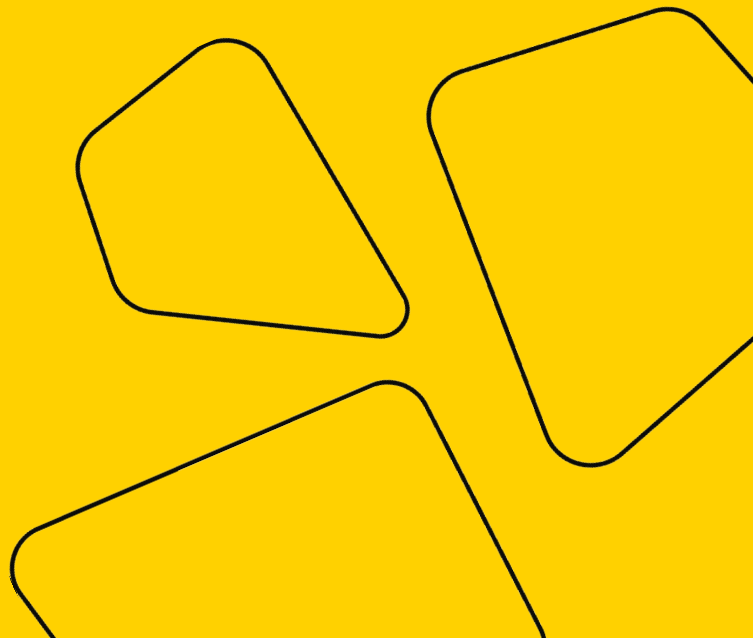
Monocular Depth Estimation

Part 2 (advanced)

Дмитрий Коршунов

Yandex

girafe ai



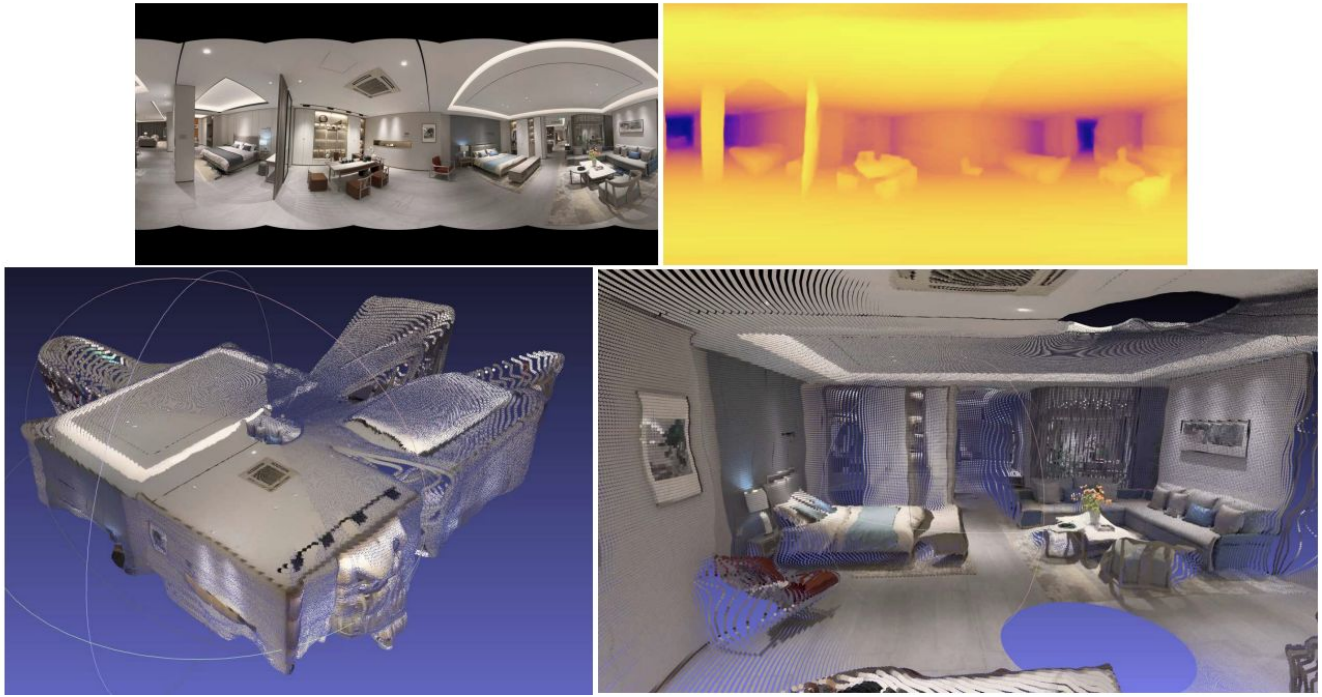
What is depth prediction? (recap)

Input video | Our depth predictions

girafe ai

* https://ai.googleblog.com/2019/05/moving-camera-moving-people-deep.html

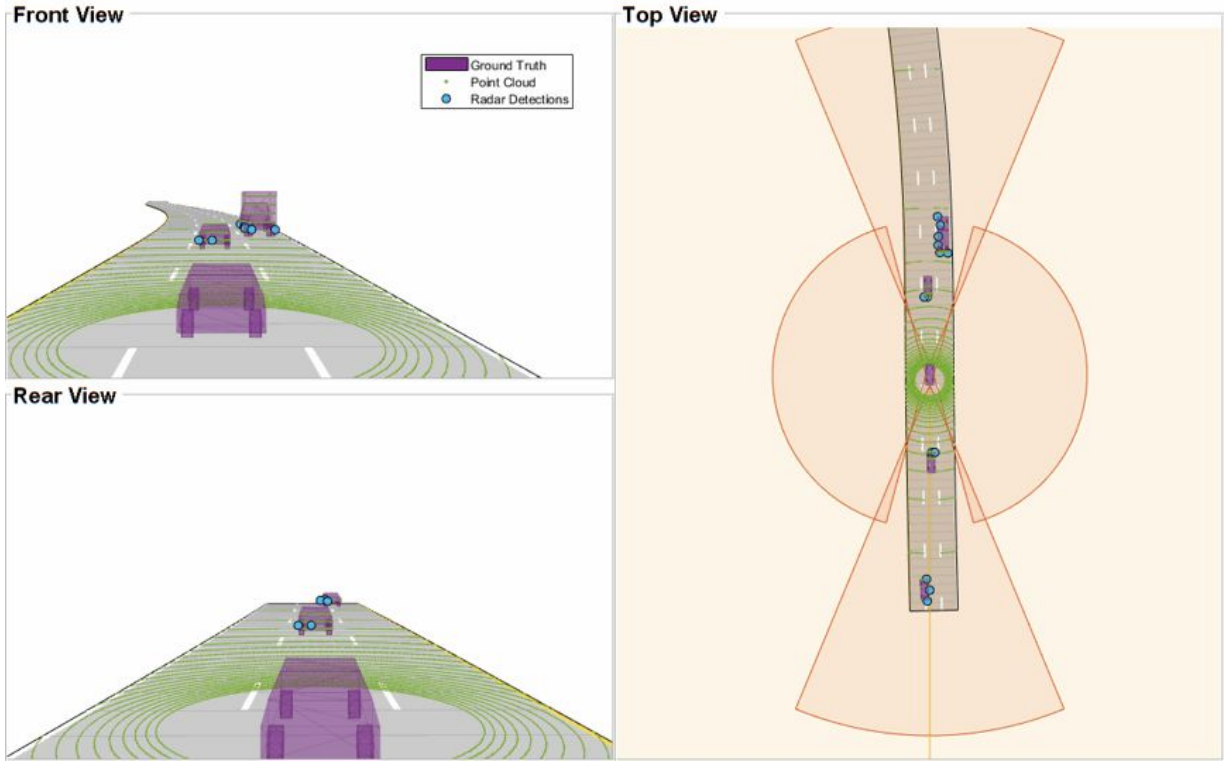
Point cloud building



girafe ai

* https://arxiv.org/pdf/2203.01502.pdf

Point cloud using

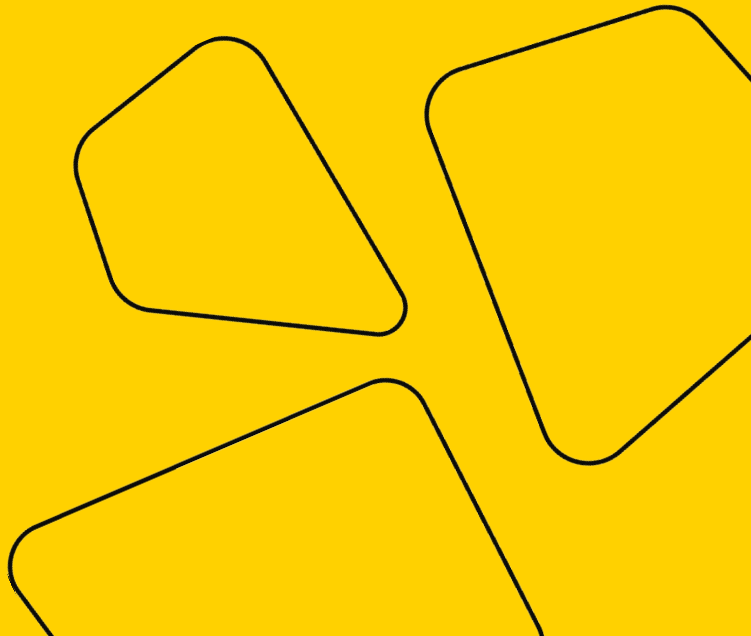


girafe ai

* https://www.mathworks.com/help/driving/ug/track-level-fusion-of-radar-and-lidar-data.html

