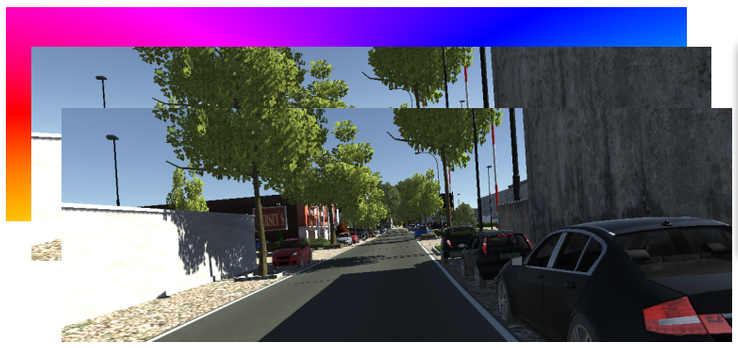


transformed 3D point cloud



camera projection

flow field

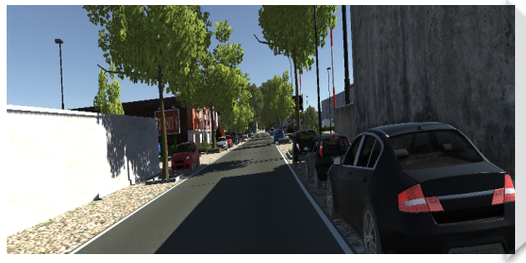


EgoNet

[R, t]

photometric loss

Depth map



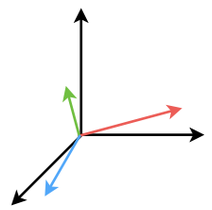
DepthNet

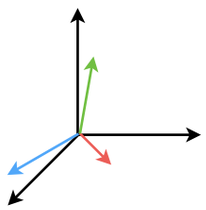
DepthDiscr

Depths

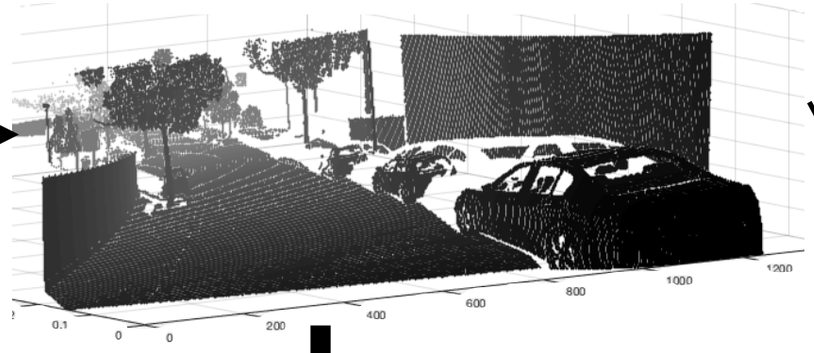
3D point cloud

Camera motions



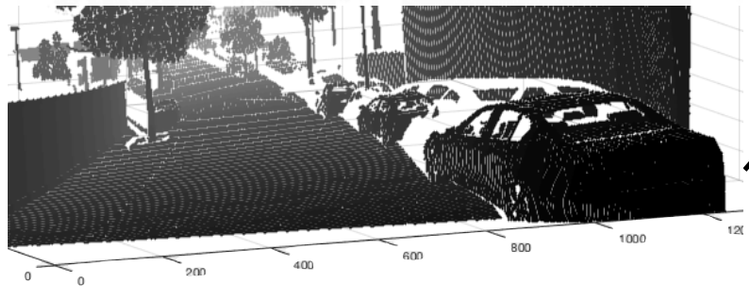


focal f

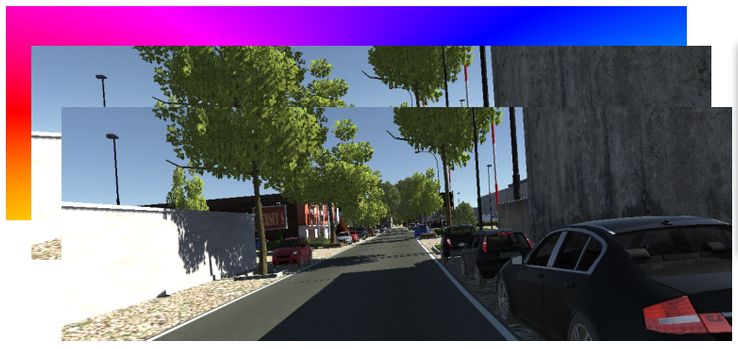


EgoDiscr

transformed 3D point cloud



camera projection



EgoNet

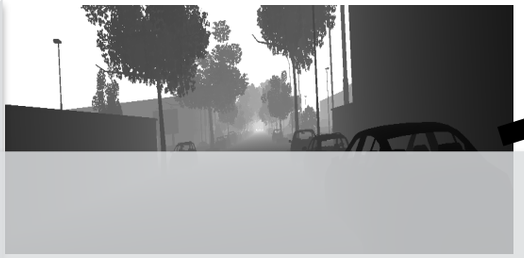
[R, t]

flow field

photometric loss

Depth map

DepthNet

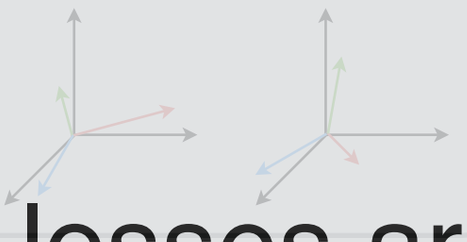


DepthDiscr

Depths

3D point cloud

Camera motions



focal f

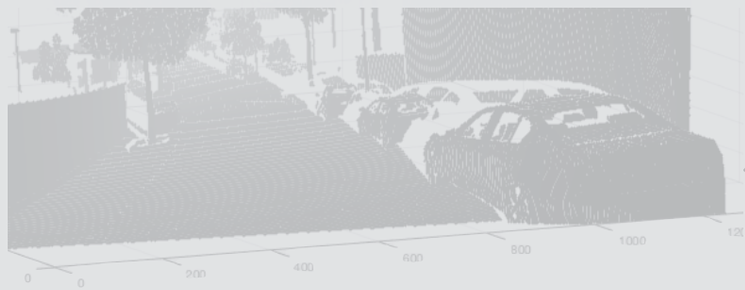
EgoDiscr

transformed 3D point cloud

All losses are unsupervised and can be applied at test time

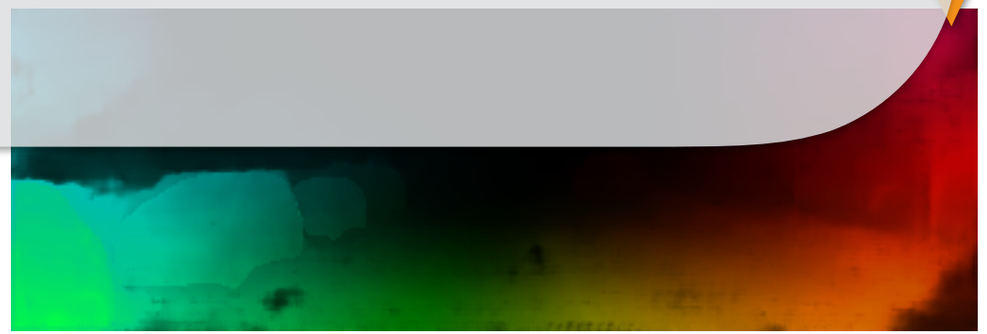
EgoNet

[R, t]



