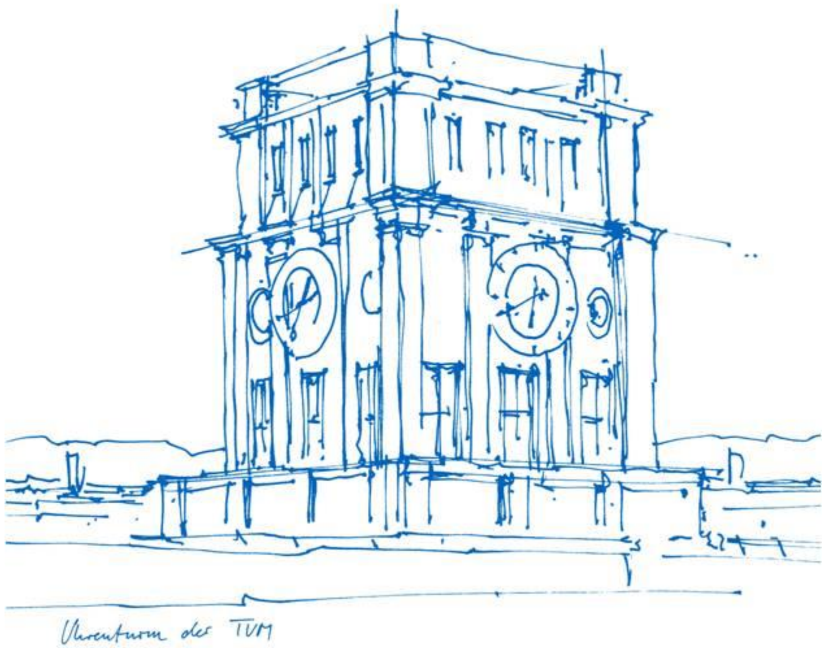
Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry

Nan Yang, Rui Wang, Jörg Stückler, Daniel Cremers

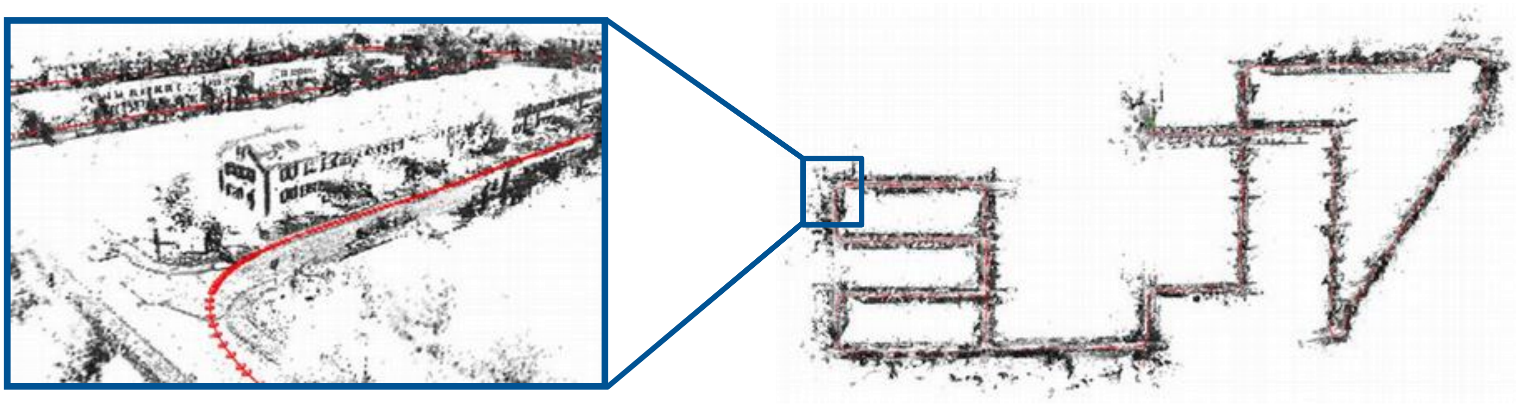
Lars Carius

Technical University of Munich

Munich, March 16th 2020



Intro – Visual Odometry

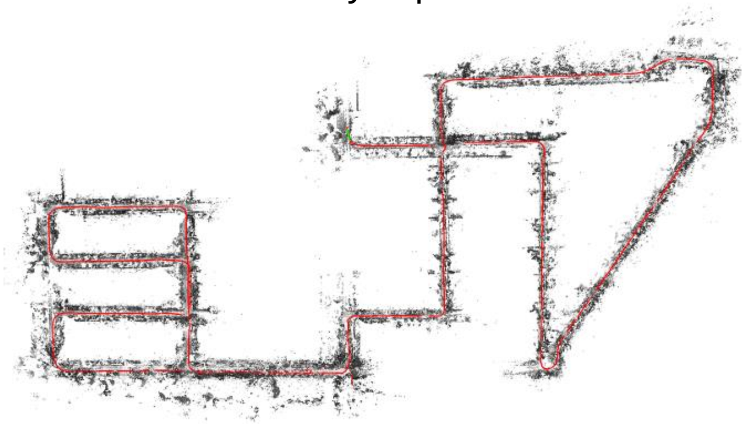


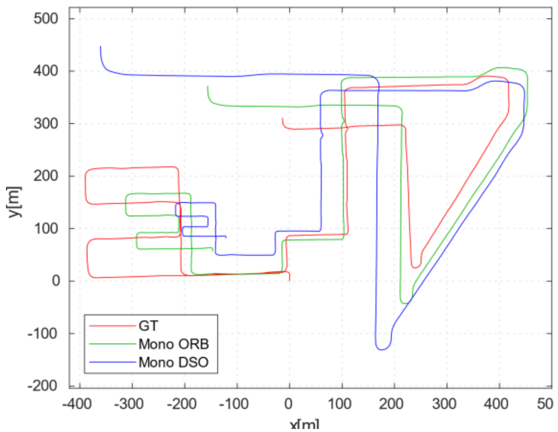
- **VO:** Determine position and orientation of camera from visual input
- **VSLAM:** Globally optimize both map and camera pose