girafe
ai

Supervised



Better than unsupervised

Rank	Model	absolute relative error ↓
1	**BinsFormer**	0.052
2	**NeWCRFs**	0.052
3	**DepthFormer**	0.052

Rank	Model	absolute relative error ↓
1	**EPCDepth** (S+1024x320)	0.091
2	**DIFFNet** (MS+1024x320)	0.094
3	**CADepth-Net** (MS+1024x320)	0.096

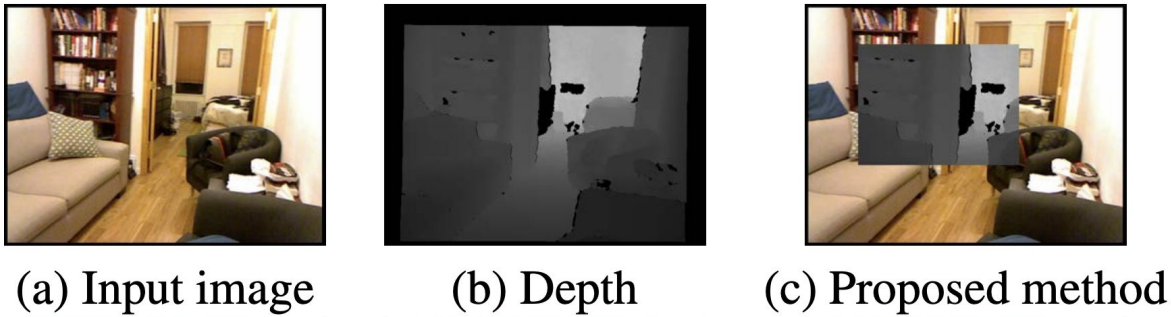
girafe ai

* https://paperswithcode.com/

Datasets (current state)

Dataset	Scenes	Ann. Lidar Fr.	Hours	Cameras	Lidars
KITTI [2]	22	15K	1.5	4	1
NuScenes [3]	1000	40K	5.5	6	1
Argo [4]	113	22K	1	9	2
Waymo 3D [5]	1150	230K	6.4	5	5

Data augmentation



(a) Input image (b) Depth (c) Proposed method

(d) CutOut (e) RE (f) CutMix

Figure 1. Examples of data augmentation

* https://arxiv.org/pdf/2107.07684.pdf

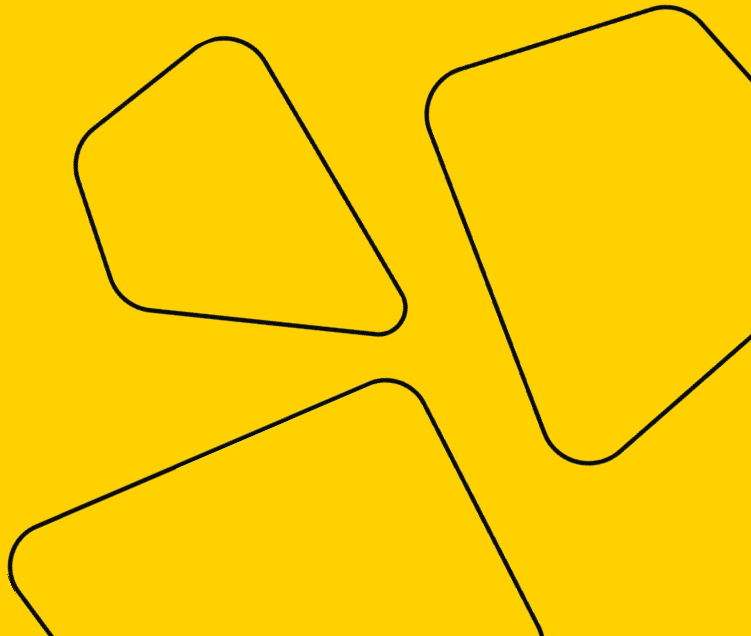
Problems

- The best models are transformers. For most robots, it’s not real-time approach

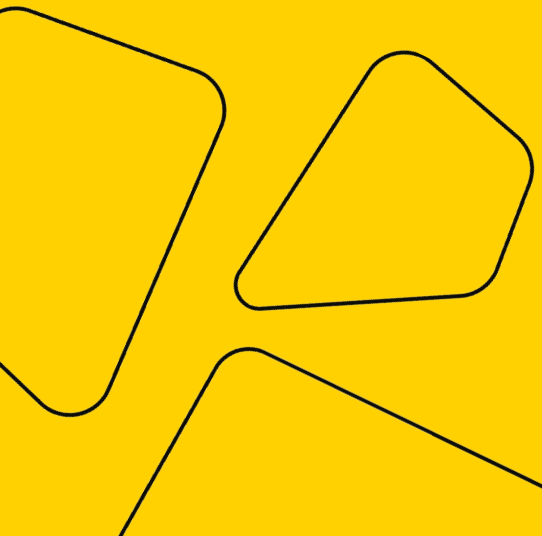
- If you want to adopt the network to your robot you need to install lidar and collect depth data