camera projection

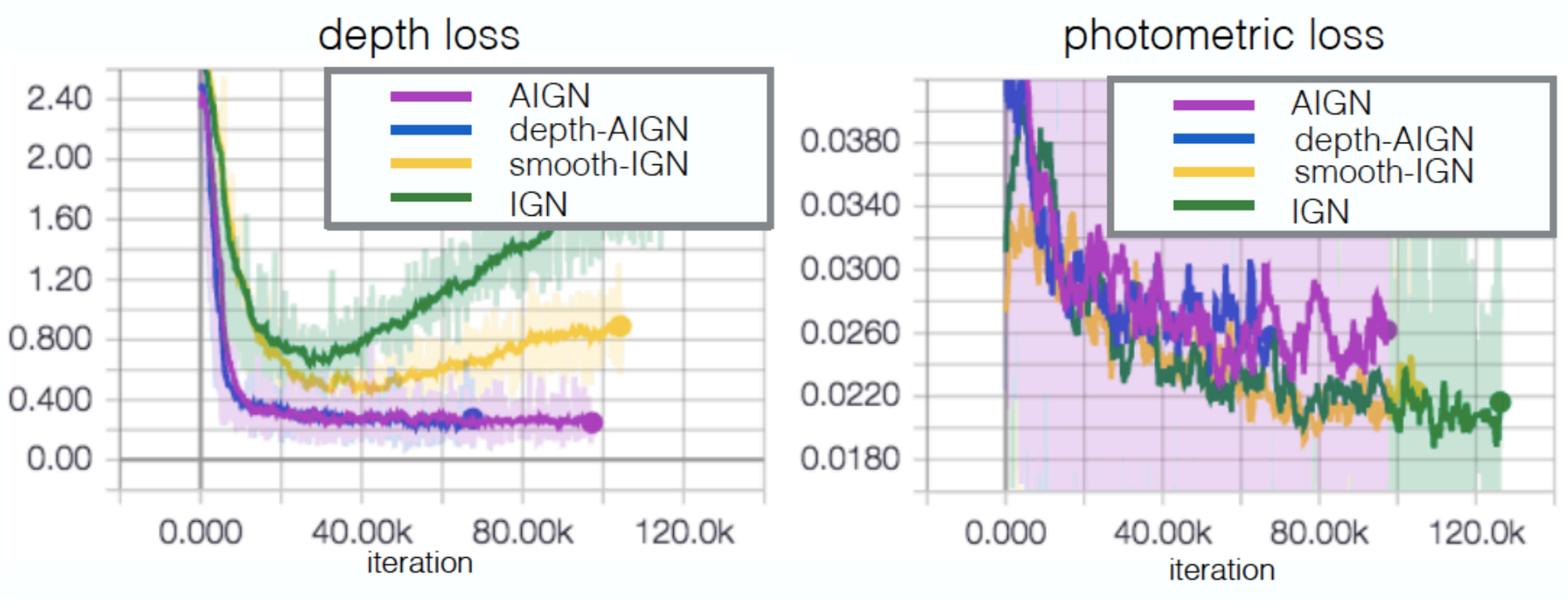
flow field



photometric loss

Adversarial Inverse Graphics Networks

Tung et. al.,ICCV 2017

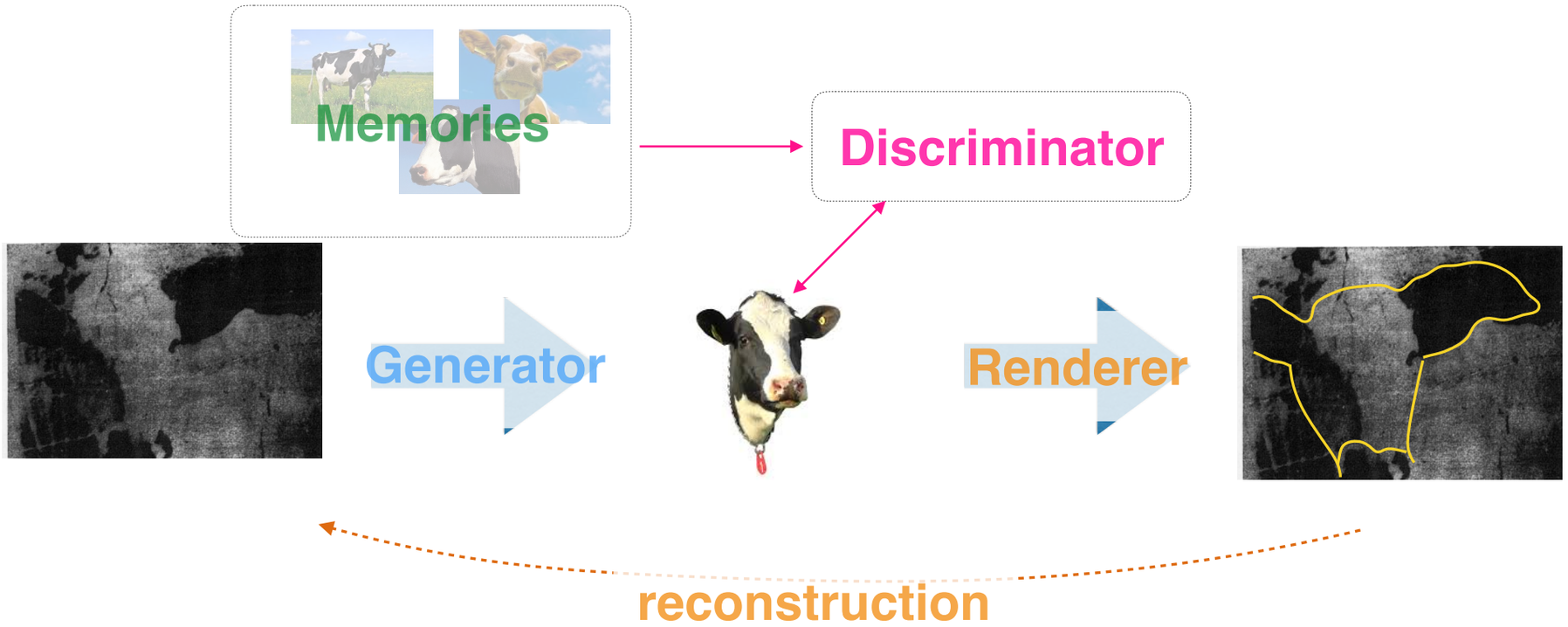


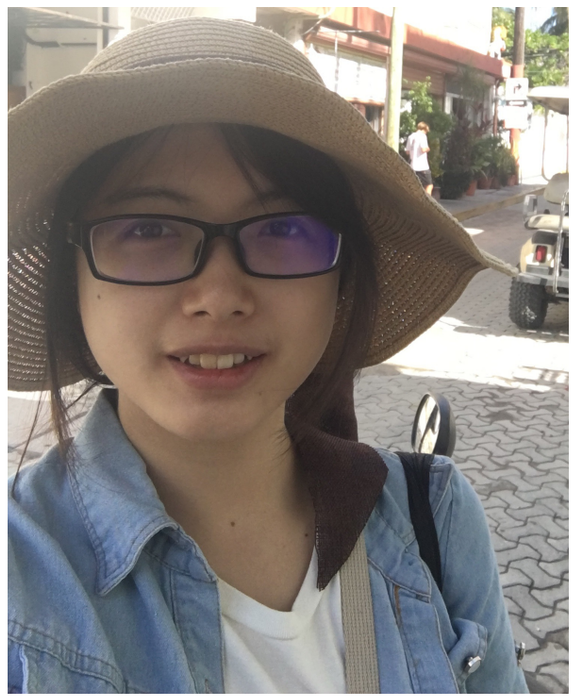
Fish Tung

- Scale Ambiguity causes Drifting
- Adversarial priors stabilize training
- Recovered depth is metric

Adversarial Inverse Graphics Networks

Tung et. al.,ICCV 2017

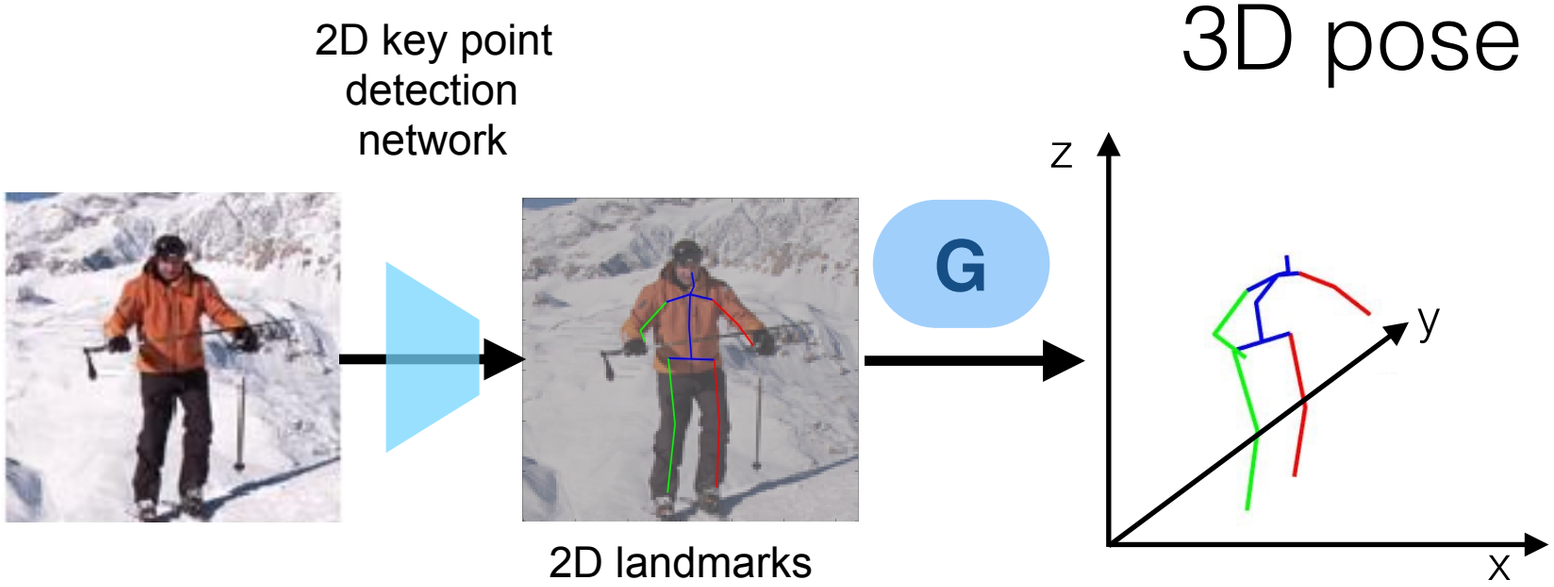




Fish Tung

- Parameter-free decoder (renderer)
- Discriminators on the predicted parameters
- Reprojection losses