Image source: Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry, Nan Yang, Rui Wang, Jörg Stückler, Daniel Cremers, https://arxiv.org/abs/1807.02570

Intro – Shortcomings of Existing Methods

- **Mono camera setups**: Scale Drift
  - ➢ Geometrically impossible to recover scale!





- **End-to-end trained Neural Networks**: Performance
- **Stereo Camera, RGB-D, LiDAR:** Cost & Calibration

➢ **This paper**: Hybrid method (Single camera enhanced with Deep Learning)

Image source: Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras, Rui Wang, Martin Schwörer, Daniel Cremers, https://arxiv.org/abs/1708.07878

Outline

- Introduction
- Method
  - High-Level Concept
  - Depth Estimation
  - Bundle Adjustment
  - Deep Virtual Stereo Odometry
- Experiments & Results
  - Monocular Depth Estimation
  - Monocular Visual Odometry
- Personal Comments
- Summary

Method – High-Level Concept

Deep Learning Pipeline

Mono camera image

SimpleNet → ResidualNet →

Left disparity map

Right disparity map

Consistent metric scale initialization

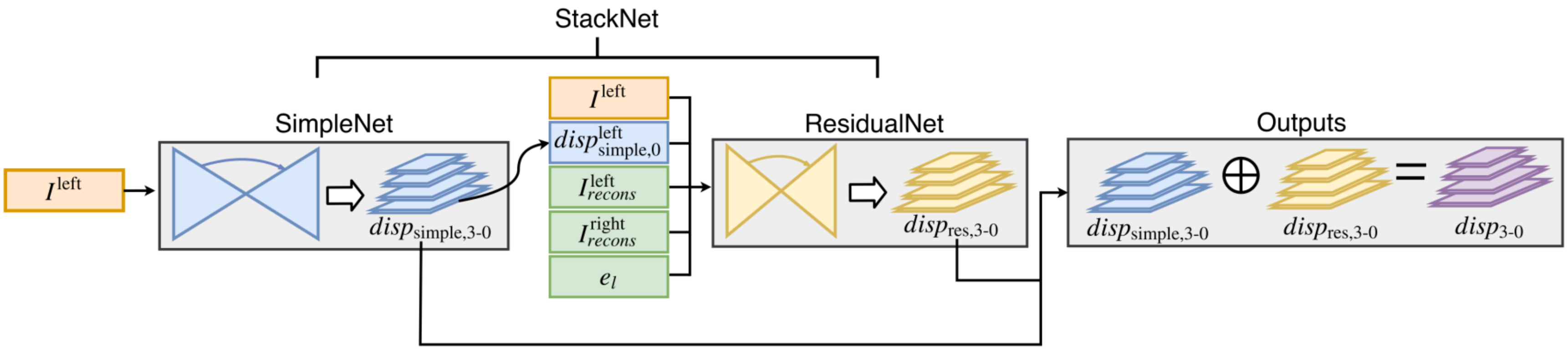
Virtual stereo photometric error

Classic Visual Odometry Pipeline

Mono camera image

Depth Initialization → Photometric Error Optimization → 3D-model and localization →

Method – Depth Estimation

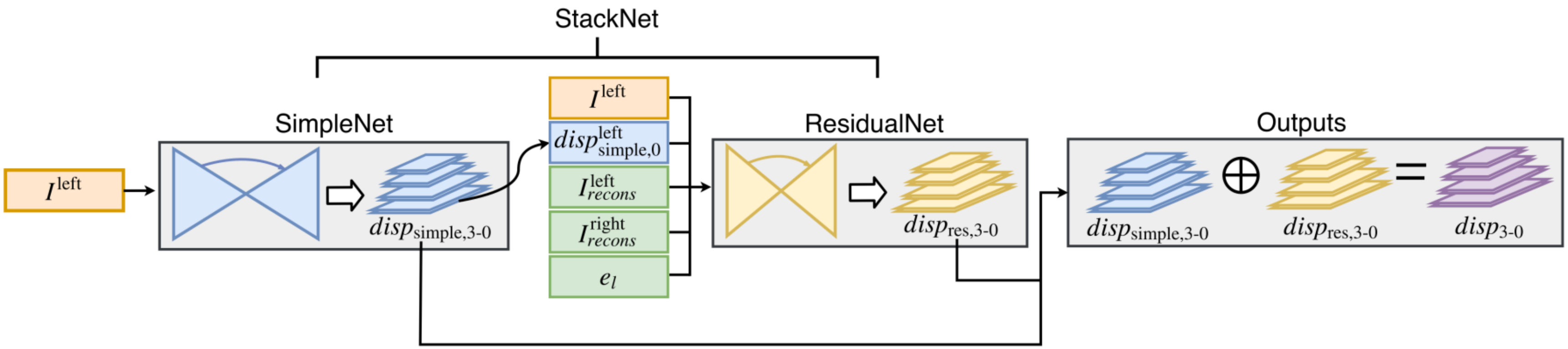


Architecture:

- 2 fully-convolutional encoder-decoder networks with skip connections
- ResidualNet refines disparity maps output by SimpleNet

Image source: Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras, Rui Wang, Martin Schwörer, Daniel Cremers, https://arxiv.org/abs/1708.07878

Method – Depth Estimation



Architecture:

- 2 fully-convolutional encoder-decoder networks with skip connections
- ResidualNet refines disparity maps output by SimpleNet

Loss Function:

$$\mathcal{L}_s = \underbrace{\alpha_U \left(\mathcal{L}_U^{left} + \mathcal{L}_U^{right}\right)}_{\text{Self-supervised loss}} + \underbrace{\alpha_S \left(\mathcal{L}_S^{left} + \mathcal{L}_S^{right}\right)}_{\text{Supervised loss}} + \underbrace{\alpha_{lr} \left(\mathcal{L}_{lr}^{left} + \mathcal{L}_{lr}^{right}\right)}_{\text{Left-right consistency loss}} + \underbrace{\alpha_{smooth} \left(\mathcal{L}_{smooth}^{left} + \mathcal{L}_{smooth}^{right}\right)}_{\text{Disparity smoothness regularization}} + \underbrace{\alpha_{occ} \left(\mathcal{L}_{occ}^{left} + \mathcal{L}_{occ}^{right}\right)}_{\text{Occlusion regularization}}$$

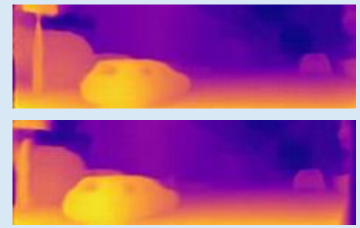
Image source: Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras, Rui Wang, Martin Schwörer, Daniel Cremers, https://arxiv.org/abs/1708.07878

Method – Deep Virtual Stereo Odometry

Deep Learning Pipeline

Mono camera image

SimpleNet → ResidualNet



Left disparity map

Right disparity map

Consistent metric scale initialization

Virtual stereo photometric error

Classic Visual Odometry Pipeline

Mono camera image

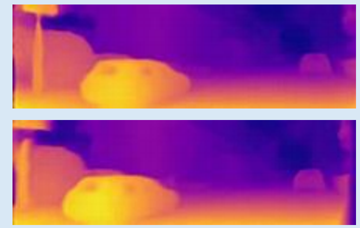
Depth Initialization → Photometric Error Optimization → 3D-model and localization

Method – Deep Virtual Stereo Odometry

Deep Learning Pipeline

Mono camera image

SimpleNet → ResidualNet



Left disparity map

Right disparity map

Consistent metric scale initialization

→ mitigates scale drift

Virtual stereo photometric error

Classic Visual Odometry Pipeline

Mono camera image

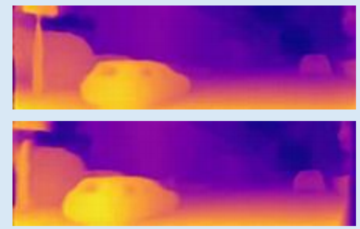
Depth Initialization → Photometric Error Optimization → 3D-model and localization

Method – Deep Virtual Stereo Odometry

Deep Learning Pipeline

Mono camera image

SimpleNet → ResidualNet



Left disparity map

Right disparity map

Consistent metric scale initialization
→ mitigates scale drift

Virtual stereo photometric error
→ increases accuracy & mitigates scale drift

Classic Visual Odometry Pipeline

Mono camera image

Depth Initialization → Photometric Error Optimization → 3D-model and localization

Method – Definitions

- **Bundle adjustment (BA):** Optimize 3D points, camera motion & camera params based on 2D images
- **Direct bundle adjustment:** No reprojection error, optimize in image space
- **Sparse bundle adjustment:** Use only a subset of pixels
- **Windowed bundle adjustment:** BA with sliding-window (heuristic keyframes)