girafe ai

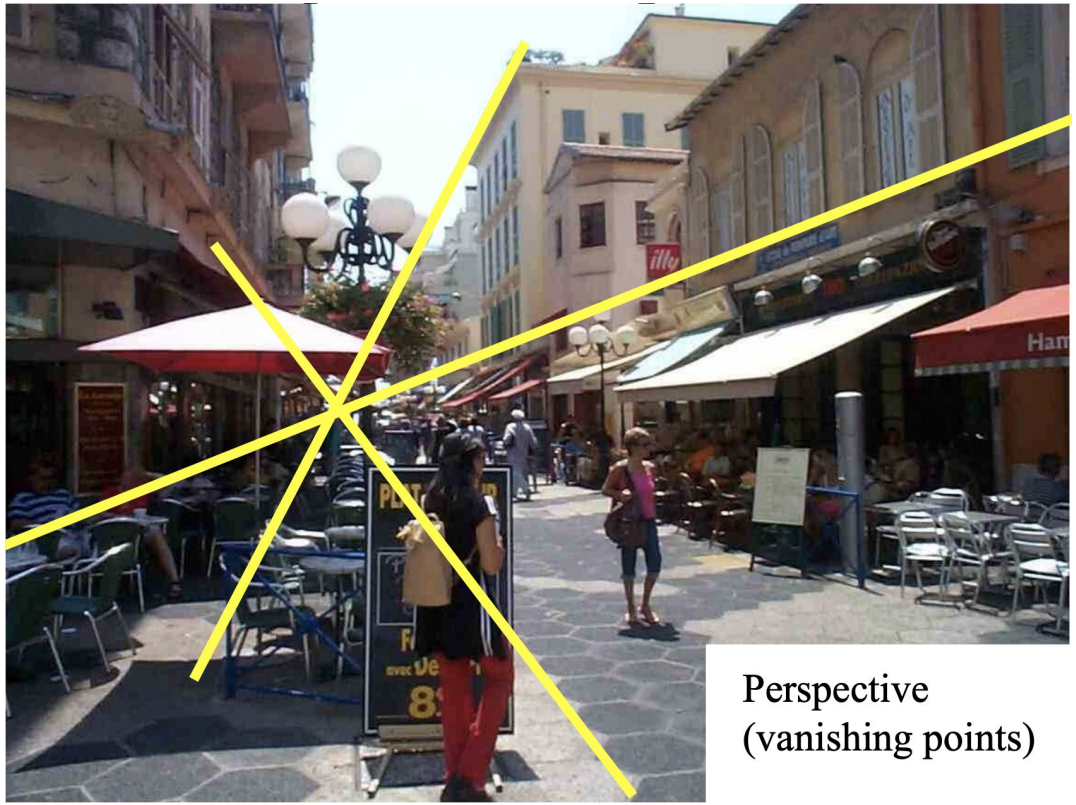
Unsupervised



Base CV

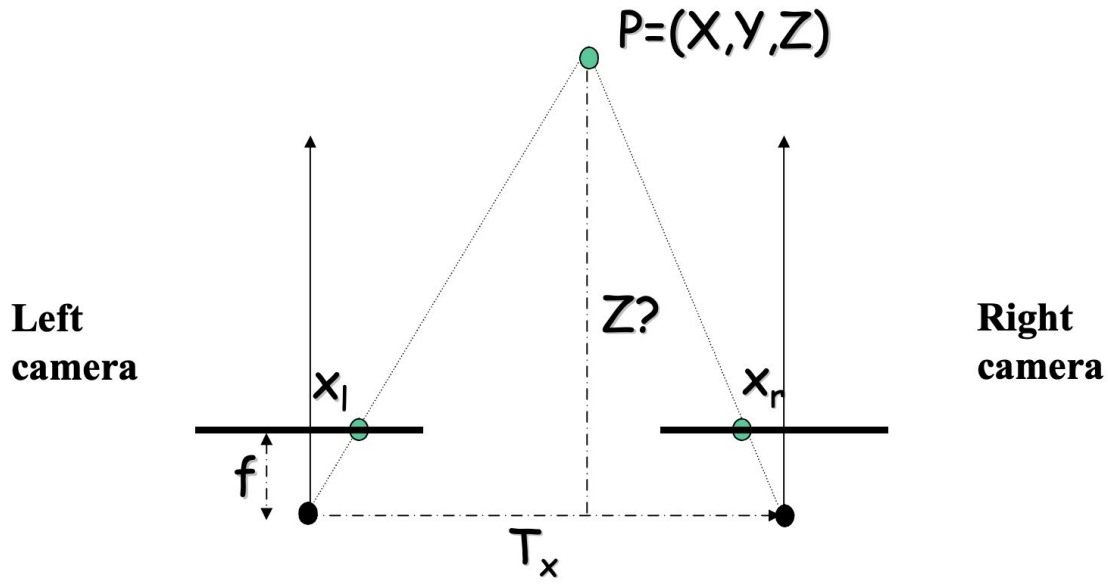


Higher-level Depth Cues



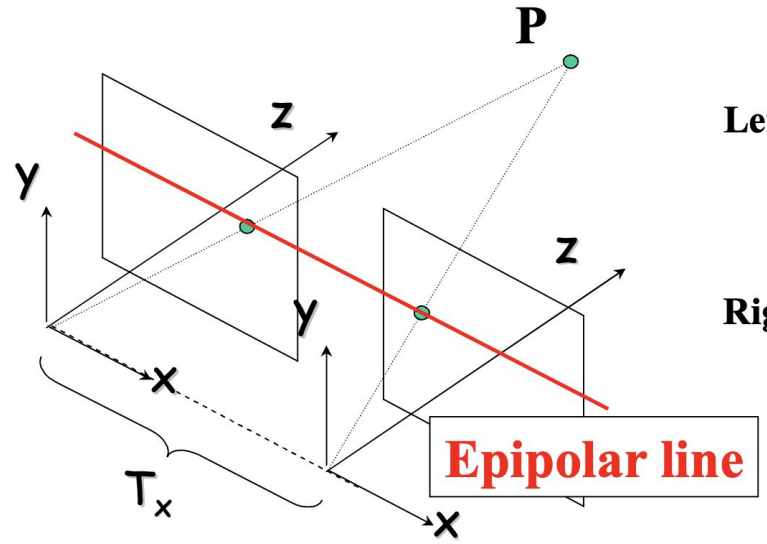
Perspective
(vanishing points)

A Simple Stereo System



Translated by a distance T_x along X axis
(T_x is also called the stereo "baseline")

Epipolar line



Left camera

$$x_l = f\frac{X}{Z} \qquad y_l = f\frac{Y}{Z}$$

Right camera

$$x_r = f\frac{X - T_x}{Z} \qquad y_r = f\frac{Y}{Z}$$

Same Y Coord!

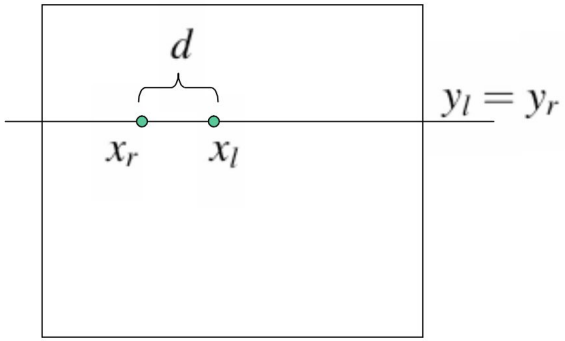
Stereo Disparity

Left camera

$$x_l = f\frac{X}{Z} \qquad y_l = f\frac{Y}{Z}$$

Right camera

$$x_r = f\frac{X - T_x}{Z} \qquad y_r = f\frac{Y}{Z}$$



Stereo Disparity

$$d = x_l - x_r = f\frac{X}{Z} - \left(f\frac{X}{Z} - f\frac{T_x}{Z}\right)$$

$$d = \frac{f\,T_x}{Z}$$

depth **baseline**

$$Z = \frac{f\,T_x}{d}$$

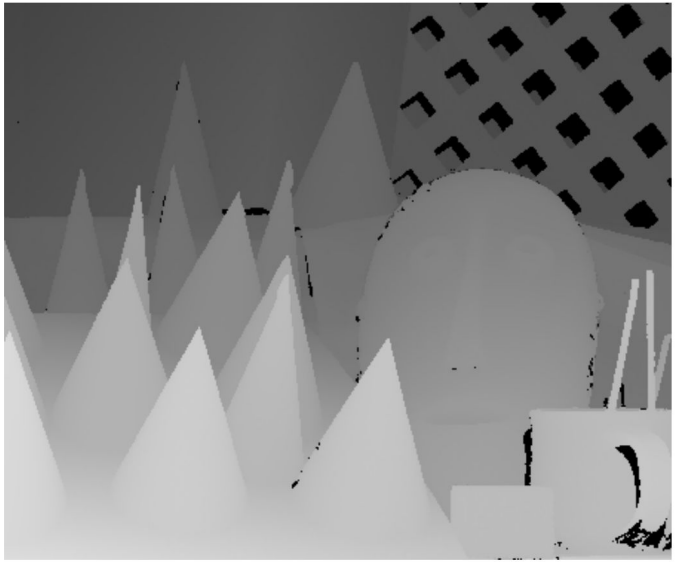
disparity

Important equation!

girafe ai

Stereo Example

Disparity values (0-64)

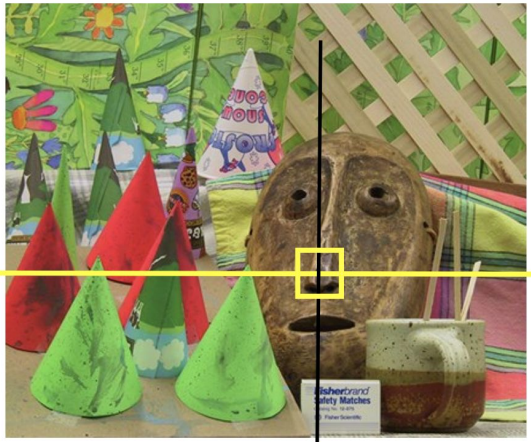


Note how disparity is larger (brighter) for closer surfaces.

