Computer Vision for Autonomous Navigation:

Self-supervised Joint Learning of Depth, Optical Flow and Semantic Contours

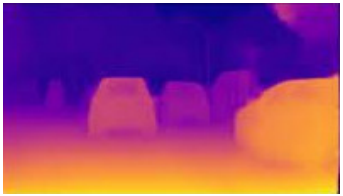
Antoine Manzanera

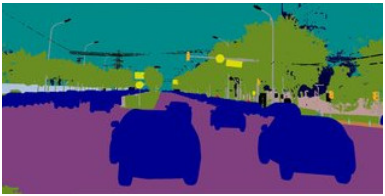
ENSTA Paris

Workshop USP/IPP 2025
São Paulo, Brazil
March 2025

Context: Computer Vision for Autonomous Navigation

The objective of this research is to allow a mobile robot to *learn autonomously* the vision tasks that are fundamental for its navigation:





- DEPTH MAPS: Get a 3d map of its environment, find navigable paths, avoid obstacles...
- OPTICAL FLOW: Recover ego-motion (Odometry), localise moving objects,...
- SEMANTIC MAPS: Recognise things (Vehicles, Humans, Buildings,...) and stuffs (Road, Sky, Forest,...)

Context: Computer Vision for Autonomous Navigation

We consider the *monocular* case (one single camera). We aim to design a *fully self-supervised* deep learning approach, that can be naturally extended towards continuous learning, which supposes:

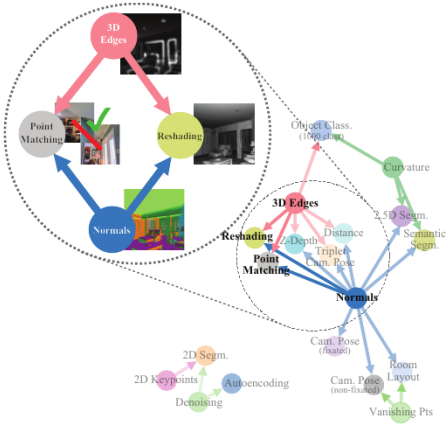
- Zero annotated data
- Zero pre-trained module
- Adaptation and Consolidation mechanisms (out of scope here)

Presentation Outline

Presentation Outline

Synergies between Tasks in Computer Vision

Positive interferences between different computer vision tasks have been identified for years, by demonstrating strong *transfer learning* capacities from some neural networks trained on a specific task, to be retrained to perform a different task.



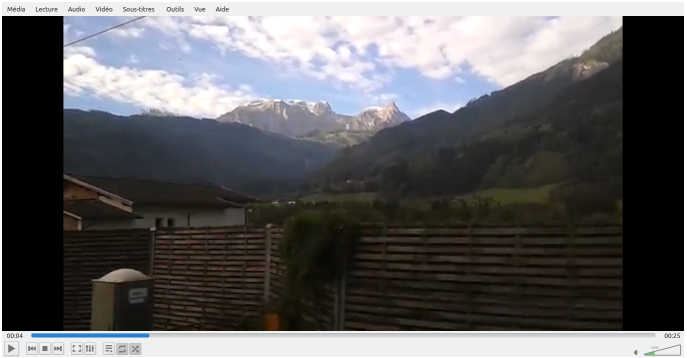
Taskonomy **[Zamir 2018]**

Transfer: Motion (Optical Flow) to Depth (Disparity)

Here, an horizontal travelling:

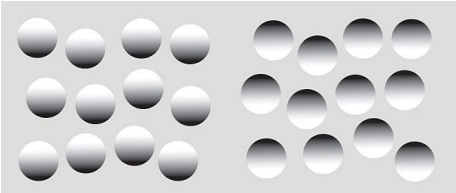
$$Z = \frac{f\dot{X}}{\dot{x}}, \text{ with}$$

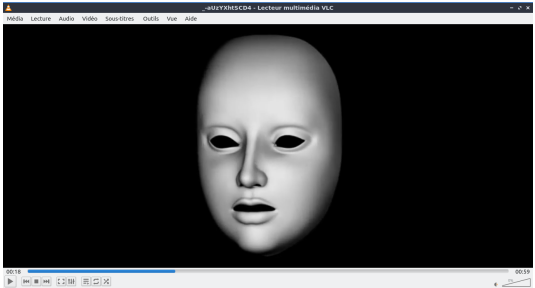
- Z the depth
- f the focal distance
- $\dot{X}$ the camera velocity (constant)
- $\dot{x}$ the apparent (pixelwise) velocity (variable)



Transfer: Shading (Normals) to Depth

Self shadowing is a strong depth cue. Without shape prior, the sense of the normal (concavity) is determined by a prior on light direction (right image).