➢ **Here: Windowed direct sparse BA**

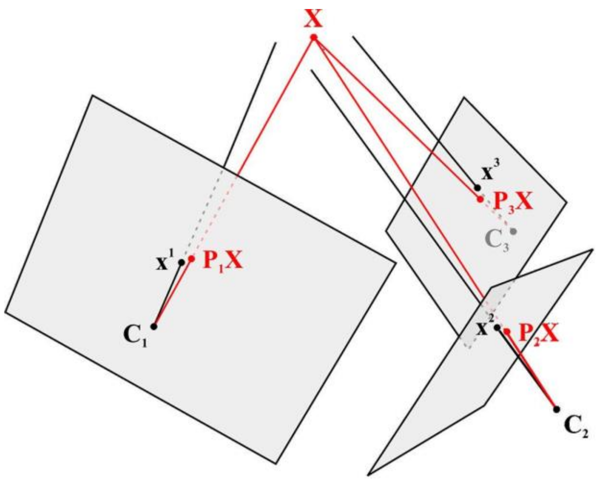
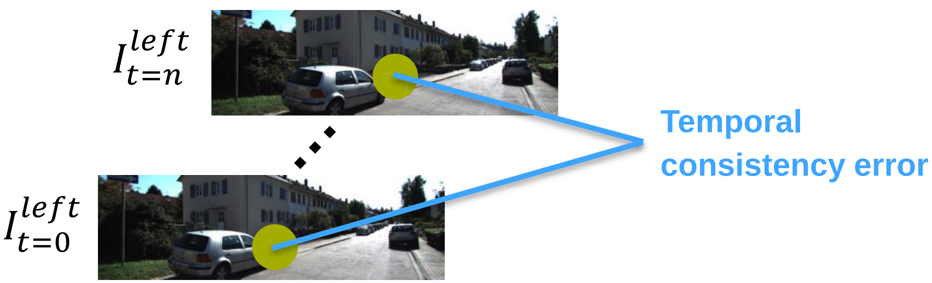
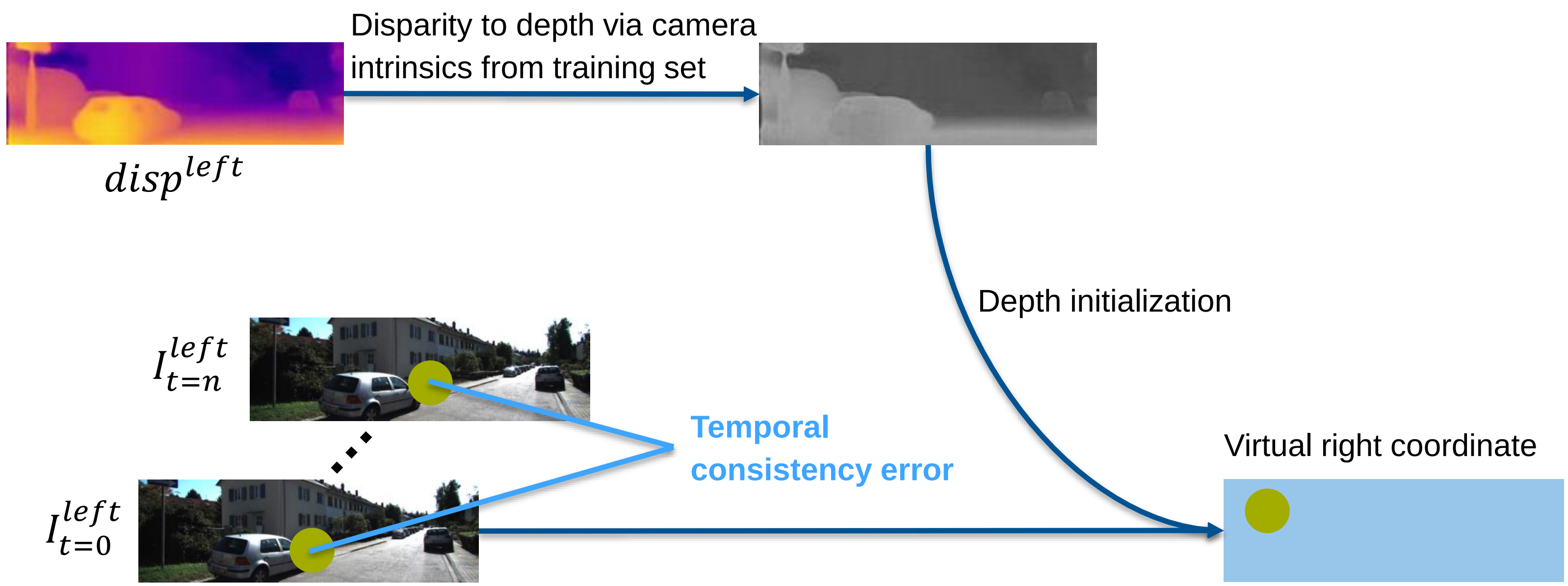