Open Pose

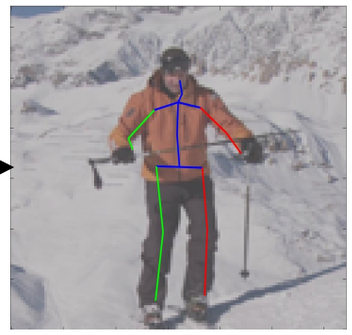


https://github.com/CMU-Perceptual-Computing-Lab/openpose

Single Image 3D Interpreter Network

Wu et al. 2016

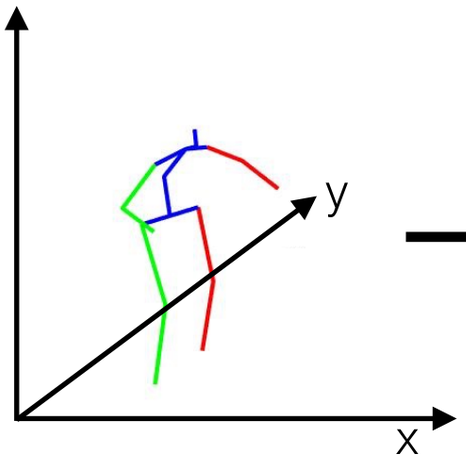
2D key point detection network



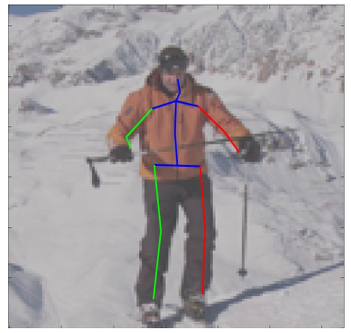
2D landmarks

G

3D pose

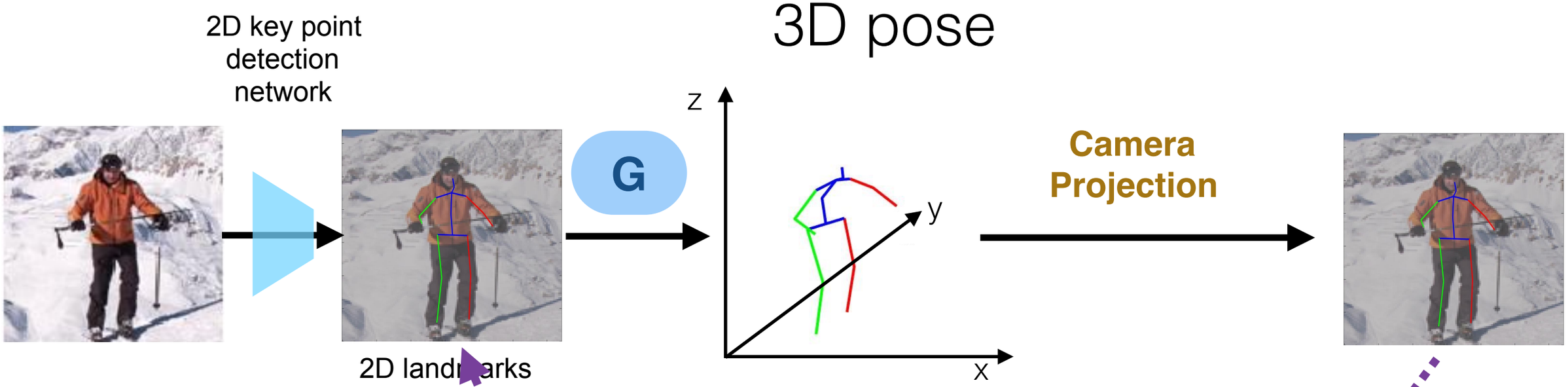


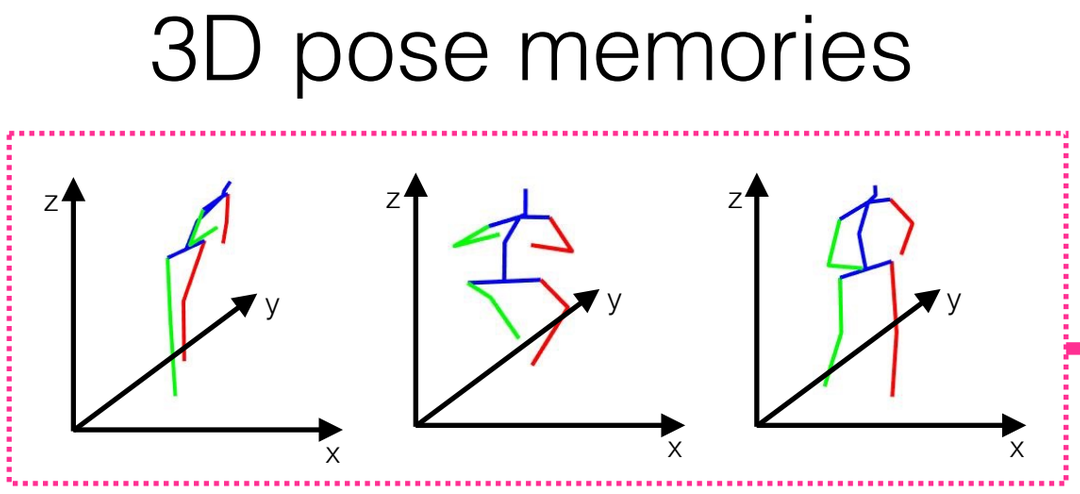
Camera Projection



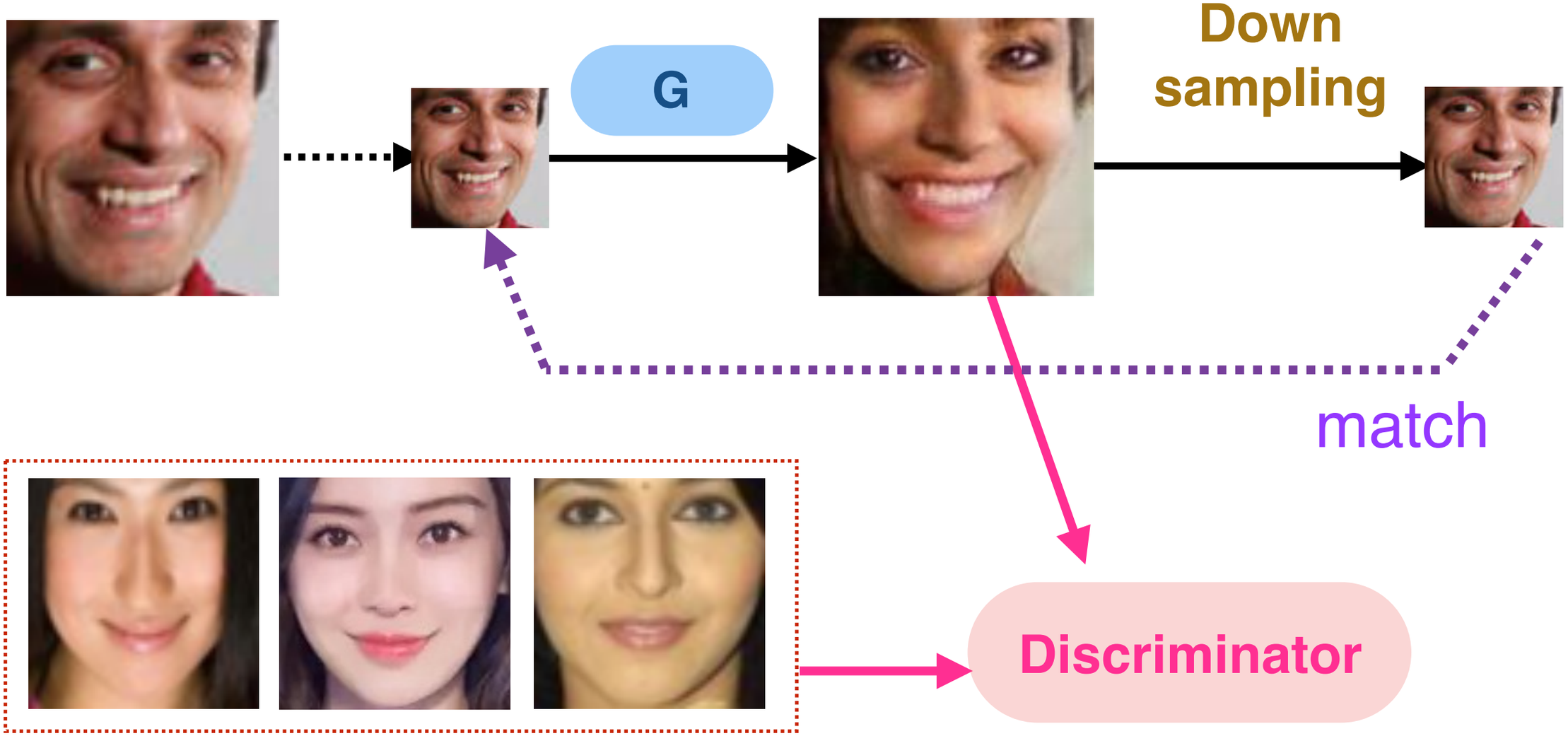
match

Adversarial Inverse Graphics Networks

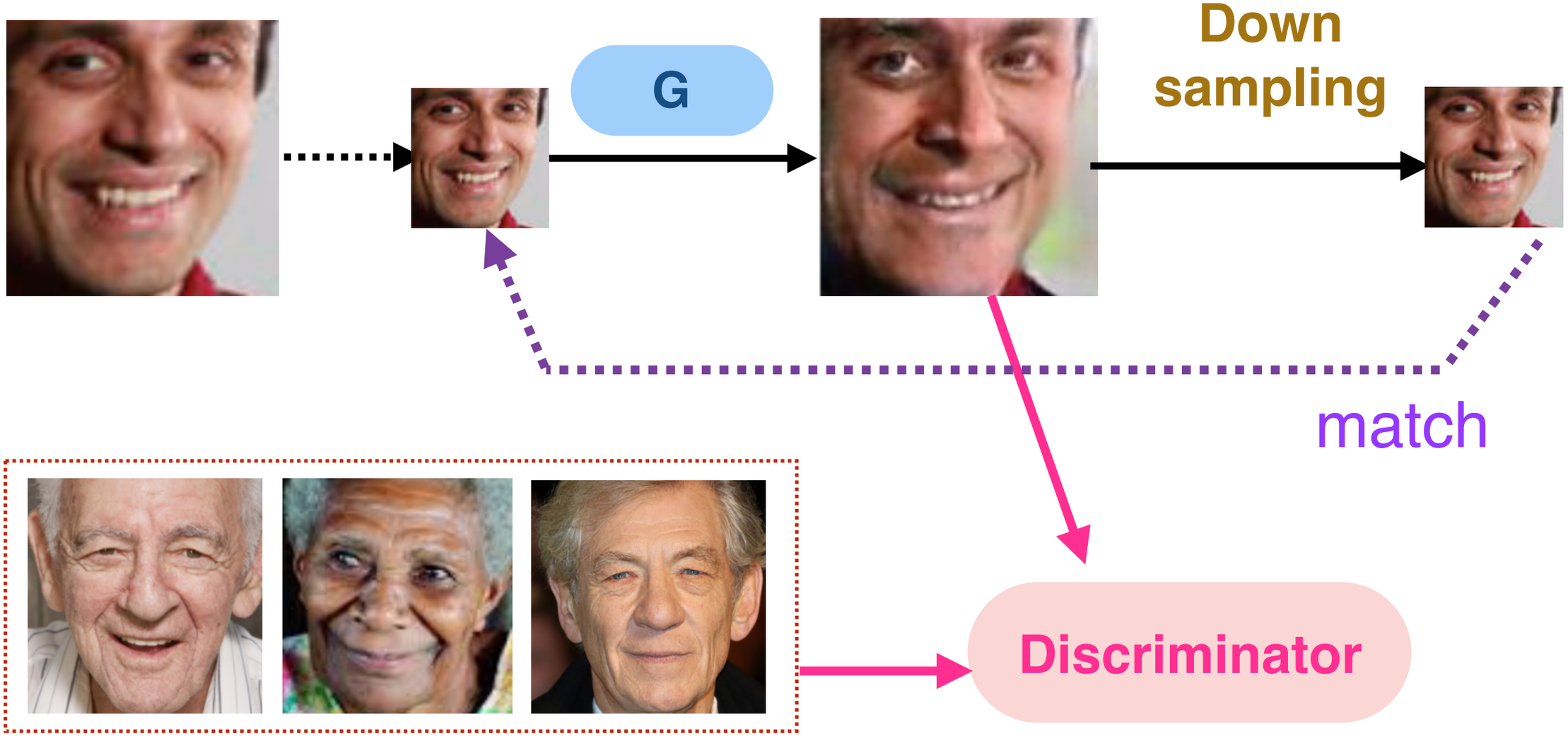




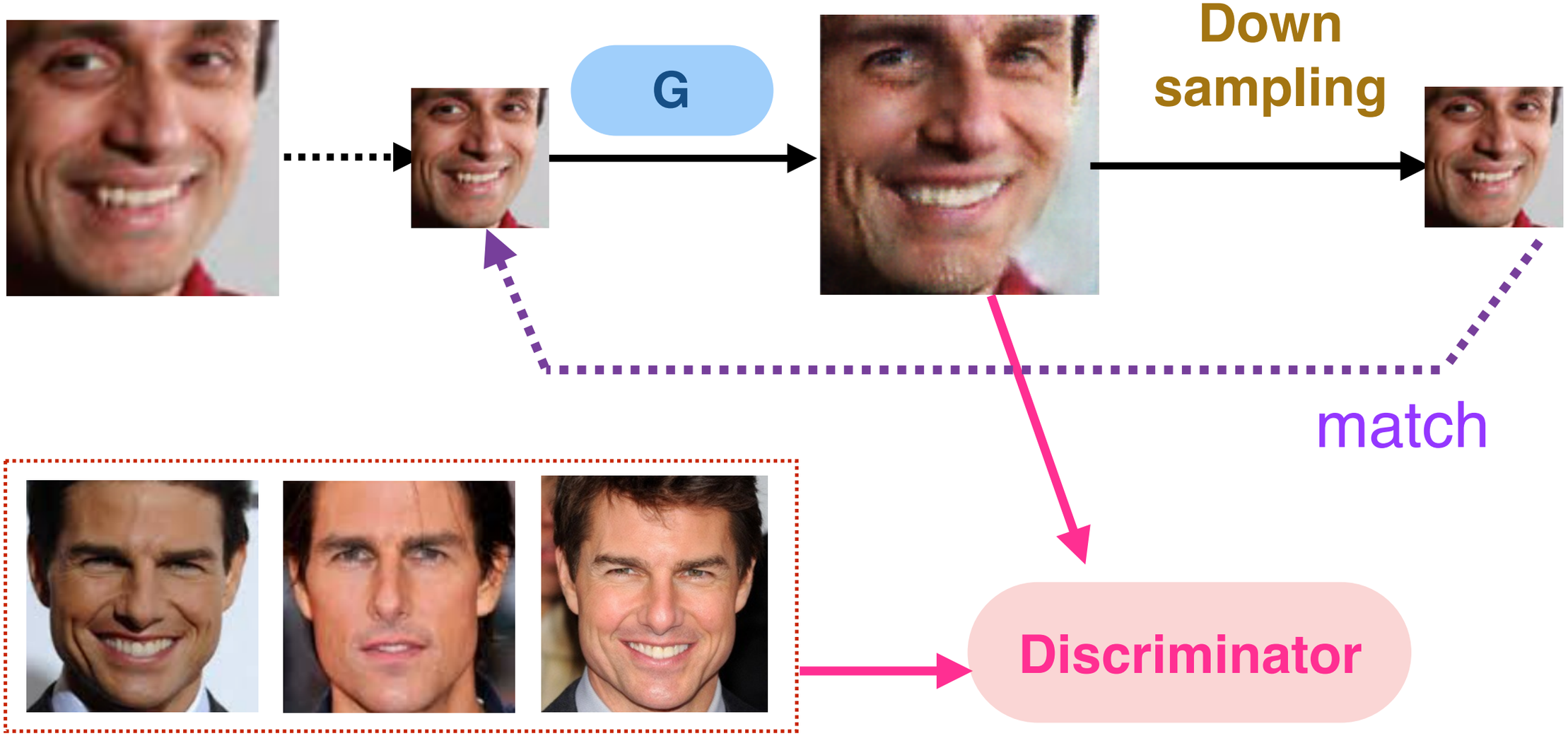
Adversarial Inverse Graphics Networks



Adversarial Inverse Graphics Networks



Adversarial Inverse Graphics Networks



Adversarial Inverse Graphics Networks

Adversarial Inverse Graphics Networks

Why Learning in SLAM