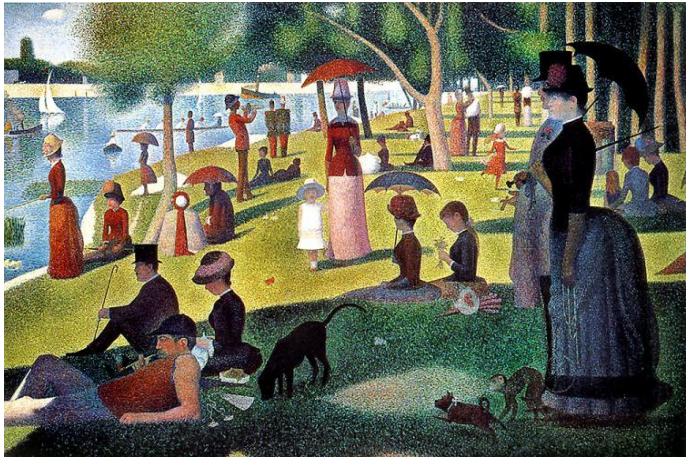
However, when the shape prior is strong, the semantic prior dominates the lighting prior (top-down effect, animation on the left).

Transfer Semantics to Depth: Occlusions

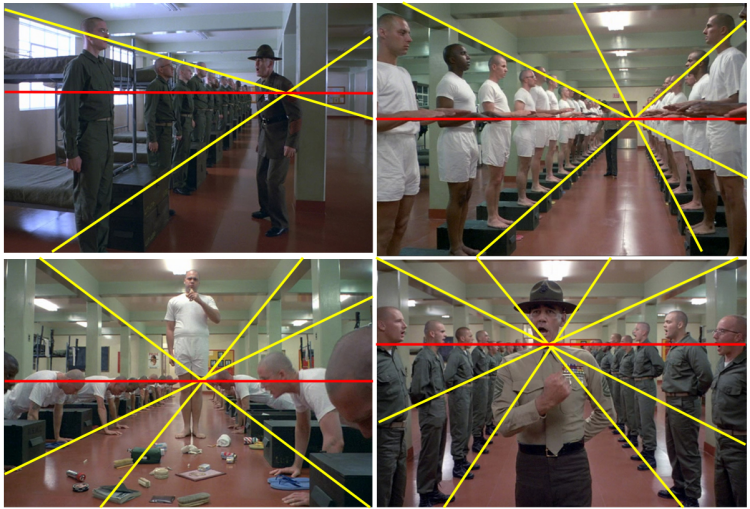
Giotto - Pentecoste
(*circa* 1305)

Transfer Semantics to Depth: Object sizes

Georges Seurat - Un après-midi à l'île de la Grande Jatte (1884-1886)



Semantics, Depth and Odometry (Pose): Vanishing point and Horizon



Stanley Kubrick – Full Metal Jacket (1987)

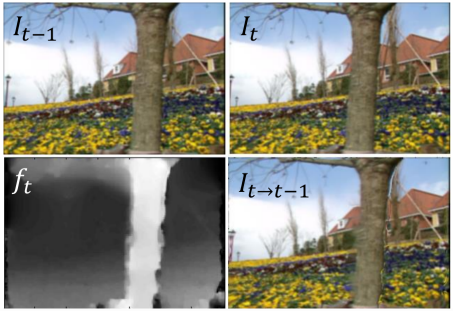
Presentation Outline

Self-supervised learning

Supervision of depth, motion or semantic maps assumes densely annotated data, which are hard to obtain for real images. Self-supervision is possible using an *auxiliary* task, that the model can perform itself to produce an alternate supervision signal.

- **Optical Flow**: Photometric loss based on straightforward image warping.
- **Depth maps**: Photometric loss based on back-projection and reprojection.
- **Semantic maps**: No reported auxiliary task; we propose to use physical cues from motion and depth to predict pre-semantic maps.