Image source: http://docplayer.net/52196787-Seminar-bundle-adjustment.html

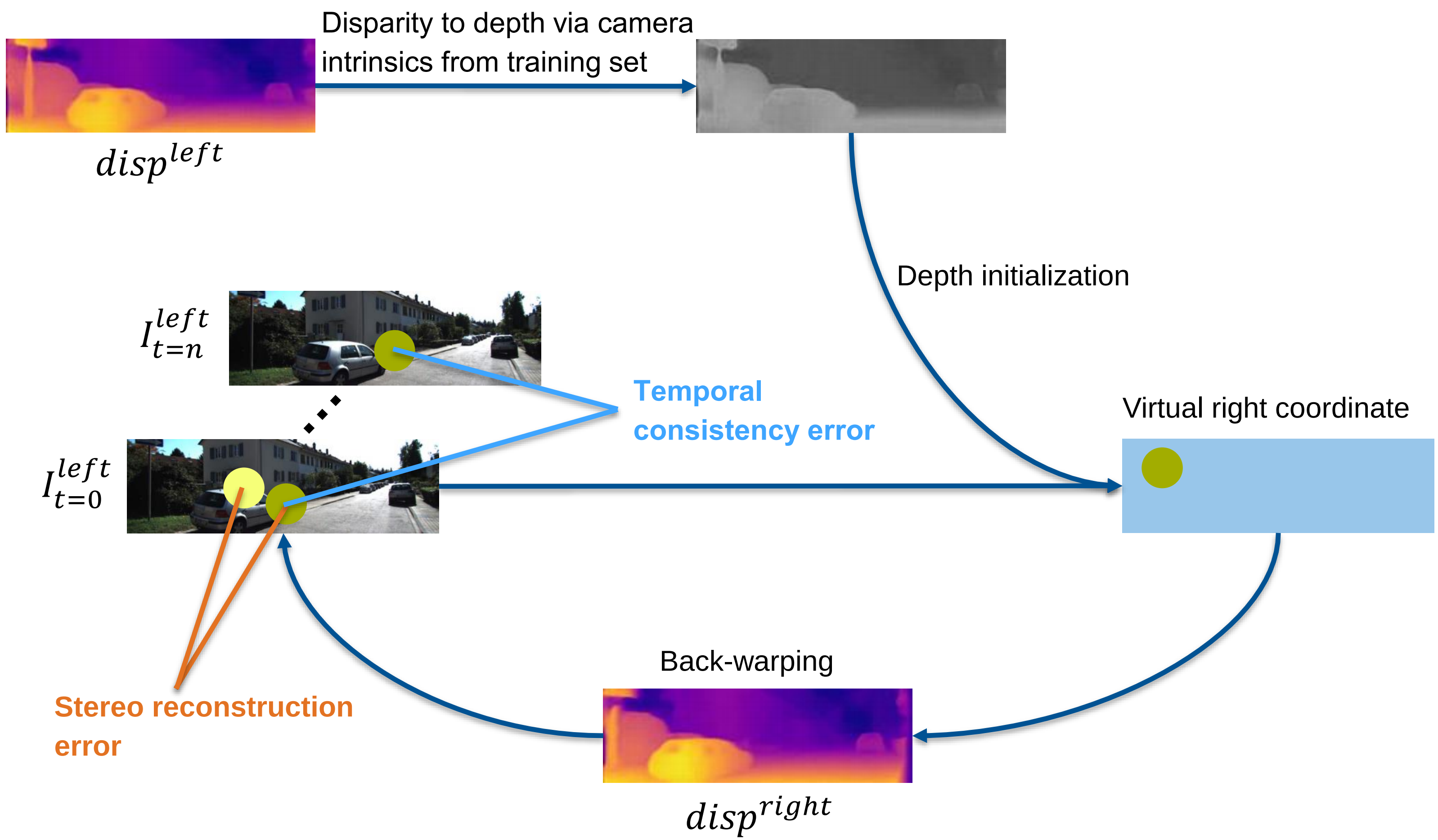
Method – Virtual Stereo Photometric Error



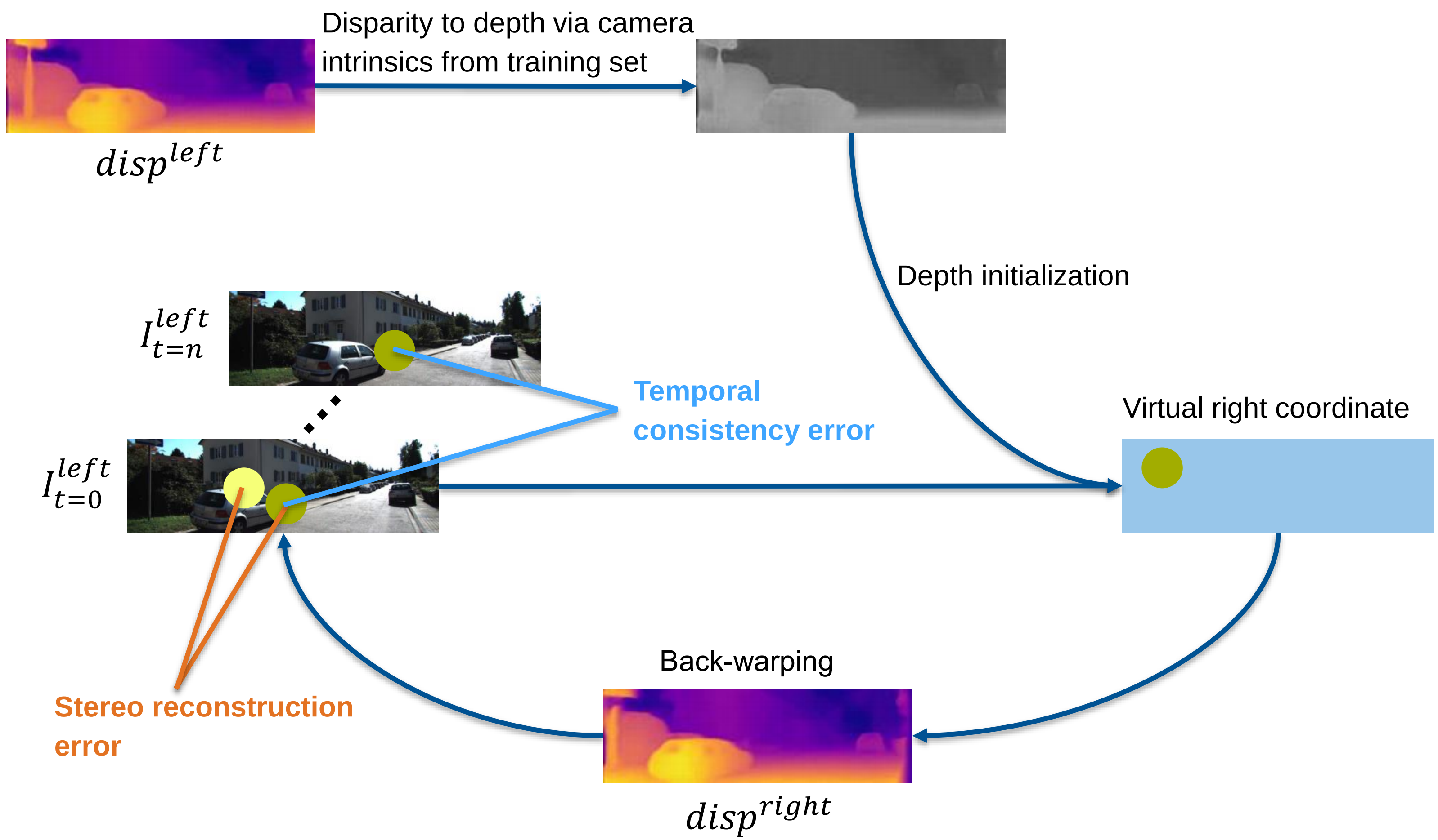
Method – Virtual Stereo Photometric Error