Matching using Epipolar Lines

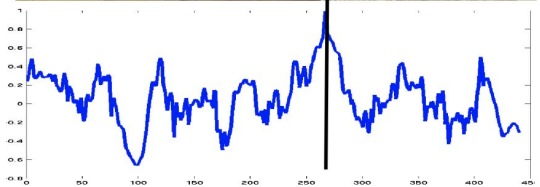
Left Image

Right Image



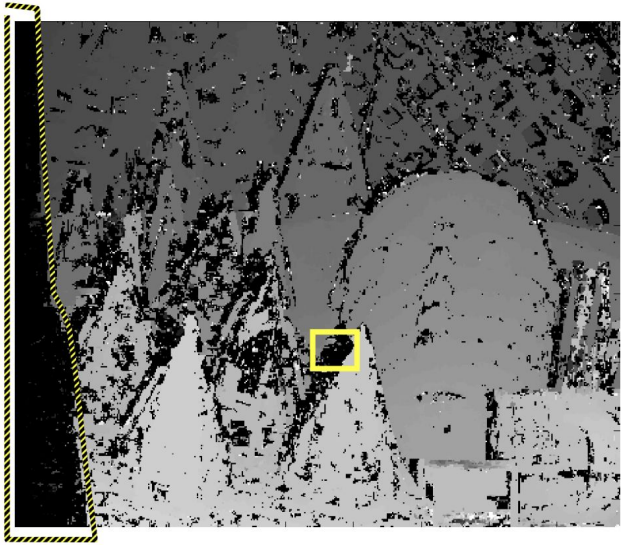
Select patch with highest match score.

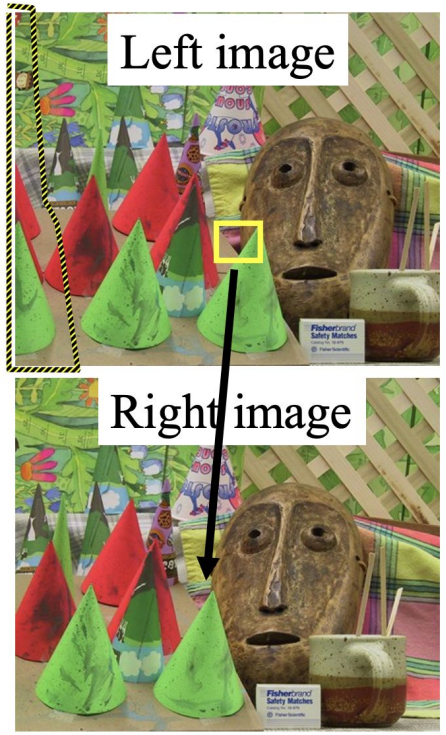
Repeat for all pixels in left image.



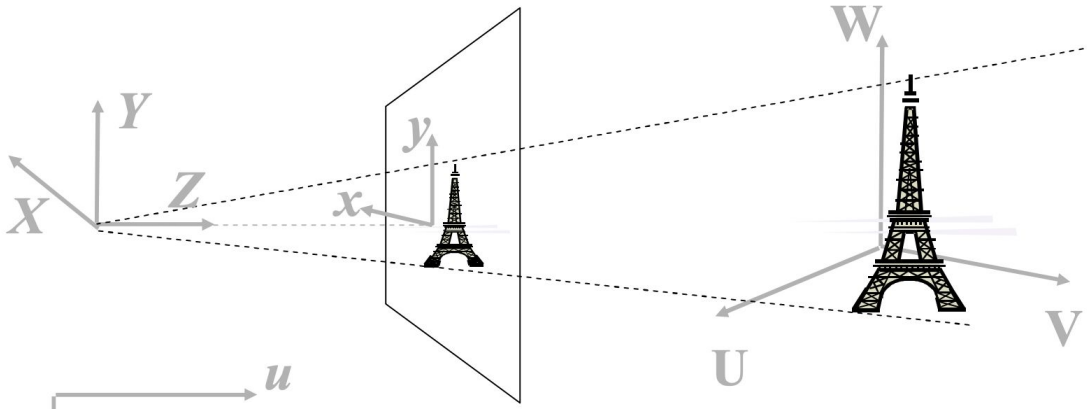
Match Score Values

Occlusions: No matches





Imaging Geometry



Our image gets digitized into pixel coordinates *(u,v)*

Perspective Matrix Equation

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

$$y = f\frac{Y}{Z}$$

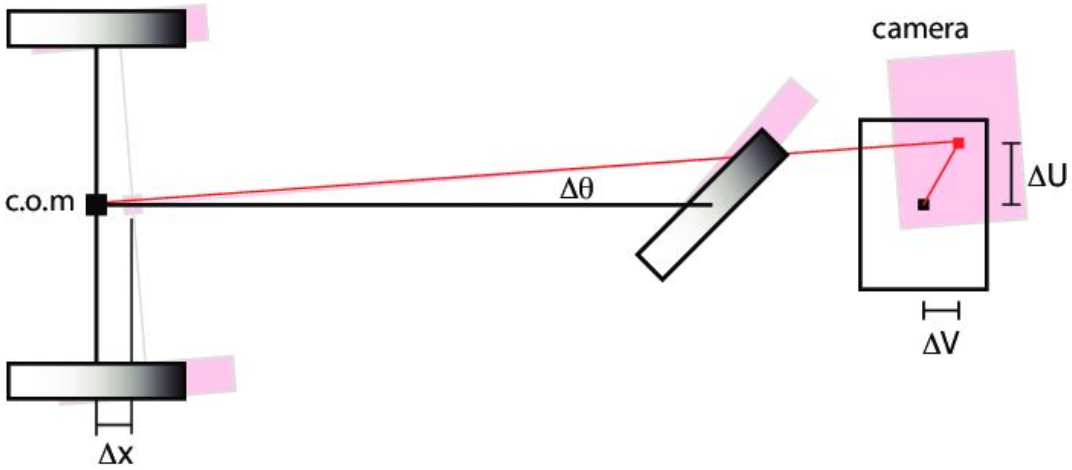
Using homogeneous coordinates:

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

$$x = \frac{x'}{z'} \quad y = \frac{y'}{z'}$$

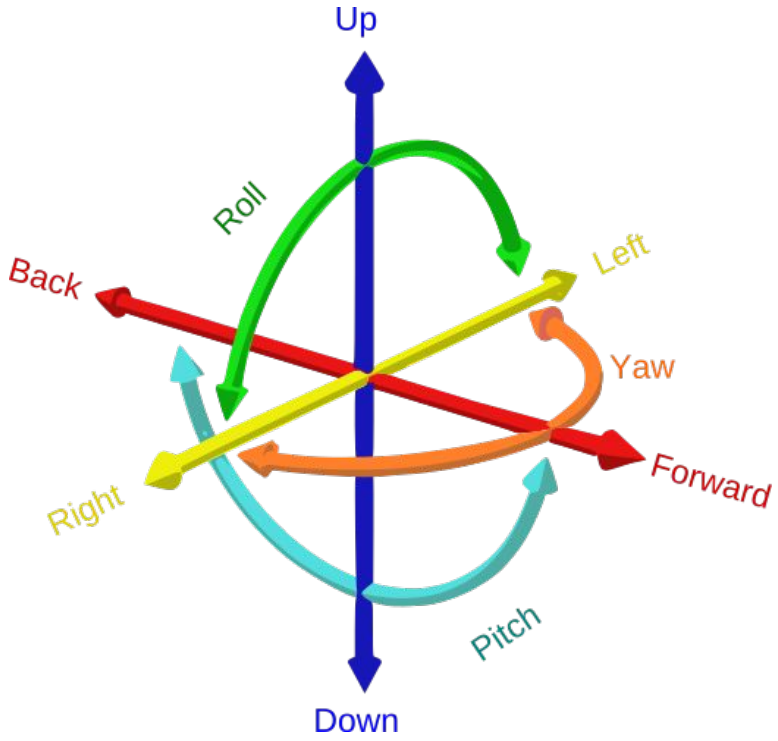
Odometry

Odometry is the use of data from motion sensors to estimate change in position over time.