- Scale Ambiguity
- Moving Objects
- Mapping the Invisible
- Geometrically-consistent deep memories for recognition in videos

Input: RGB video

Input: RGB video

Outputs:

- depth
- egomotion

Reprojection loss is not correct on independently moving objects!
Moving objects are treated as noise

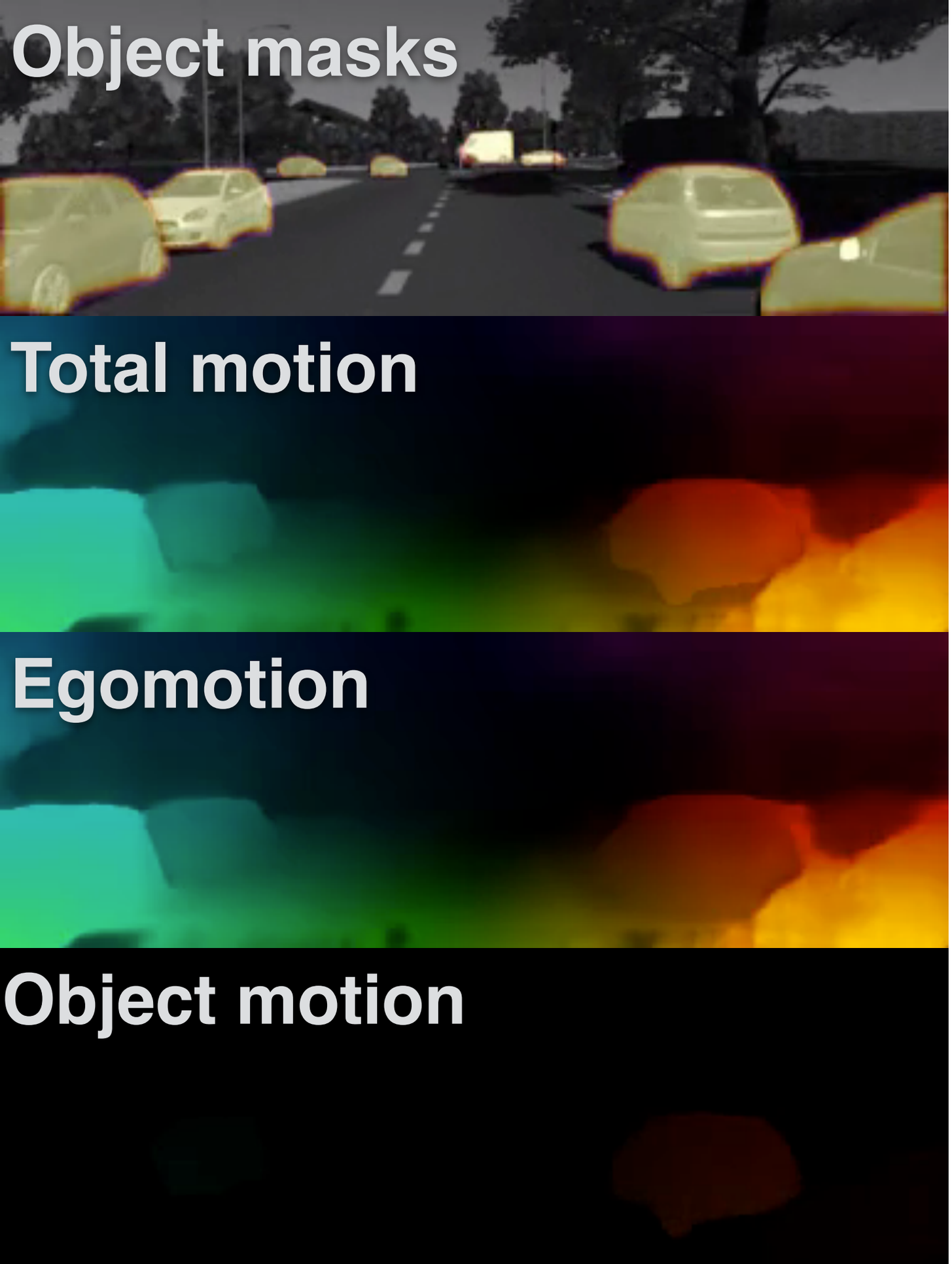
Input: RGB video

Outputs:

- depth
- egomotion

objects

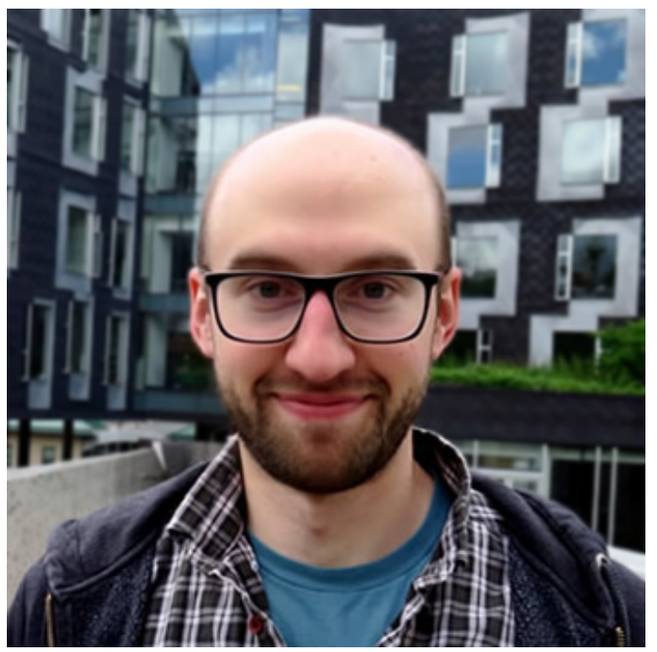
- 6D motion
- segmentation



Decompose total motion into egomotion and object motion.

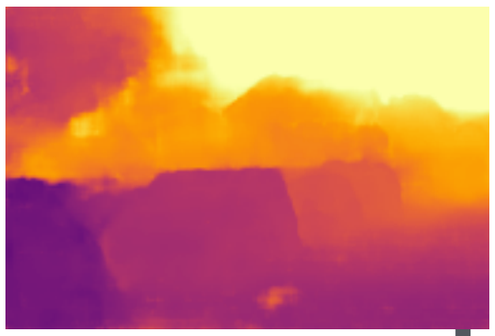
SfM-Net: Learning of Structure and Motion from Video, Vijayanarasimhan et al. 2017

Object-centric Neural Structure-from-Motion, in submission



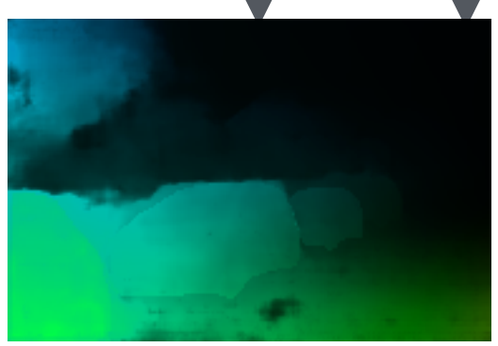
Adam Harley

Depth Net



EgoNet

[R, t]