Self-supervised Optical Flow: Photometric loss



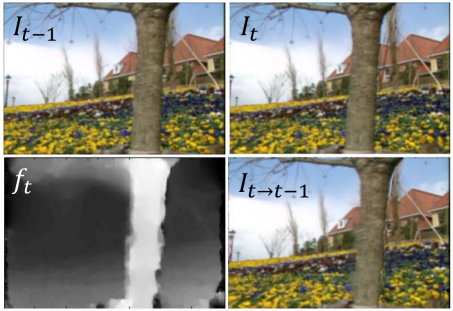
Self-supervised Optical Flow is based on photometric loss, that measures the difference between an image and its prediction based on the optical flow:

$$\mathcal{L}_{\text{photo}}^{\text{flow}} = \|I_1 - I_{0\to 1}\|,$$

with:

$$I_{0\to 1}(\mathbf{m}) = I_0\left(\mathbf{m} - f_{0\to 1}(\mathbf{m})\right).$$

Self-supervised Optical Flow: Limitations

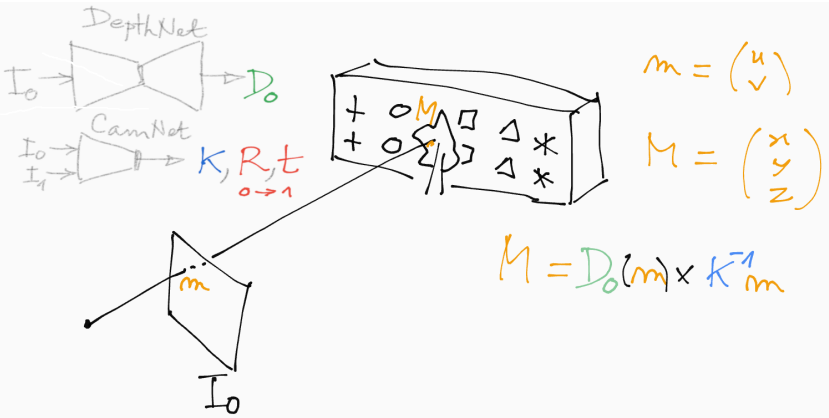


Two major issues hamper the efficiency of Photometric loss:

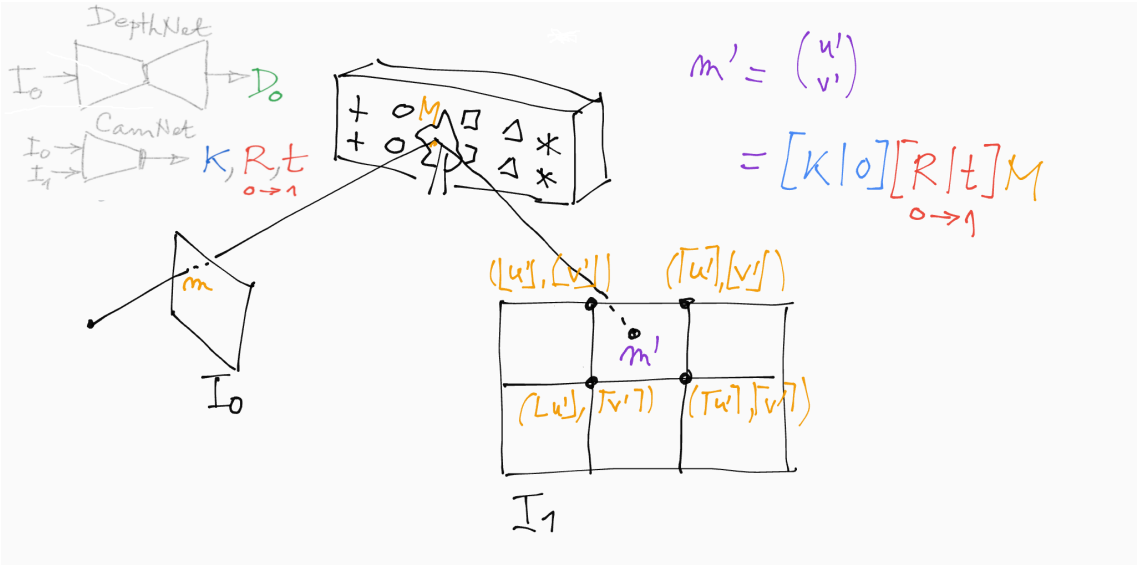
- **Occlusions**: Missing pixel values!
- **Homogeneous zones**: Small loss does not mean small errors!

Solution: Semantic cues!