Method – Virtual Stereo Photometric Error



Method – Virtual Stereo Photometric Error



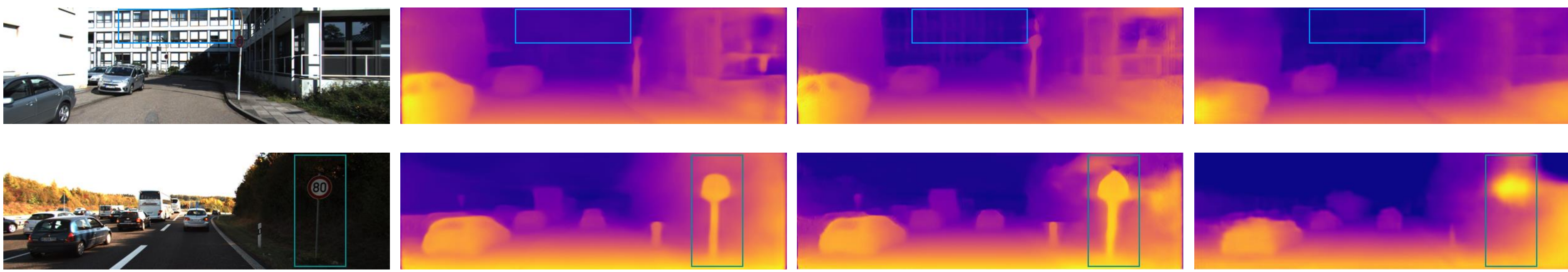
3D-points and camera motion are recovered by jointly optimizing temporal consistency and stereo reconstruction error using Gauss-Newton

Experiments & Results – Depth Estimation

Training schedule („*easy task first, then harder task, then easy again to smooth outliers*“):

1. Train SimpleNet: semi-supervised → self-supervised → semi-supervised
2. Freeze SimpleNet, train ResidualNet: semi-supervised → self-supervised → semi-supervised

Results:



Input | This paper | Self-supervised [1] | Semi-supervised by LiDAR [2]

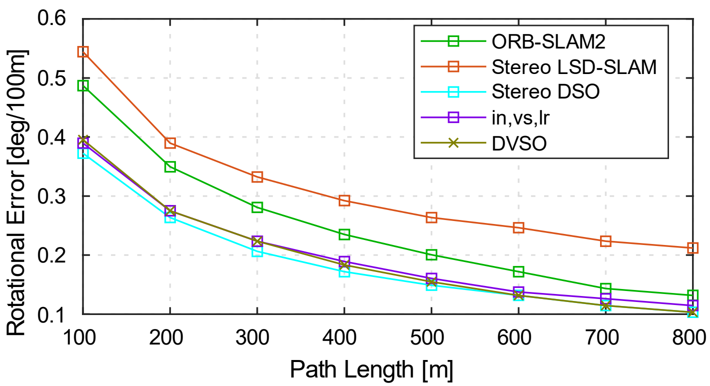
Source:
[1] Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. arXiv preprint arXiv:1609.03677 (2016)
[2] Kuznietsov, Y., St¨uckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. arXiv preprint arXiv:1702.02706 (2017)

Experiments & Results – Monocular Visual Odometry

Ablation study:

1. Initializing depth with left disparity prediction
2. Using right disparity for virtual stereo term in windowed bundle adjustment
3. Checking left-right disparity consistency for sparse point selection
4. Tuning virtual stereo baseline

Results:



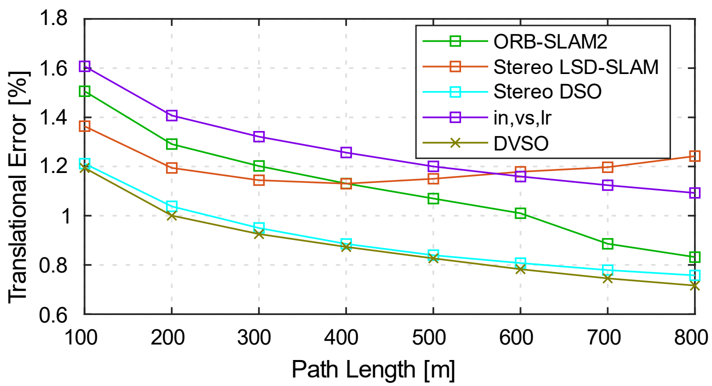
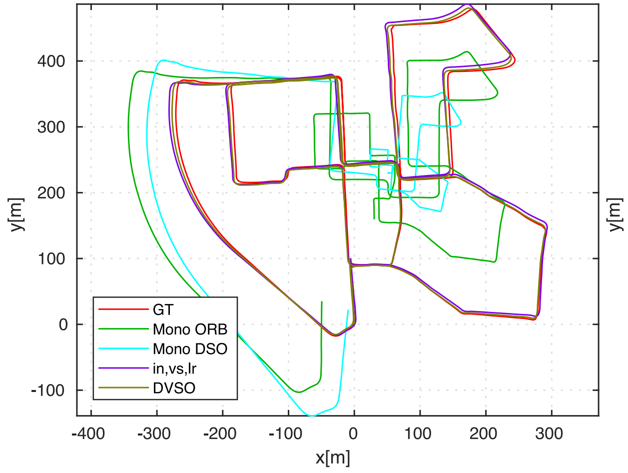
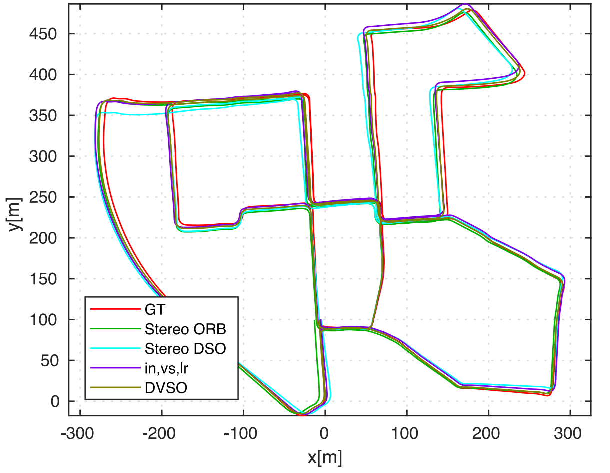
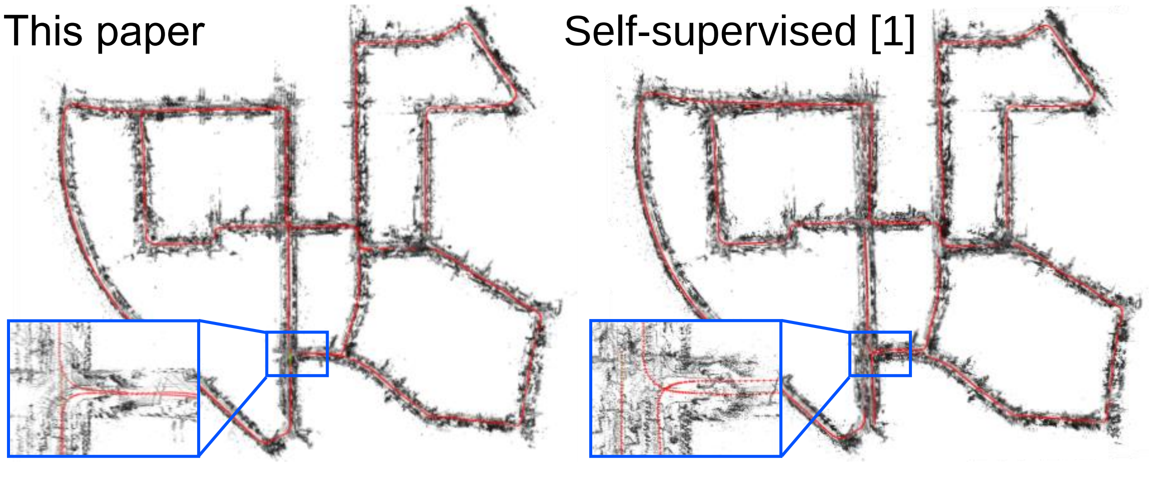


Image source: Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras, Rui Wang, Martin Schwörer, Daniel Cremers, https://arxiv.org/abs/1708.07878

Experiments & Results – Monocular Visual Odometry







Source:
[1] Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. arXiv preprint arXiv:1609.03677 (2016)

Personal Comments

Baseline tuning is cheating

- Stereo information at test time required
- Future work: Online fine tuning

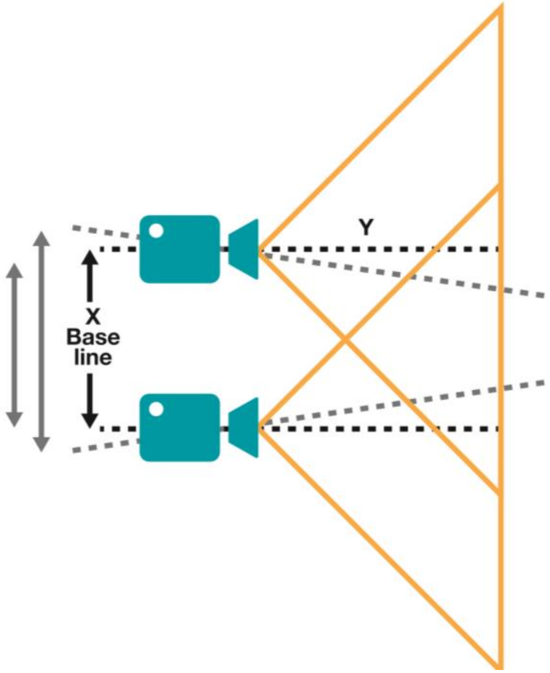


Image source: https://www.framos.com/en/news/stereo-applications-need-a-dedicated-lens-choosing

Personal Comments

Necessity for Deep Learning

- Stereo cameras cheap
- Deep Learning needs good training data
- Safety & error accountability