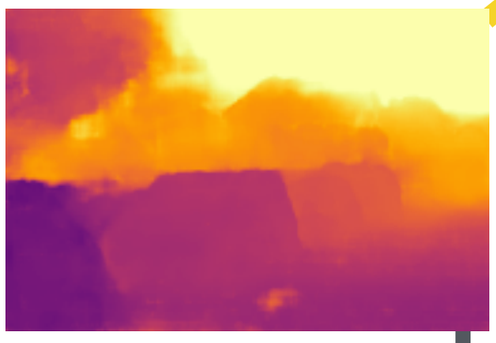


Object crops

Region Proposal Network

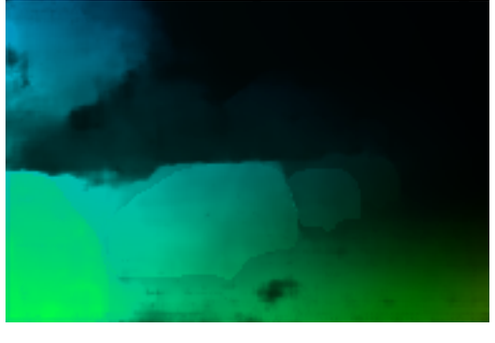
RPN

Depth Net



EgoNet

[R, t]



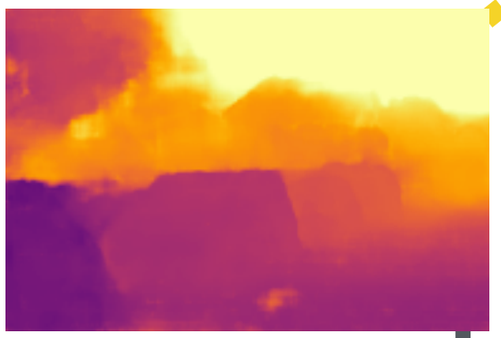
⋮

Egomotion-stabilized object crops

Region Proposal Network

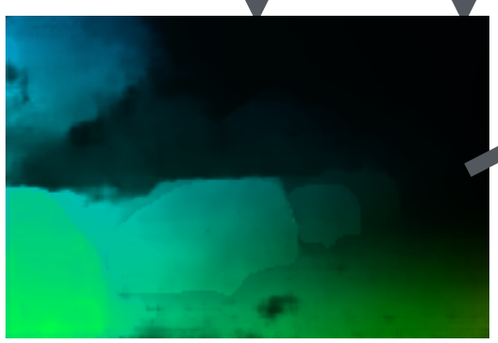
RPN

Depth Net



EgoNet

[R, t]



Egomotion-stabilized object crops

Object RT Net

[R, t]

"Residual" motion

MaskNet

Where the motion applies

Egomotion-stabilized object crops

Object RT Net

[R, t]

"Residual" motion

MaskNet

Where the motion applies

Add moving objects to the reprojection

XYZ2 = RTobj RTego XYZ1

Egomotion-stabilized object crops

Object RT Net

[R, t]

MaskNet

Class Net

class

Egomotion-stabilized object crops

Object RT Net → [R, t]

MaskNet

Class Net → class

Scale memory

Object categories have typical sizes.

Egomotion-stabilized object crops

Object RT Net

[R, t]

MaskNet

Class Net

class

XYZ Net

XYZ

Scale memory

Ego-stabilized object crops

Object RT Net

[R, t]

MaskNet

Class Net

class

Scale memory

DepthNet crop

XYZ Net

XYZ

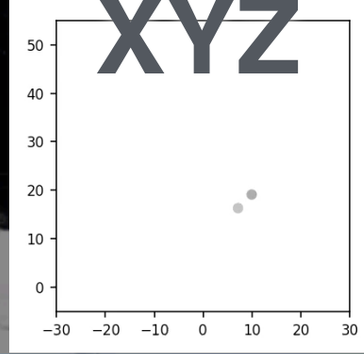
Objects provide scale supervision

scene-object depth consistency

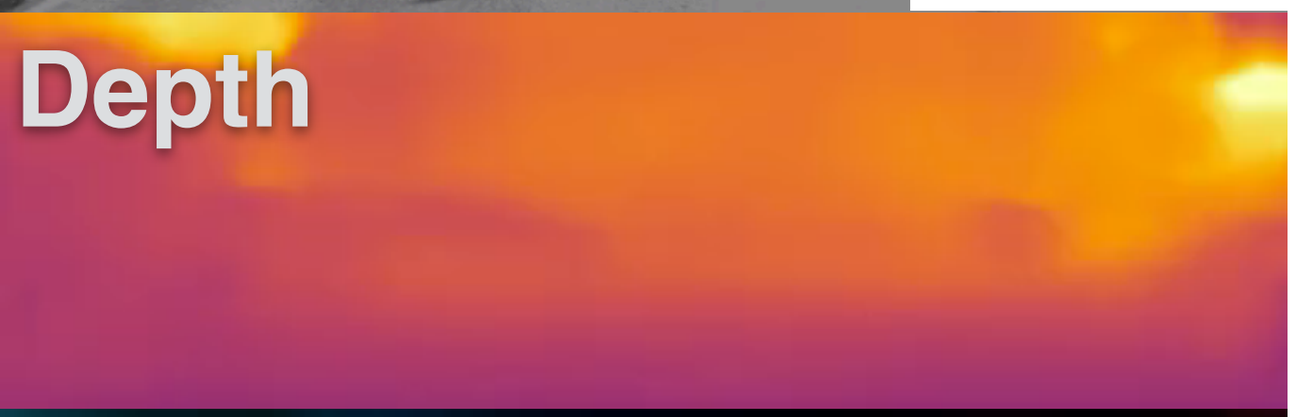
Input

Masks

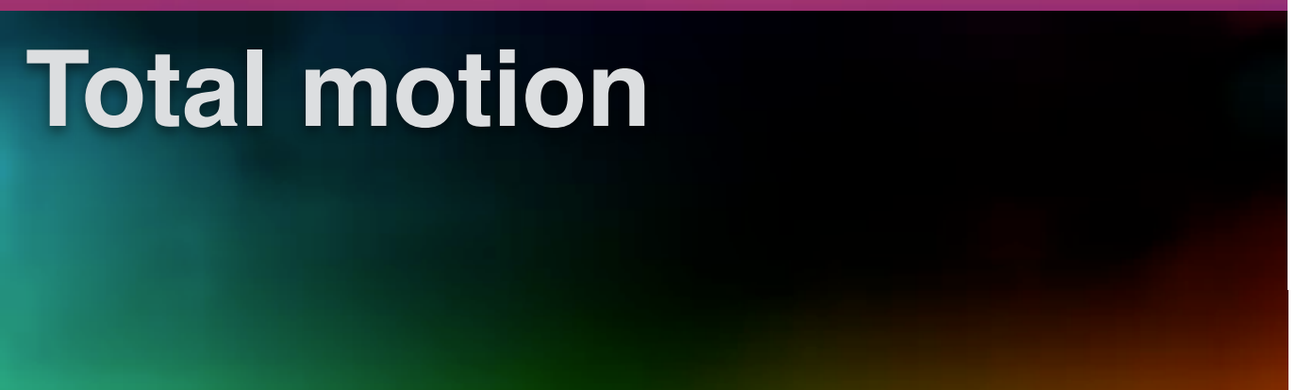
XYZ



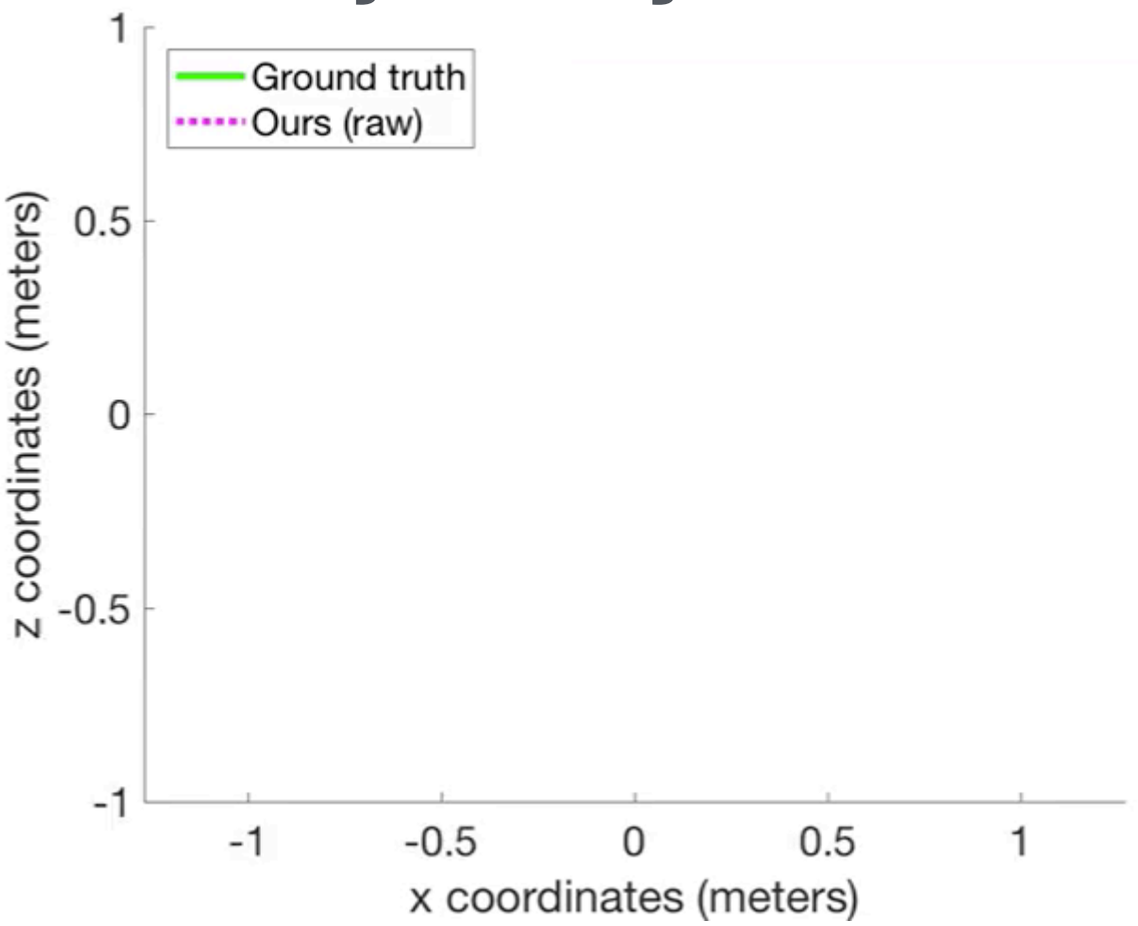
Depth



Total motion



Trajectory vs GT



Ego-stabilized object crops

Object RT Net

[R, t]