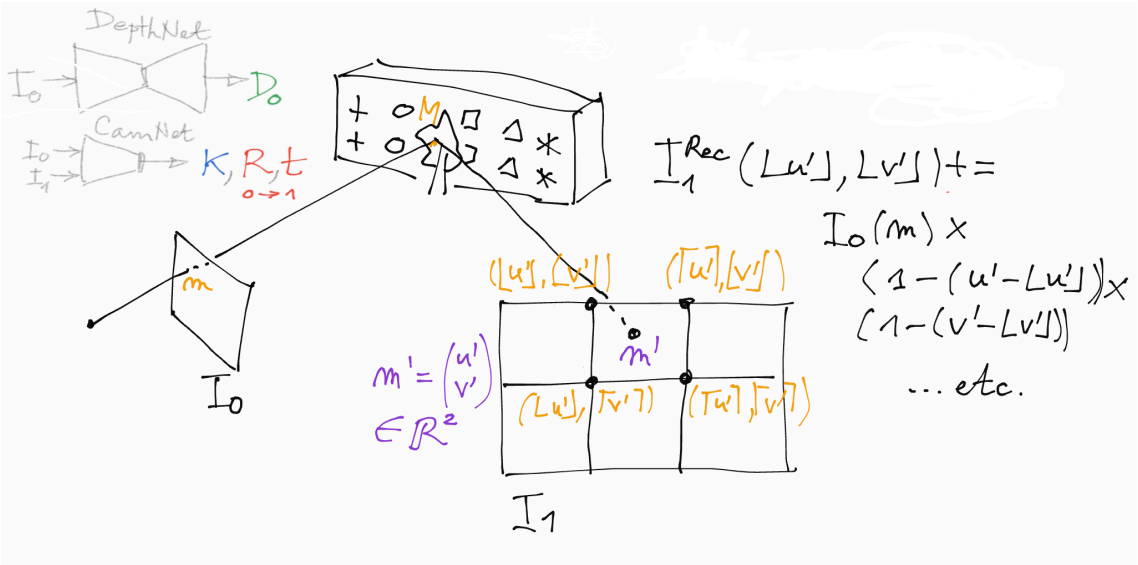
Self-supervised Depth (1): Back-projection from first image



Self-supervised Depth (2): Re-projection onto second image



Self-supervised Depth (3): Interpolation within second image



Self-supervised Depth: Summary and formula

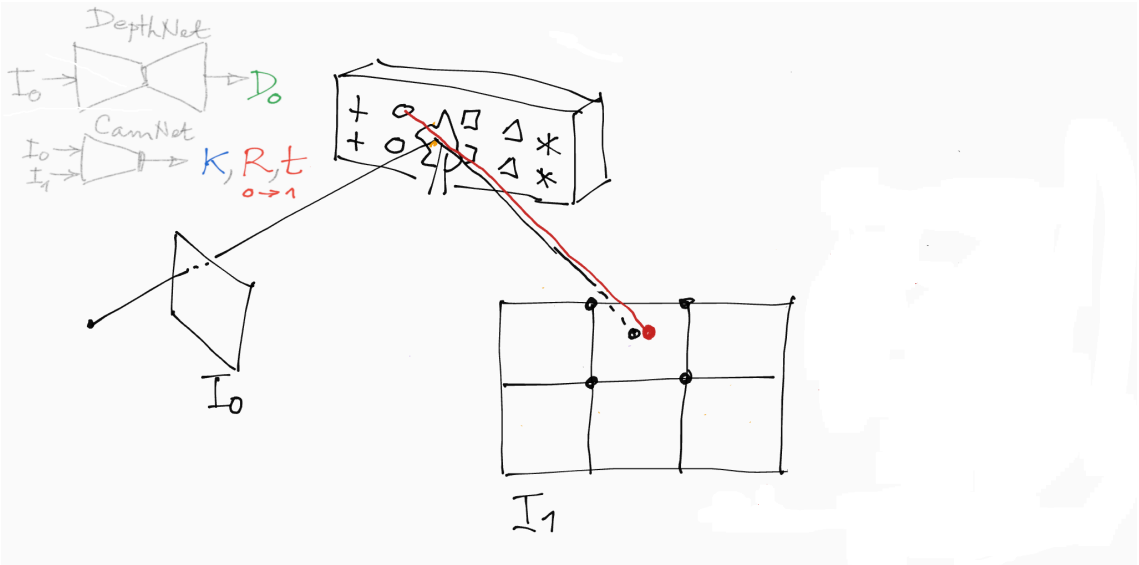
The photometric loss provides a self-supervision signal by comparing the observed image with the reconstructed image from the previous view is the depth map and odometry have been well predicted:

$$\begin{aligned}\mathcal{L}_{\text{photo}}^{\text{depth,odometry}} &= \|I_1 - I_1^{\text{Rec}}\| \\ &= \sum_{\mathbf{m}'} \left(I_1(\mathbf{m}') - I_0(\mathbf{m})\right)^2, \text{ with } \mathbf{m}' \simeq \left([\mathbf{K}|\mathbf{O}_4]\,[\mathbf{R}|\mathbf{t}]\,D_0(\mathbf{m}) \times \mathbf{K}^{-1}\mathbf{m}\right)\end{aligned}$$

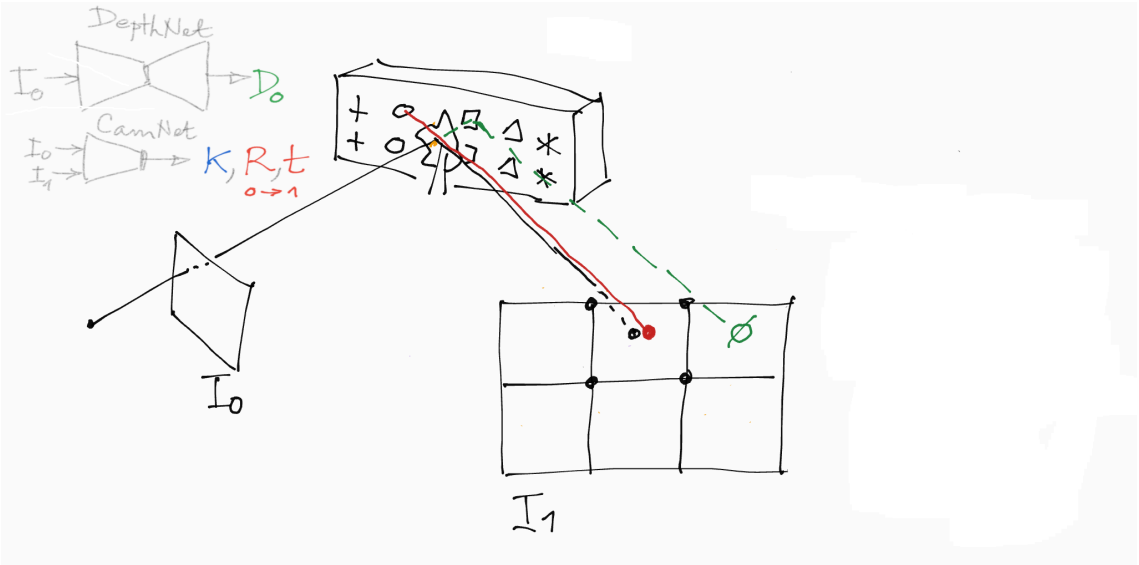
But: Even more problems than with optical flow:

- Occlusions!
- Homogeneous zones!
- Moving objects!