Image source: https://www.bosch-presse.de/pressportal/de/en/the-bosch-stereo-video-camera-provides-important-optical-information-about-the-vehicles-surroundings-121096.html

Personal Comments

Enables fleet learning

- Few vehicles with stereo camera to provide training data
- Remaining vehicles use DL and operate in a similar enough domain

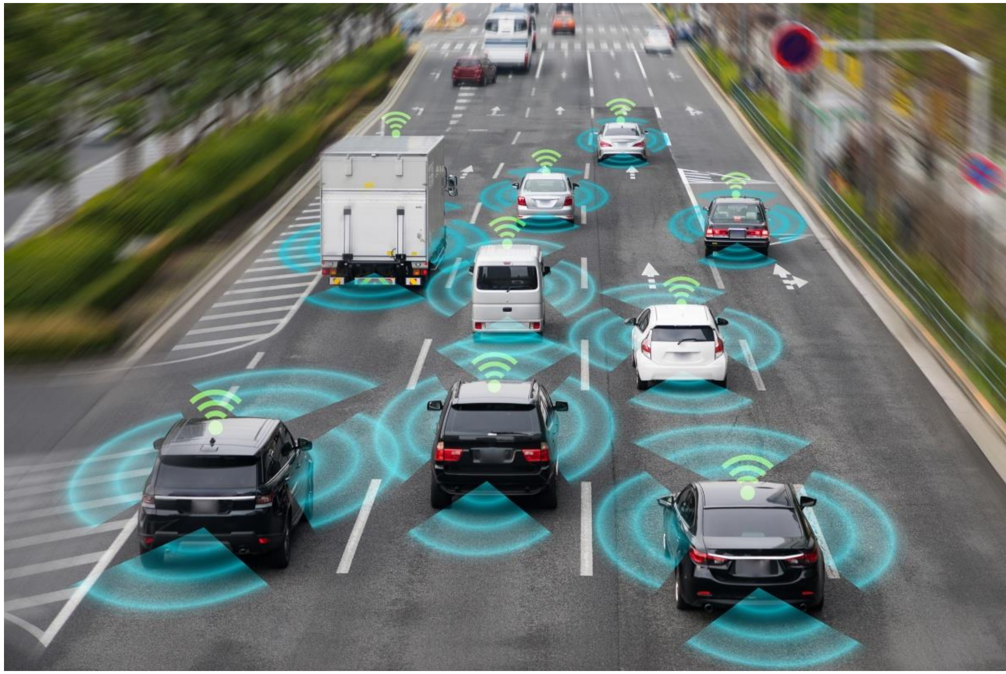


Image source: https://www.smartcitiesworld.net/news/news/oxbotica-and-cisco-partner-to-resolve-autonomous-fleet-data-challenges-5074

Summary: Deep Virtual Stereo Odometry

- **Monocular Input + Deep Depth Predictions → Stereo Performance**
  - Scale drift eliminated
- **Semi-Supervised Disparity Predictions**
  - Stereo reconstruction for self-supervision
  - Sparse DSO for supervision
- **Incorporation into Visual Odometry Pipeline**
  - Metric scale initialization
  - Virtual stereo error

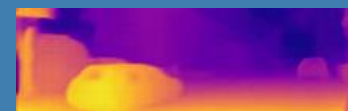
Discussion

Deep Learning Pipeline

Mono camera image

SimpleNet → ResidualNet

Left disparity map



Right disparity map

Consistent metric scale initialization

Virtual stereo photometric error

Classic Visual Odometry Pipeline

Mono camera image

Depth Initialization → Photometric Error Optimization → 3D-model and localization

Backup: From Disparity to Depth

$$d(\vec{p}) = \frac{b \cdot f_x}{disp(\vec{p})}$$

- $d(\vec{p})$: Depth at point **p**
- b: Baseline
- f_x: Focal length (camera intrinsic)
- $disp(\vec{p})$: Disparity at point **p**

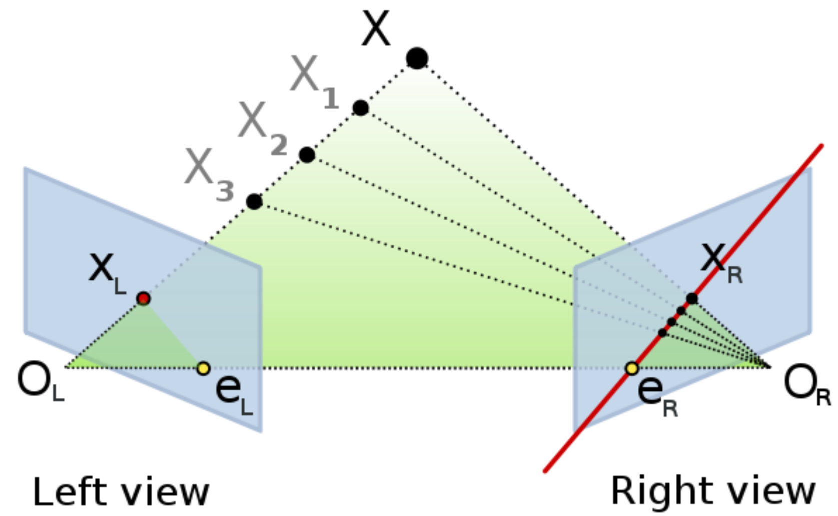


Image source: https://stackoverflow.com/questions/17607312/what-is-the-difference-between-a-disparity-map-and-a-disparity-image-in-stereo-m