*https://www.researchgate.net/figure/A-simplified-motion-and-odometry-model-for-a-vehicle-with-Ackerman-like-steering-model_fig1_224557194

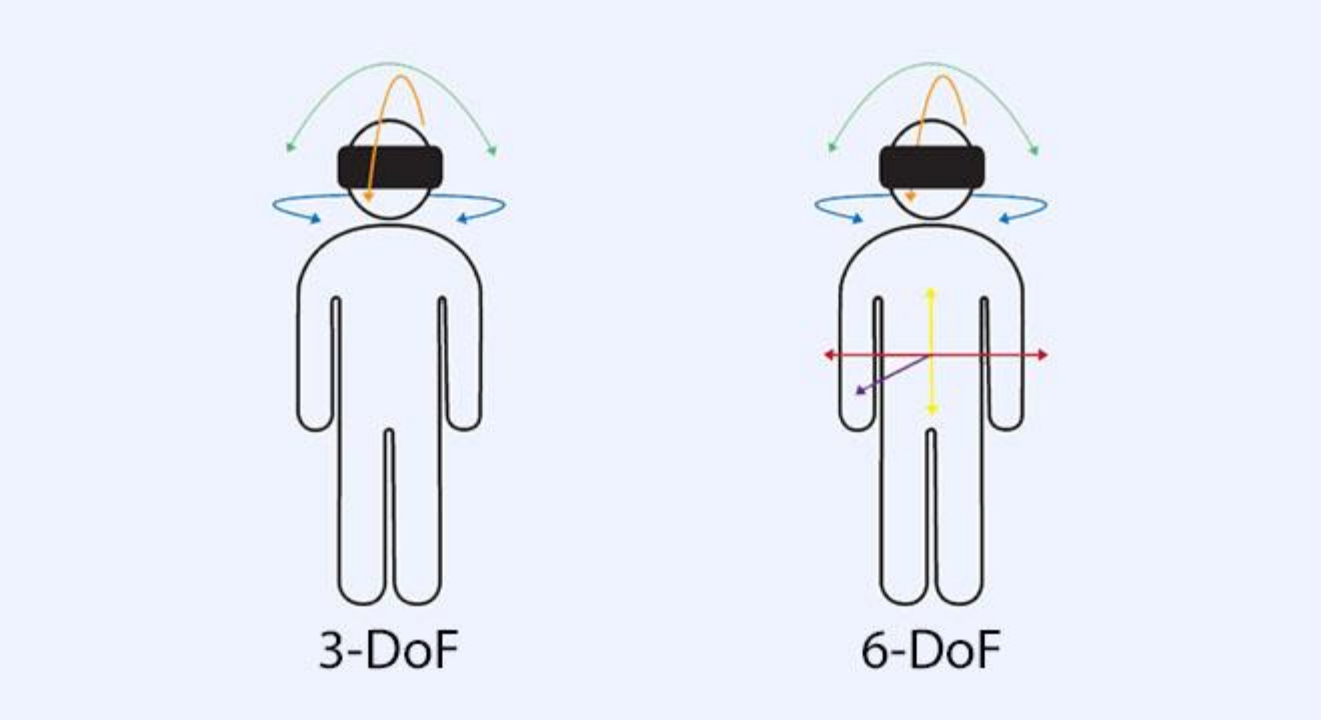
6-DOF (network output example)



girafe ai

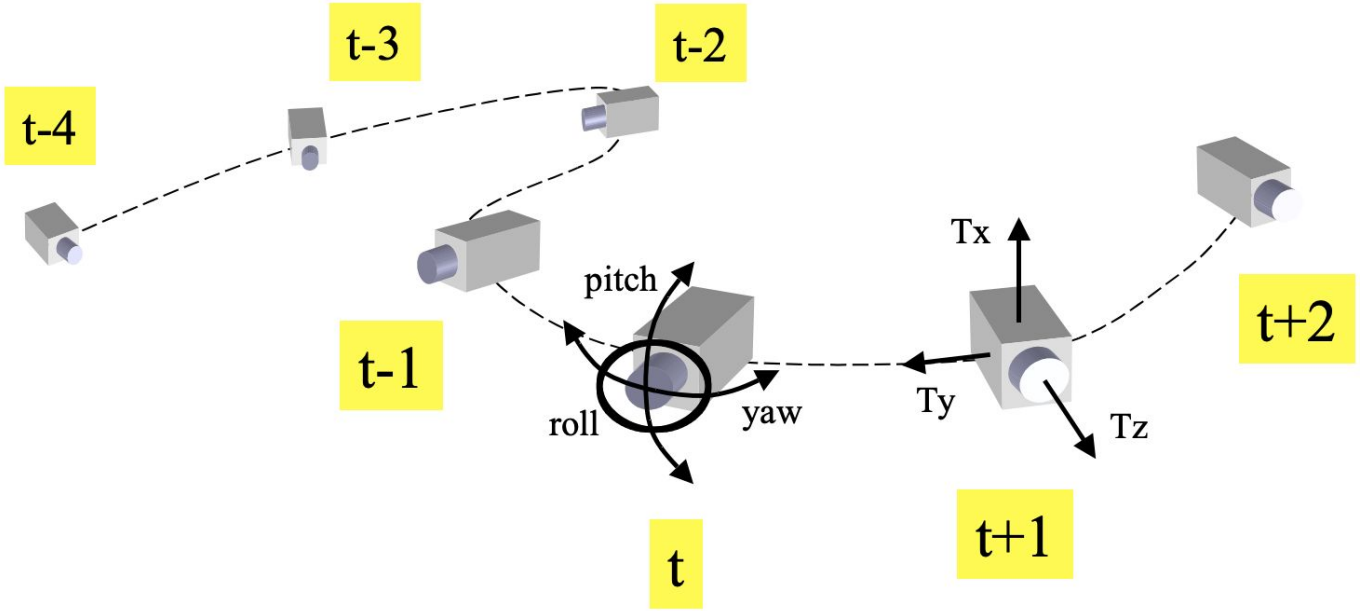
* https://en.wikipedia.org/wiki/Six_degrees_of_freedom

3-DOF VS 6-DOF



* https://virtualspeech.com/blog/degrees-of-freedom-vr

6-DOF (diff prediction)