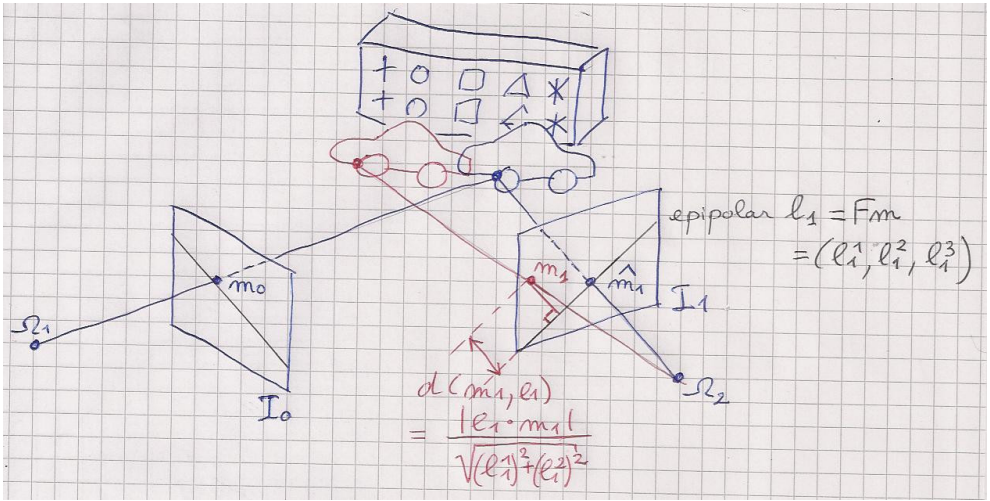
Self-supervised Depth: Occlusion issue



Self-supervised Depth: Un-occlusion issue

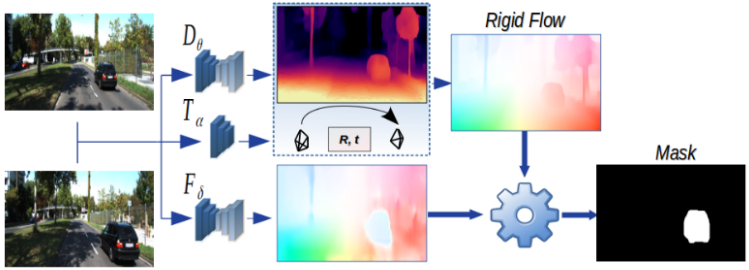


Self-supervised Depth: Moving objects issue



Presentation Outline

CoopNet: Joint training of Optical Flow, Odometry and Depth

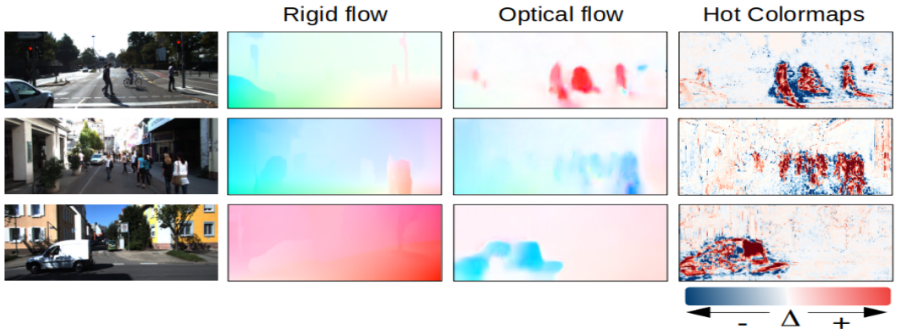


CoopNet **[Hariat WACV'23]**

By estimating (or predicting) the optical flow, moving objects can also be predicted by comparing the optical flow with the *rigid flow*, which is the apparent velocity field under rigid assumption scene (i.e. only due to camera motion), defined as:

$$[\mathbf{K}|\mathbf{O}_4]\,[\mathbf{R}|\mathbf{t}]\,D_0(\mathbf{m}) \times \mathbf{K}^{-1}\mathbf{m} - \mathbf{m}$$

CoopNet: Joint training of Optical Flow, Odometry and Depth



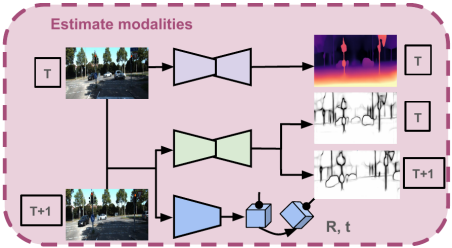
CoopNet **[Hariat WACV'23]**

The CoopNet network is trained based on the difference between the photometric losses from the optical flow and from the depth networks:

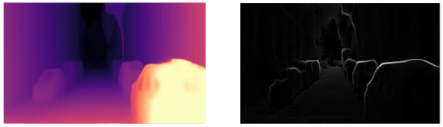
$$\Delta(\mathbf{m}) = \mathcal{L}_{\text{photo}}^{\text{depth,odometry}} - \mathcal{L}_{\text{photo}}^{\text{flow}}$$

Now, back to Semantics!