MaskNet

Class Net

class

XYZ Net

XYZ

Pose Net

az, el, th

Motion-pose memory

motion-pose consistency

Input

Scene nets

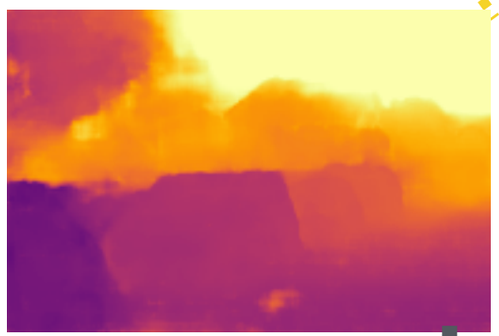
Object nets

RPN

Object RT Net

[R, t]

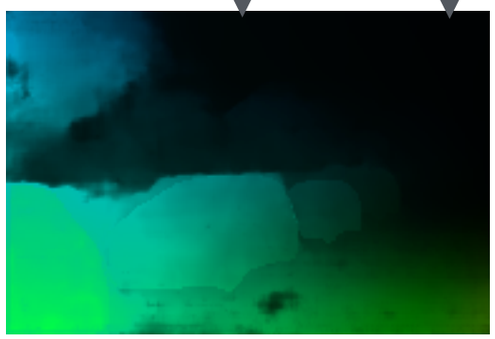
DepthNet



MaskNet

EgoNet

[R, t]



Class Net

class

XYZ Net

XYZ

Pose Net

az, el, th

Input

Scene nets

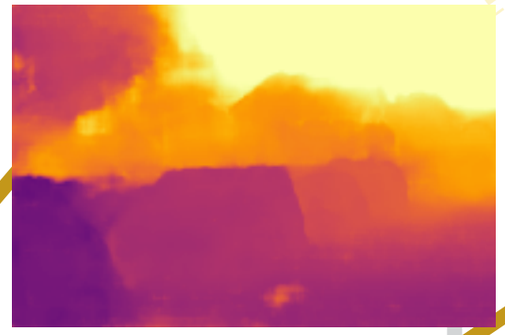
Object nets

RPN

Object RT Net

[R, t]

Depth Net



MaskNet

EgoNet

[R, t]

Class Net

class

XYZ Net

XYZ

Pose Net

az, el, th

Reproj

Input

Scene nets

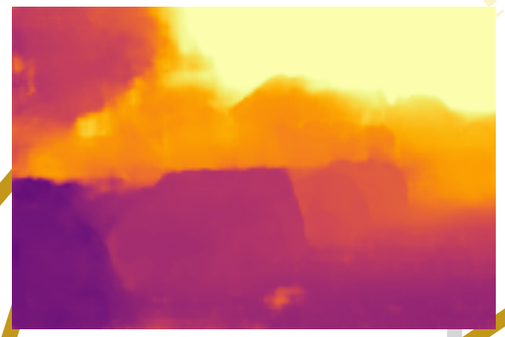
Object nets

RPN

Object RT Net

[R, t]

Depth Net



MaskNet

EgoNet

[R, t]

Class Net

class

XYZ Net

XYZ

Pose Net

az, el, th

Reproj

Scale

Input

Scene nets

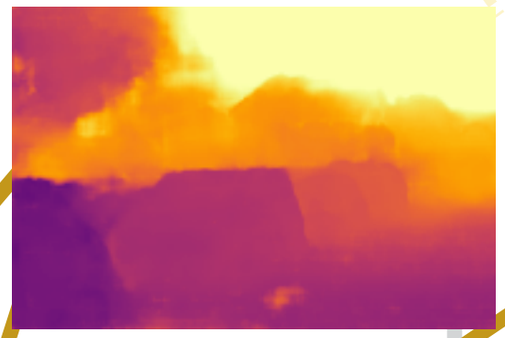
Object nets

RPN

Object RT Net

[R, t]

Depth Net



MaskNet

EgoNet

[R, t]

Class Net

class

XYZ Net

XYZ

Pose Net

az, el, th

Reproj

Scale

Pose

Self supervision helps, even if all networks are pretrained (strongly) supervised.

	Depth metrics		Egomotion		Object-centric	
Net type	**L1 static**	**L1 moving**	**R**	**T**	**Vel**	**Pose**
Pretrained	10.96	10.48	0.006	0.46	0.19	1.69
Ours	8.14	3.49	0.002	0.07	0.14	0.608
Error reduction	**25%**	**67%**	**67%**	**85%**	**26%**	**64%**

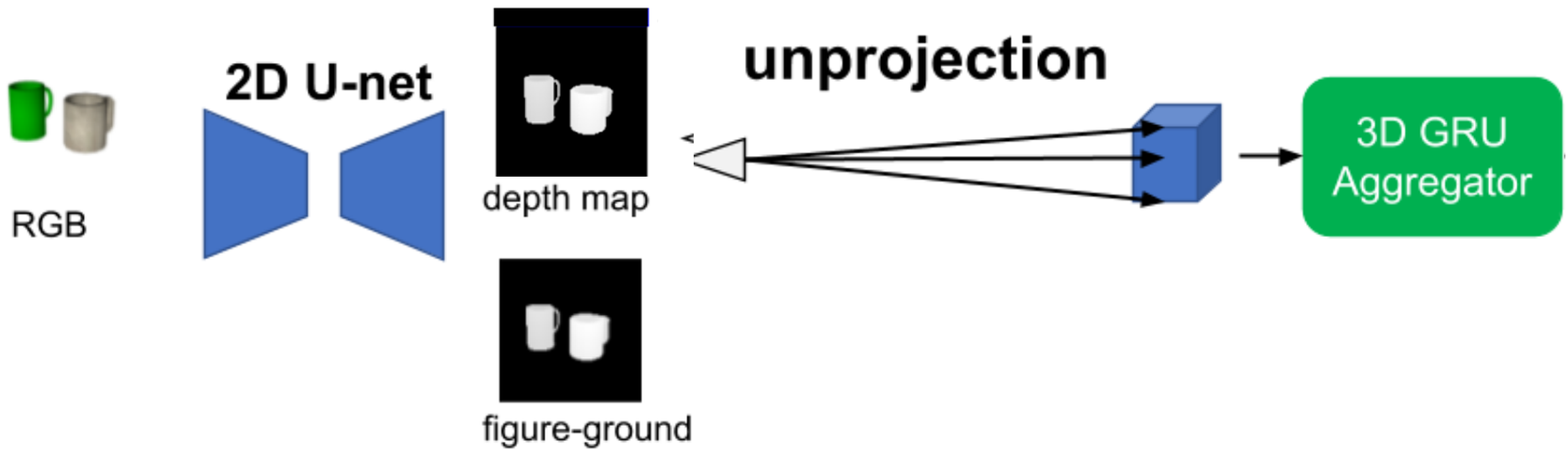
Our outputs are **metric,** but our model wins on **relative** evaluations as well, due to improved training stability.

	Relative depth metrics							Rel. Ego
Net type	**Abs. relative**	**Sq. relative**	**RMSE**	**Log RMSE**	**D1**	**D2**	**D3**	**ATE**
Baseline	0.24	4.81	6.49	0.28	0.72	0.90	0.95	0.15
Ours	0.19	1.41	5.21	0.26	0.71	0.91	0.97	0.14
Error reduction	**21%**	**71%**	**20%**	**7%**	**-1%**	**1%**	**2%**	**6%**

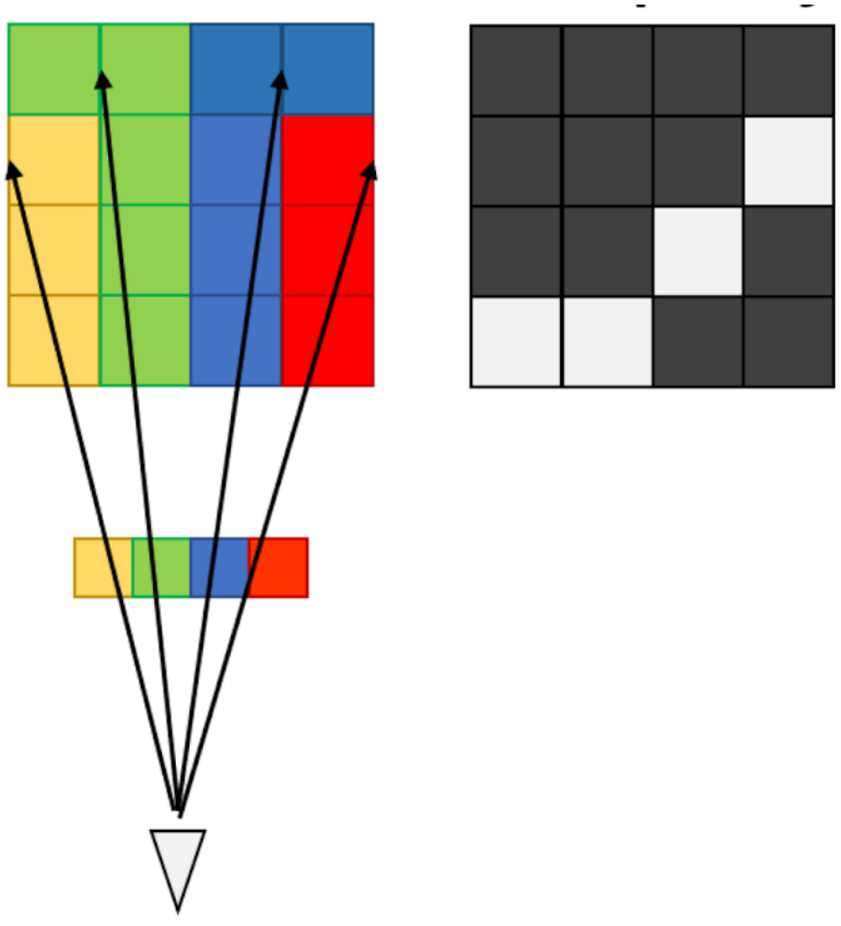
Why Learning in SLAM

- Scale Ambiguity
- Moving Objects
- Mapping the Invisible
- Geometrically-consistent deep memories for recognition in videos