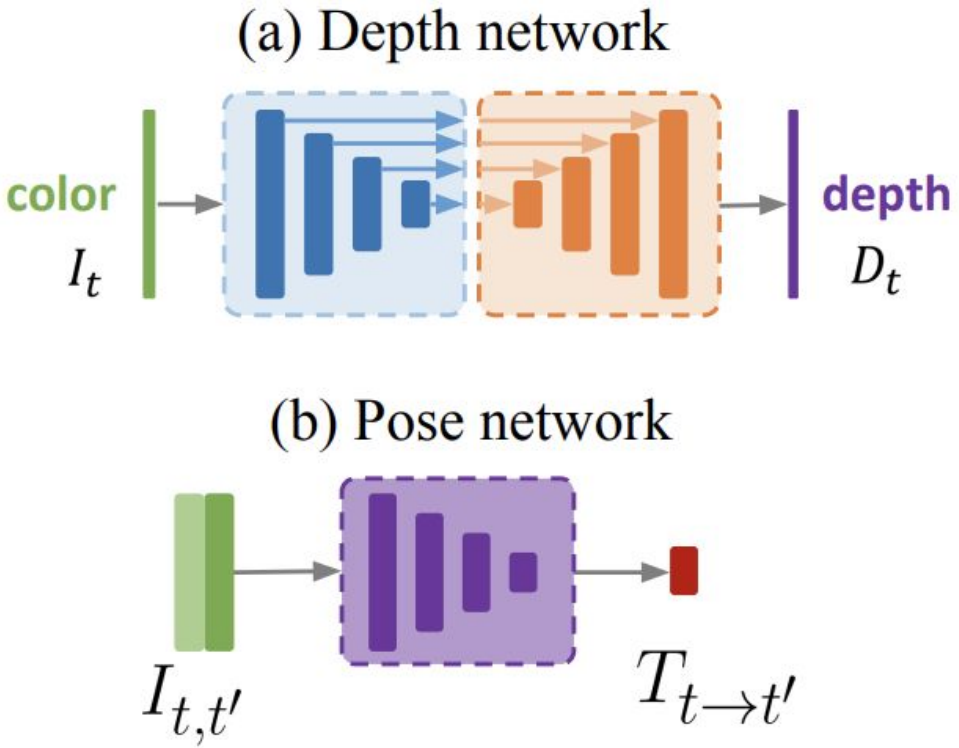
girafe ai

* https://www.cse.psu.edu/~rtc12/CSE486/

Odometry problem

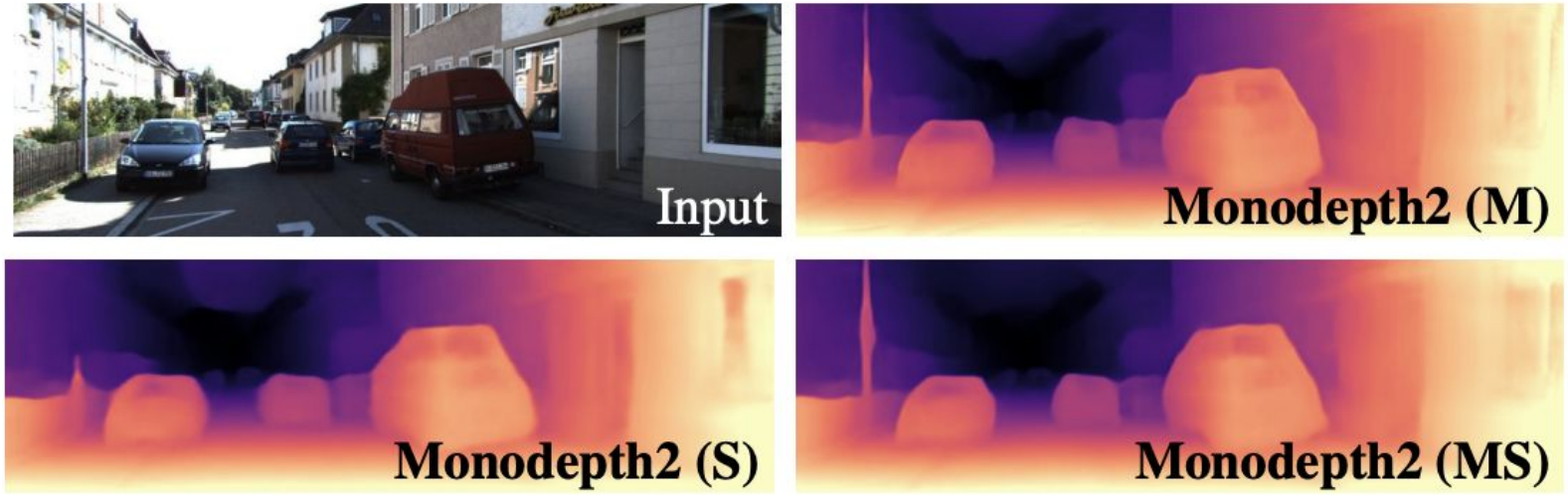
* https://arxiv.org/pdf/1704.07813v2.pdf

MonoDepth2



* https://arxiv.org/pdf/1806.01260.pdf

Mono and stereo approaches



Loss function

Photometric reconstruction error

$$pe(I_a, I_b) = \frac{\alpha}{2}(1 - \text{SSIM}(I_a, I_b)) + (1 - \alpha)\|I_a - I_b\|_1$$

where,

I - reconstructed and reference images

SSIM - Structural similarity index

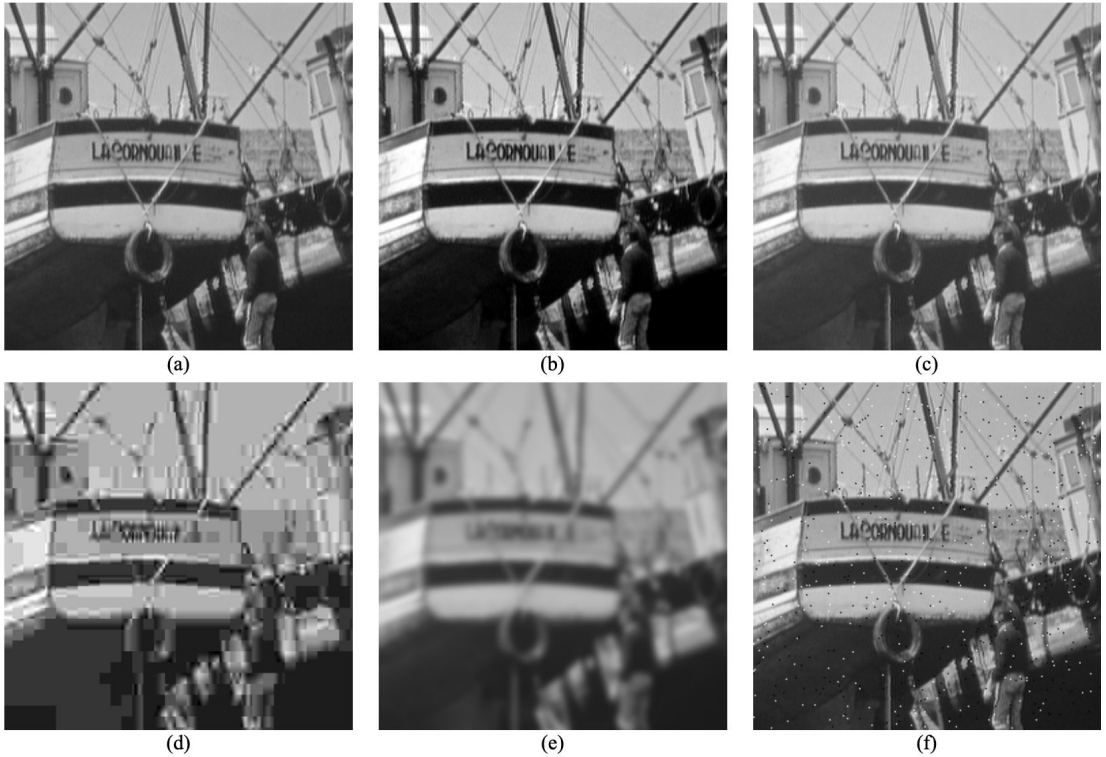
α = 0.85

$$L_p = \min_{t'} pe(I_t, I_{t' \to t}).$$

final per-pixel photometric loss

* https://papers.nips.cc/paper/2014/file/7bccfde7714a1ebadf06c5f4cea752c1-Paper.pdf

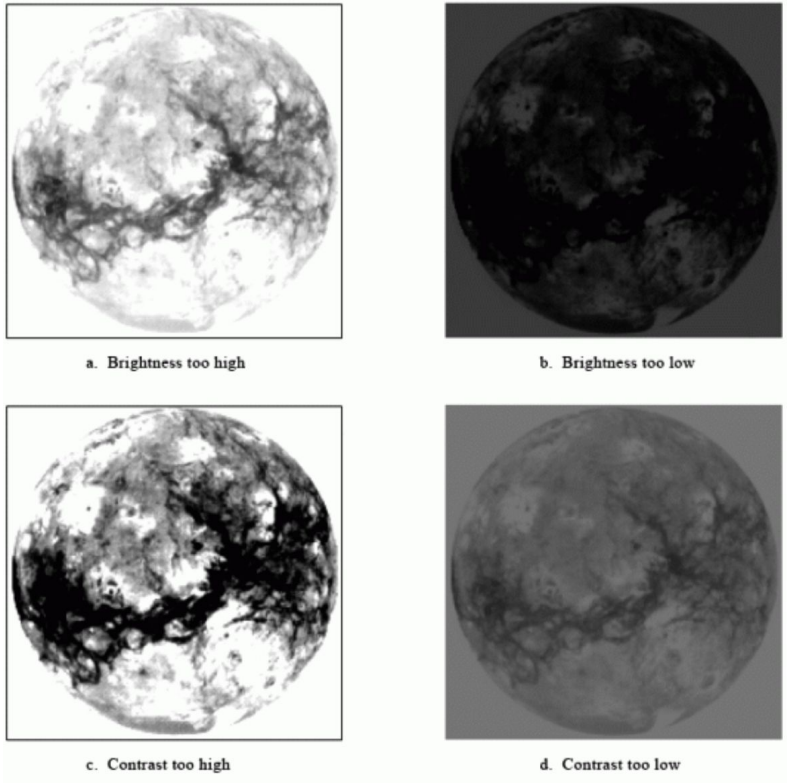
L1 loss problem



girafe ai

* https://www.cns.nyu.edu/pub/lcv/wang03-preprint.pdf

Problem generalization



a. Brightness too high

b. Brightness too low

c. Contrast too high

d. Contrast too low

* https://www.dspguide.com/ch23/5.htm

Structural SIMilarity (SSIM) Index

$$S(\mathbf{x}, \mathbf{y}) = f(l(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y}))$$

Where

l - luminance comparison function (luminance masking)