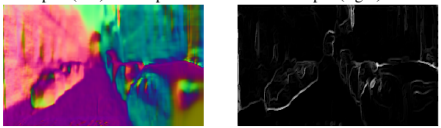
Depth and Optical Flow, both learned in a self supervised way, actually provide *physical cues* to separate objects or surfaces (Pre-semantic maps):



CoopNet2 **[Hariat CVPR'25]**



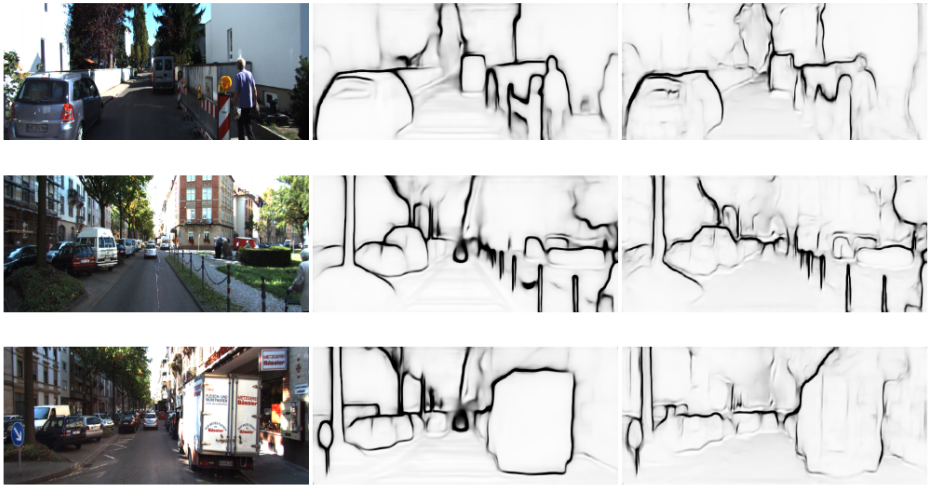
depth (left) and Laplacian activation of depth (right).



normal (left) and gradient activation of normal (right).

The loss function for the edge map is designed to promote edges around the inflexion points (2nd derivative of depth maps) and the orientation changes (1st derivative of the normal maps) of the surfaces.

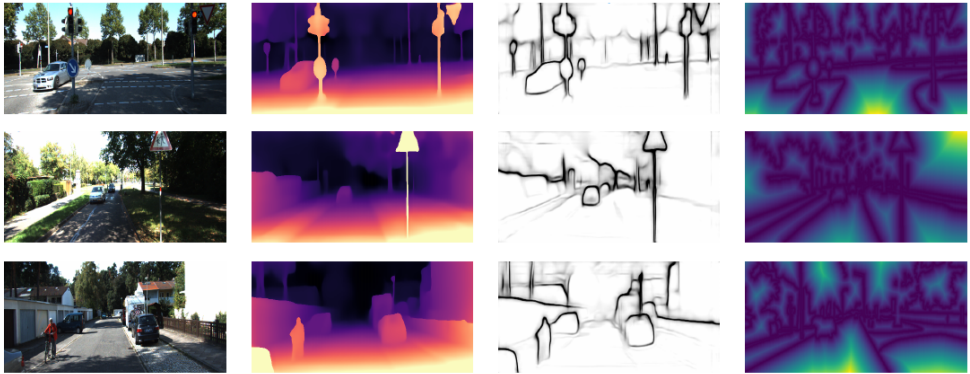
Comparative results for Contours



Center column: our work **[Hariat CVPR'25]** compared with Lego **[Yang 2018]**

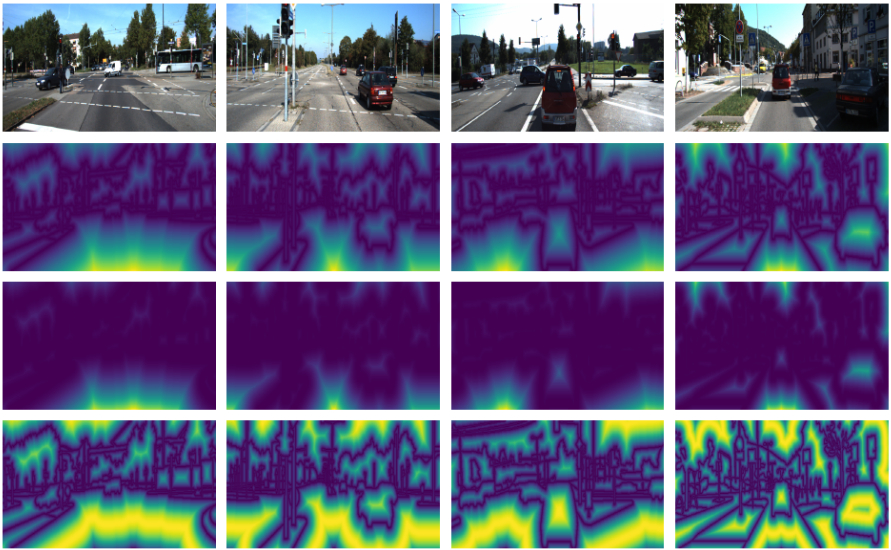
Now, what about homogeneous areas?