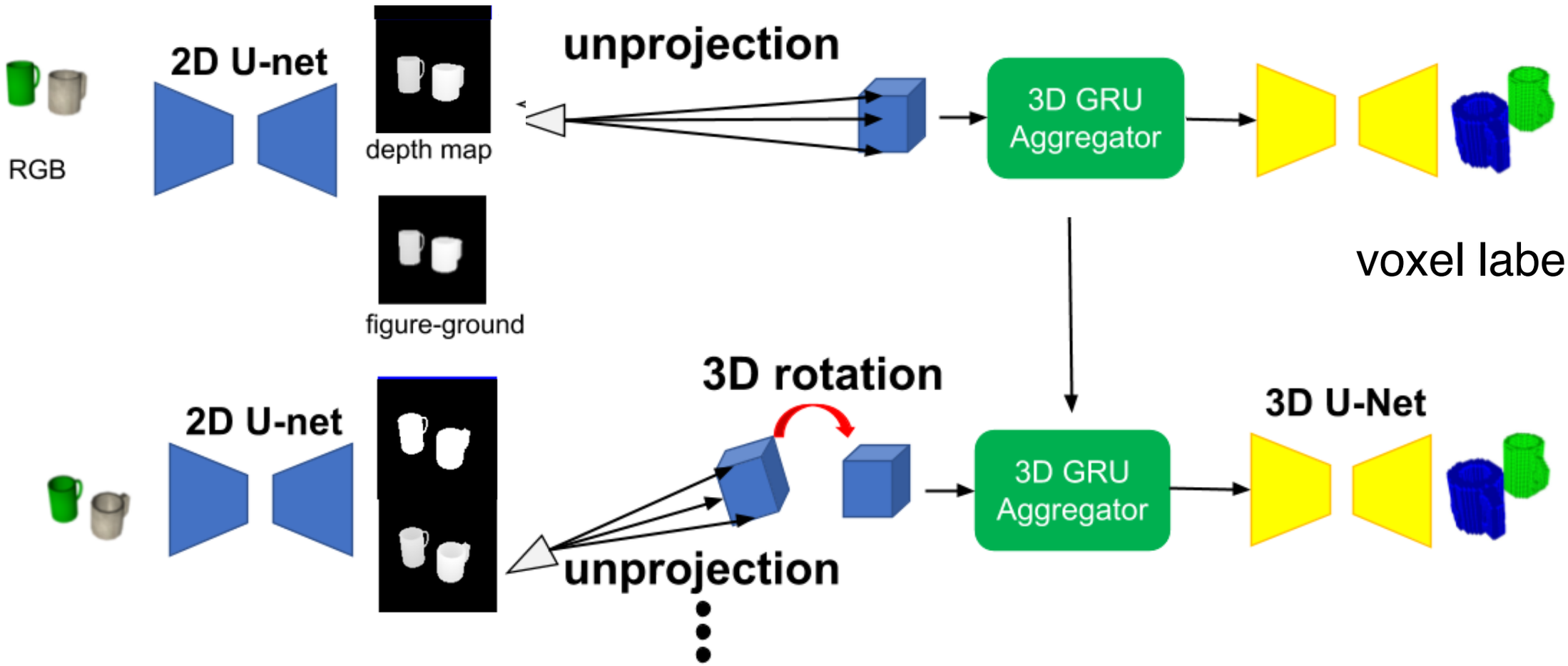
Full 3D scene reconstruction

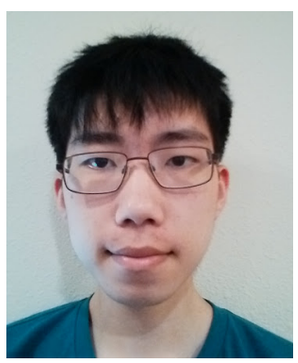


Unprojection



Full 3D scene reconstruction





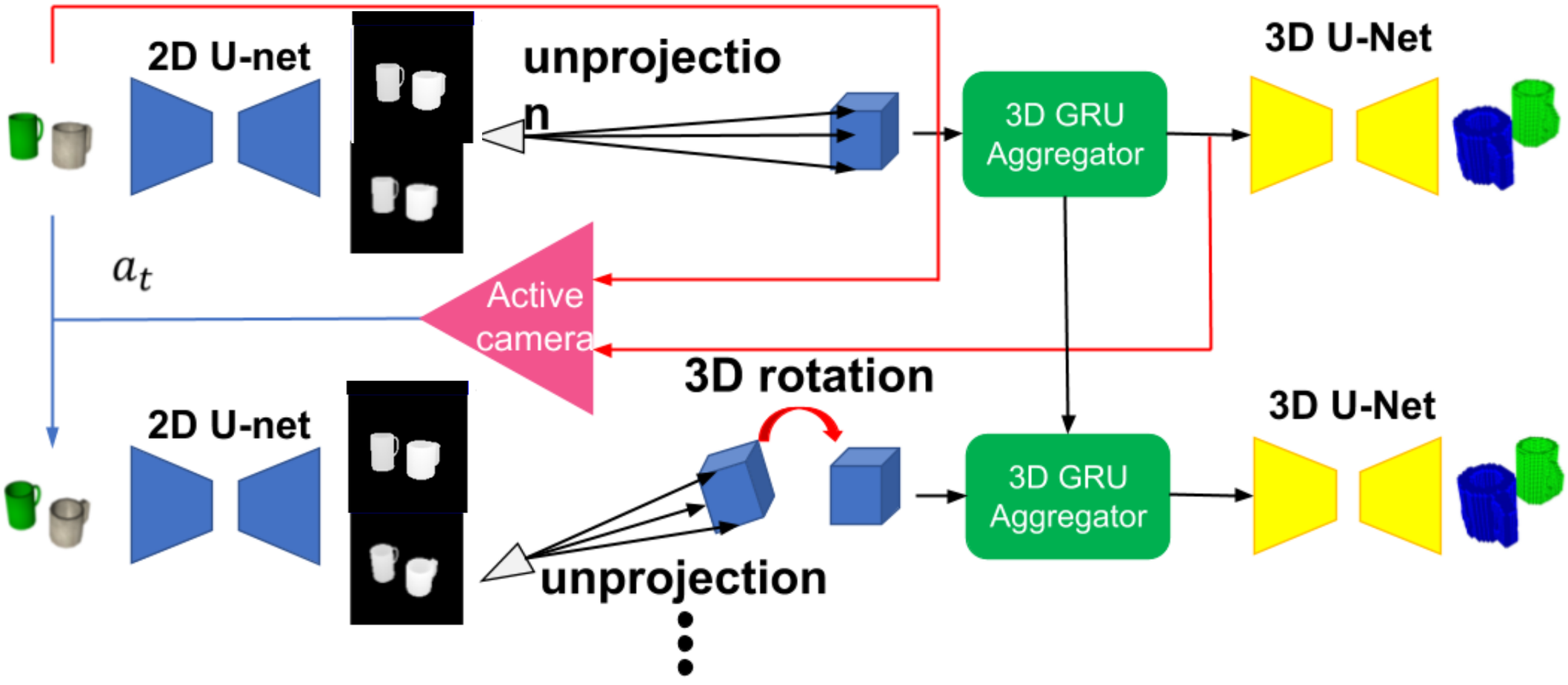
Ricson Chen

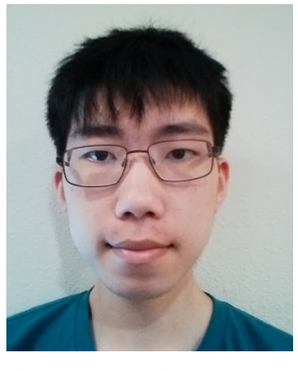
Ziyan Wang

Geometry-aware Active Visual Recognition under Occlusions

In submission

Active full 3D scene reconstruction





Ricson Chen

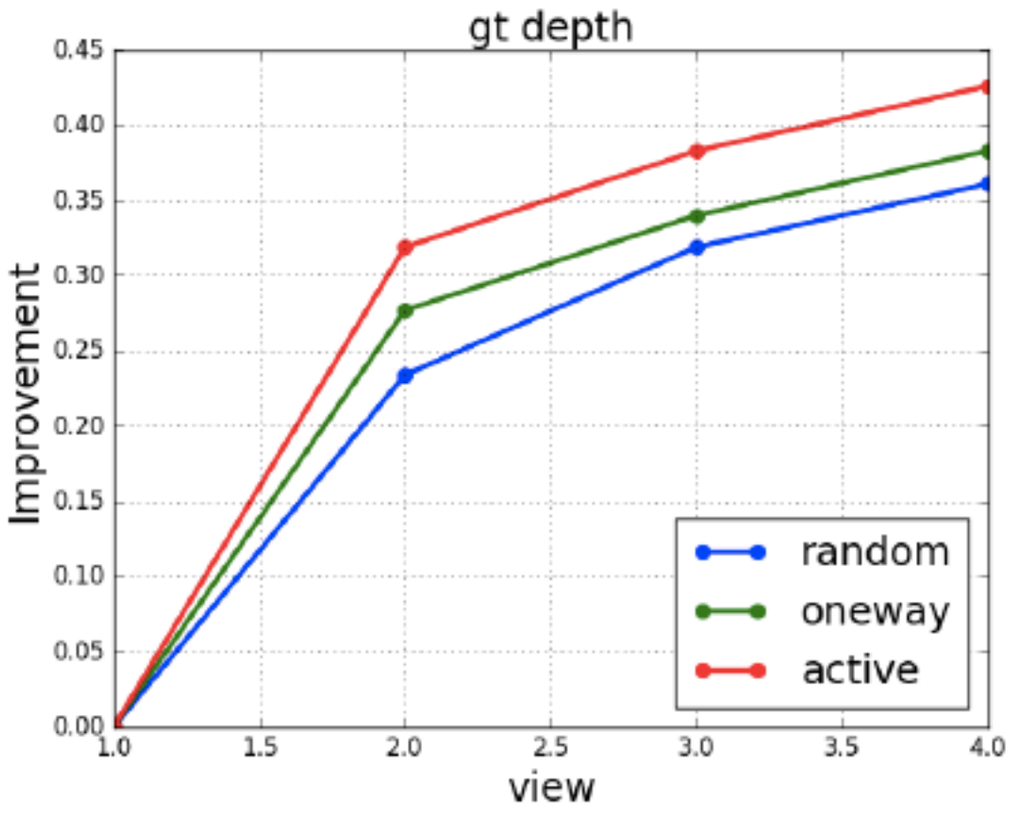
Ziyan Wang

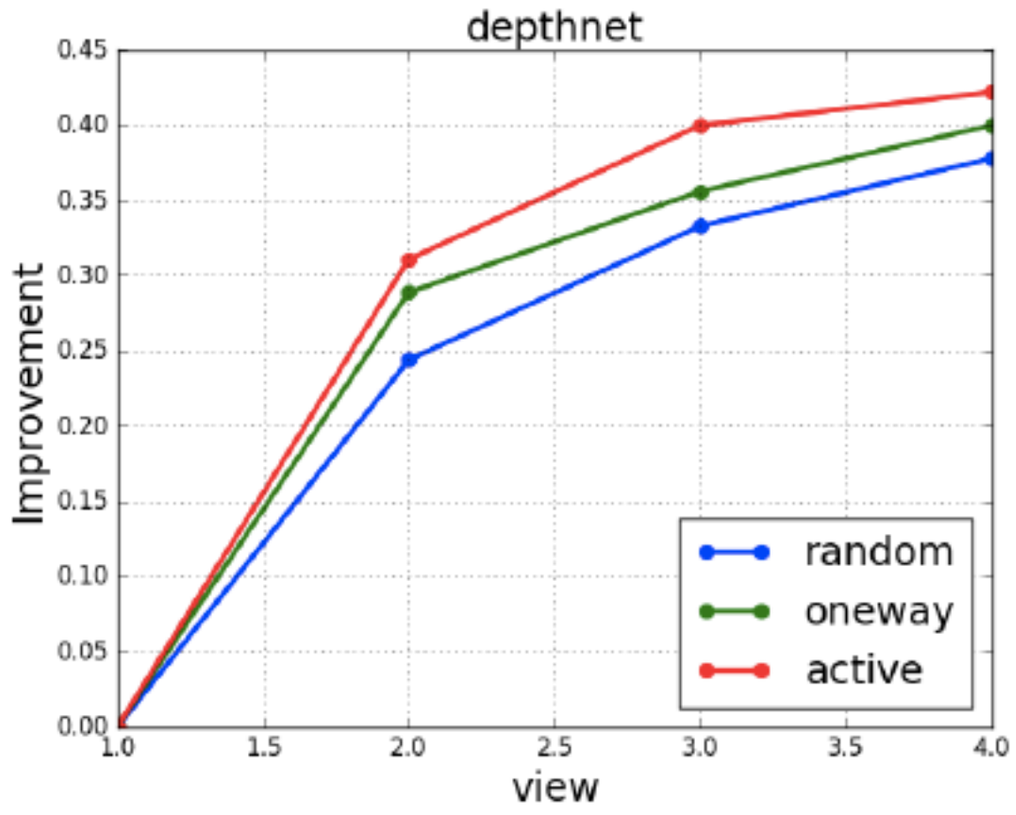
Geometry-aware Active Visual Recognition under Occlusions

In submission

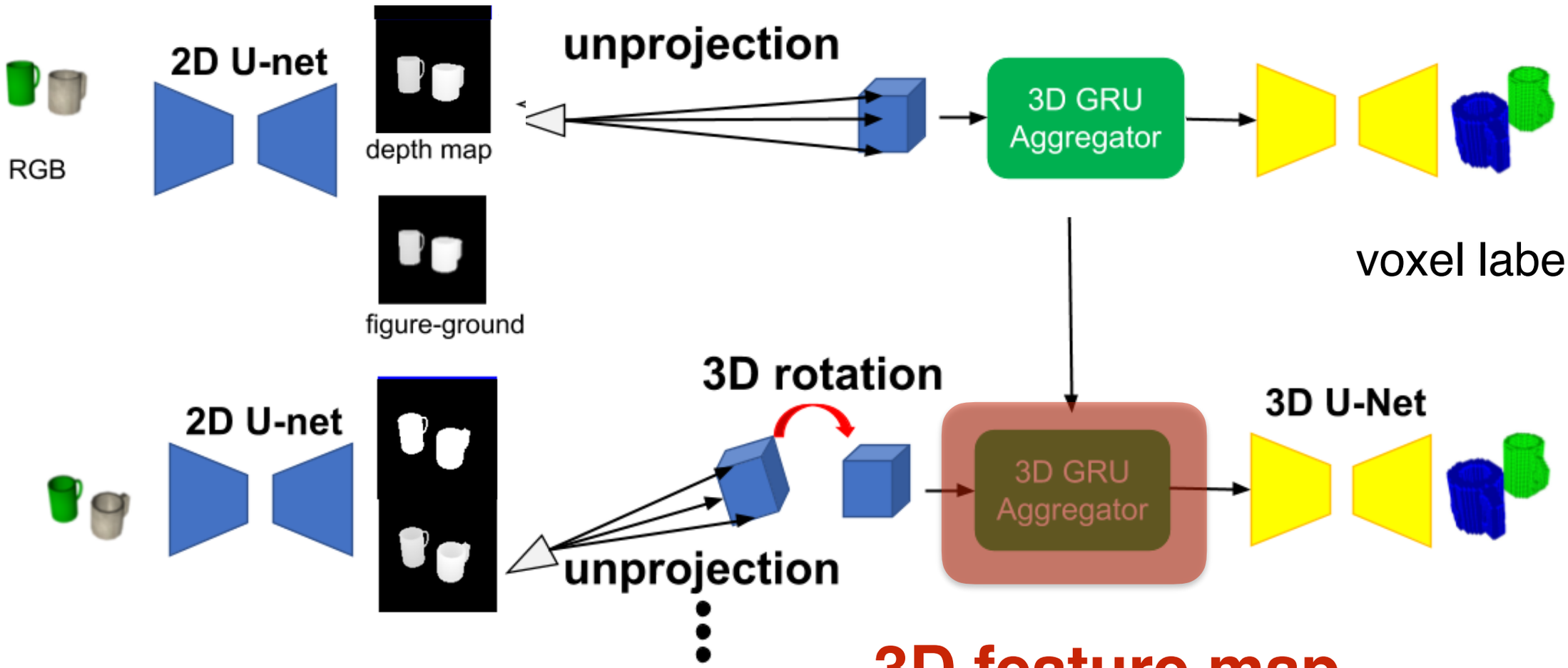
Geometry-aware Active Visual Recognition under Occlusions

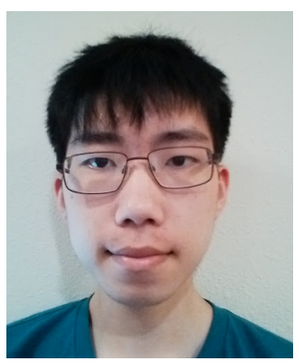
In submission





Full 3D scene reconstruction





Ricson Chen

Ziyan Wang

Geometry-aware Active Visual Recognition under Occlusions

In submission

Benefit of Geometric consistency

	single object				multi-objects			
	view-1	view-2	view-3	view-4	view-1	view-2	view-3	view-4
2D LSTM	0.57	0.59	0.60	0.60	0.11	0.15	0.17	0.20
LSM	0.63	0.66	0.68	0.69	0.43	0.47	0.51	0.53
LSM+gt depth	**0.65**	0.68	0.69	0.70	**0.48**	0.51	0.54	0.56
Ours+gt depth	0.55	**0.69**	**0.72**	**0.73**	0.47	**0.58**	**0.62**	**0.64**
Ours+learnt depth	-	-	-	-	0.45	0.56	0.60	0.62

Why Learning in SLAM

- Scale Ambiguity