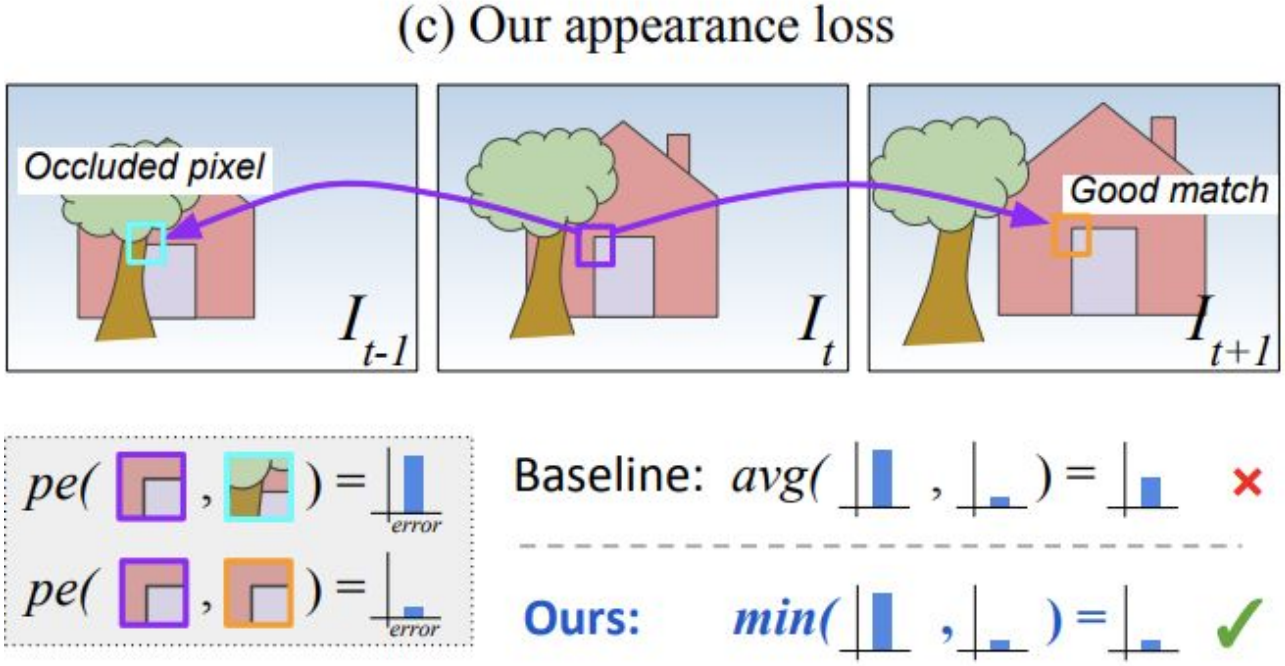
c - contrast comparison function

s - structure comparison function (geometrically correlation)

girafe ai

* https://www.cns.nyu.edu/pub/lcv/wang03-preprint.pdf

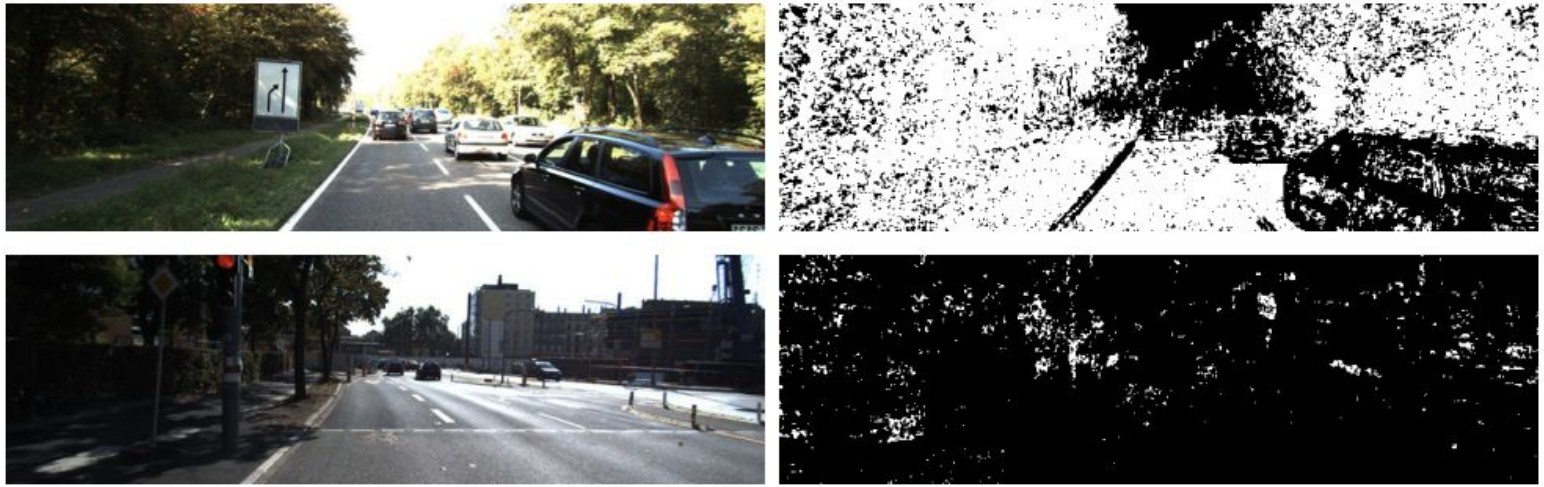
Per-pixel minimum reprojection



girafe ai

* https://arxiv.org/pdf/1806.01260.pdf

Auto-masking



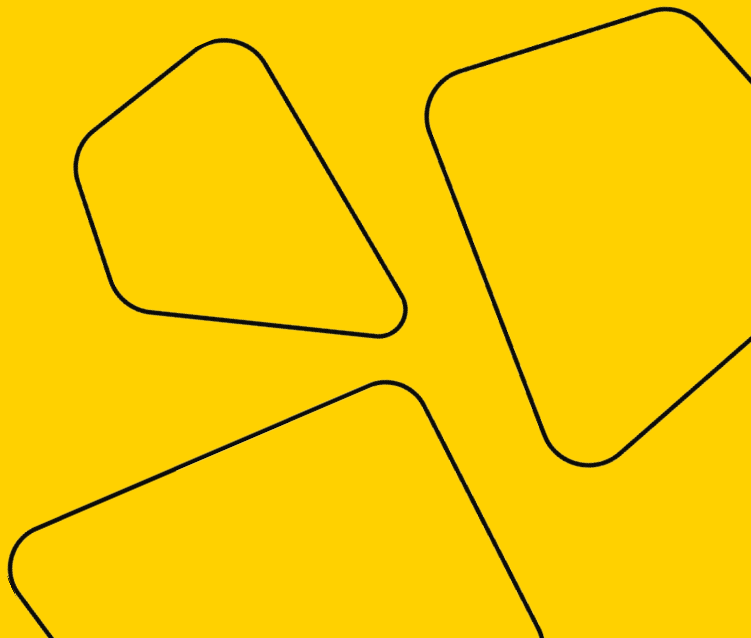
The mask prevents objects moving at similar speeds to the camera (top) and whole frames where the camera is static (bottom) from contaminating the loss

All together

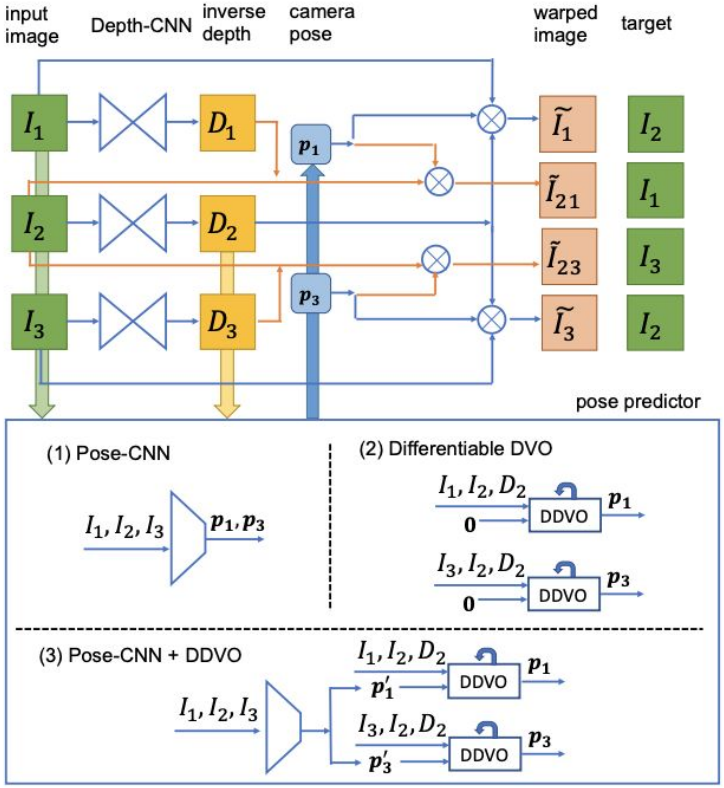
- Predicting depth from single image (for left image or current frame)
- Predicting odometry data diff between neighbors frames
- Using depth prediction, odometry predictions, and camera calibration data reproject pixels from the left image to the surface of the right image.
- Using the photometric loss function calculating an error between the right image and created reprojection on the right image surface.