- Moving Objects
- Mapping the Invisible
- Geometrically-consistent deep memories for recognition in videos

Active full 3D scene reconstruction

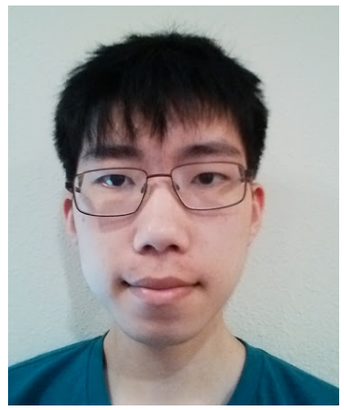
Along 3D reconstruction losses, add segmentation loss and classification loss, directly on the 3D feature map

Object detection can be carried out directly from the 3D feature memory map, as opposed to 2D views

Geometry-aware Active Visual Recognition under Occlusions

In submission

Projecting the object detection results in each 2D view, we get amodal boxes



Ricson Chen

Ziyan Wang

Conclusions

- GANs provide rich data-driven priors to regularize inverse problems

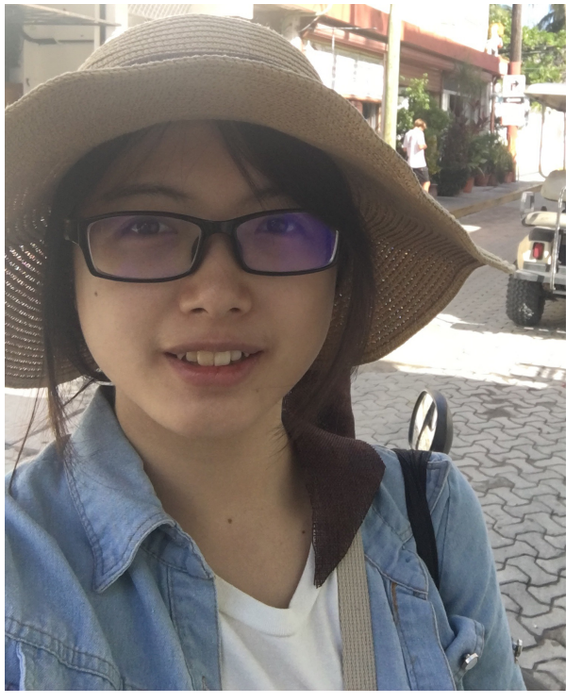
- Object semantics help scale ambiguity in 3D reconstruction, and can deal with moving object in 3D reconstruction

- Egomotion-aware 3D visual feature memory maps produce stable in time object recognition

Thank you!

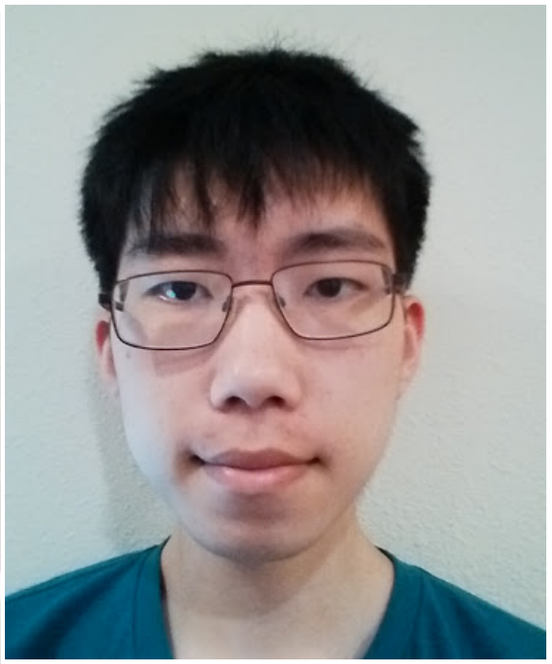
MACHINE LEARNING DEPARTMENT

THE ROBOTICS INSTITUTE



Fish Tung

Adam Harley



Ricson Chen

Ziyan Wang

- Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation with Unpaired Supervision, F. Tung, A. Harley, W. Sato, K.F. et. al, ICCV 2017
- Geometry-Aware Recurrent Neural Networks for Active Visual Recognition , R. Cheng, Z. Wang, K.F., NIPS 2018