girafe ai

More approaches



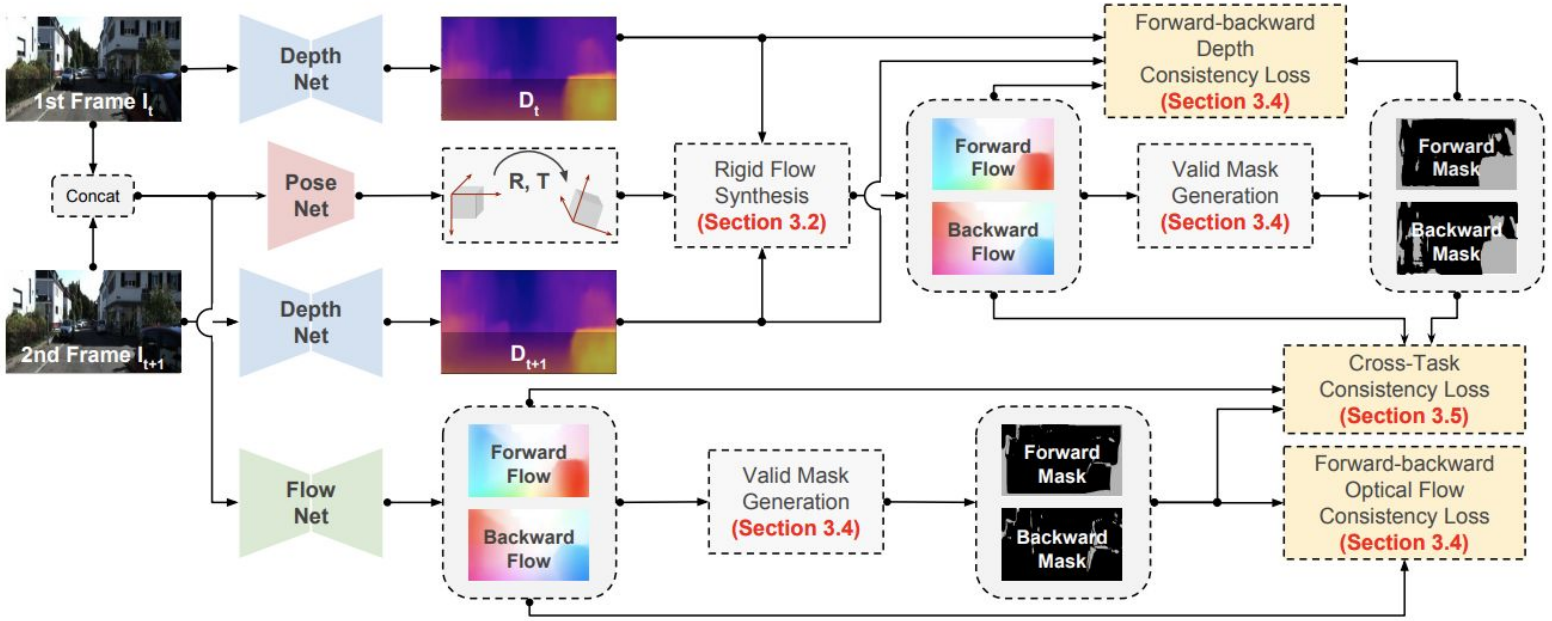
Depth from Videos (Direct Methods)



girafe ai

* https://arxiv.org/pdf/1712.00175.pdf

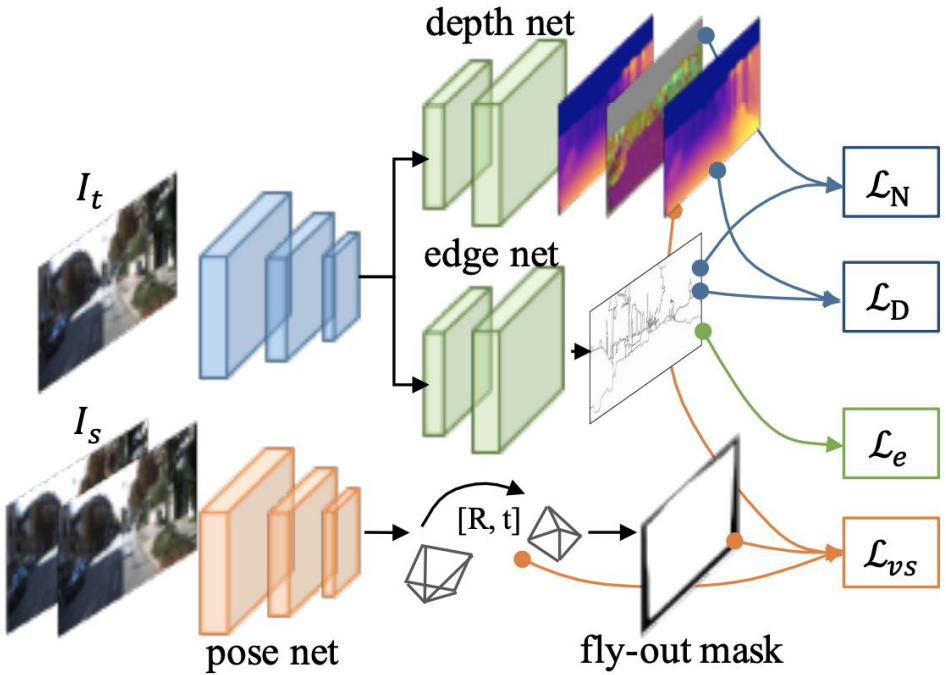
DF-Net: Depth and Flow



girafe ai

* https://arxiv.org/pdf/1809.01649.pdf

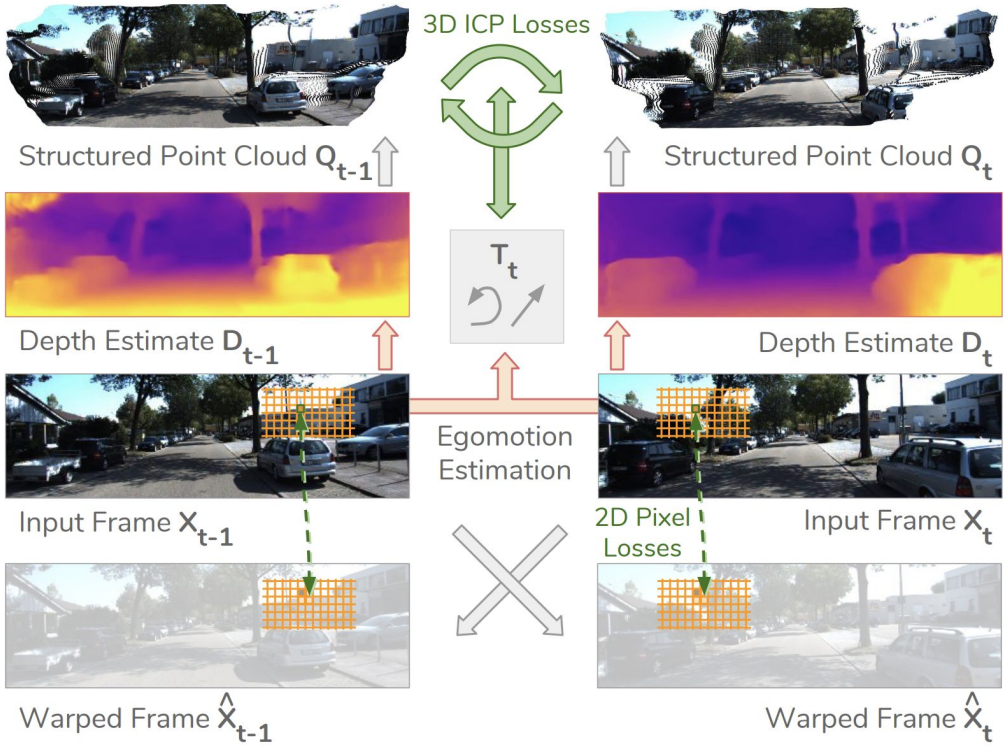
LEGO: Learning Edge with Geometry



girafe ai

* https://arxiv.org/pdf/1803.05648.pdf

Depth Using 3D Geometric Constraints



girafe ai

* https://arxiv.org/pdf/1802.05522v2.pdf

Conclusions

- Supervised networks need depth labeled data which is hard to collect
- To adopt a pre-trained supervised network to the autonomous vehicles need to collect depth data with lidars
- Unsupervised approaches show better speed inference for real-time using
- To adopt the model to a robot not necessary to install lidars

Tips and tricks

If you don’t have the ability to install lidar or collect odometry data, you can adopt pre-trained models pre calculating the median error and subtracting it.

girafe
ai

Thank you!