The post-processed edge maps provide contours which are used to calculate distance transform maps, that are combined with RGB images to *add structure* within the homogeneous areas.



CoopNet2 **[Hariat CVPR'25]**

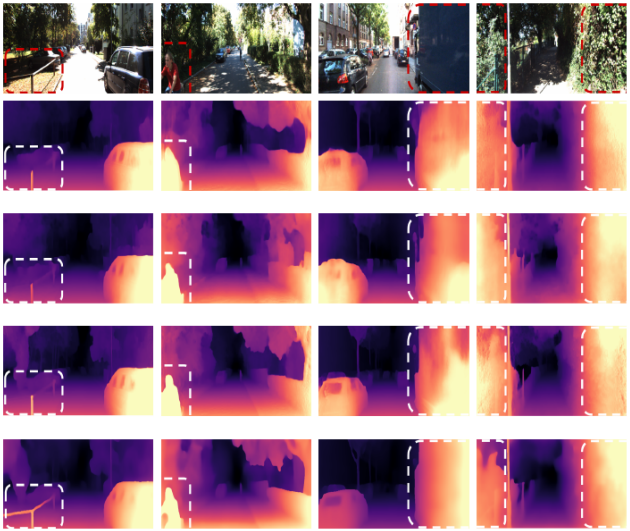
Variations on Eikonal



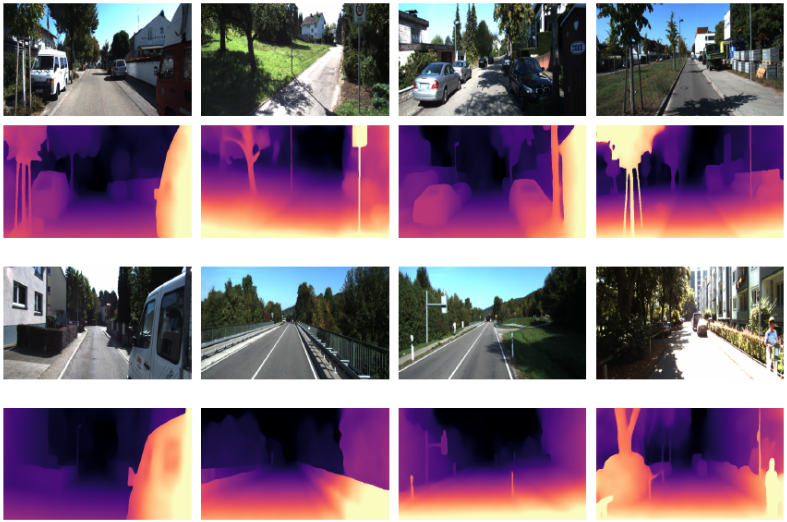
Comparative results for Depth maps

Our work **[Hariat CVPR'25]** (last row) compared with competitors.

- column 1: thin object
- column 2: moving object
- column 3: large untextured area
- column 4: complex shape



More qualitative results



[Hariat CVPR'25]

Presentation Outline

Conclusion and Future Works

Contributions

- Fully self-supervised joint learning of Depth, Odometry, Optical Flow and Pre-semantic Contours.
- Eikonal-augmented images for enhanced photometric loss.
- Optimality results on distance transforms.
- n-d extensions of distance maps based on random walks.

Future Works

- Extension of the framework to continuous learning.
- Extension of the photometric loss to non Lambertian models.

References (1)

[Zamir 2018] A.R. Zamir and A. Sax and W.B. Shen and L.J. Guibas and J. Malik and S. Savarese
Taskonomy: Disentangling Task Transfer Learning
Computer Vision and Pattern Recognition (CVPR), 2018.

[Zhou 2017] T. Zhou and M. Brown and N. Snavely and D.G. Lowe
Unsupervised learning of depth and ego-motion from video
Computer Vision and Pattern Recognition (CVPR), 2017.

[Pinard 2018] C. Pinard and L. Chevalley and A. Manzanera and D. Filliat
Learning structure-from-motion from motion
European Conf. on Computer Vision Workshops (ECCV-W), pp.363-376, 2018

References (2)

[Hariat 2023] M. Hariat and A. Manzanera and D. Filliat
Rebalancing gradient to improve self-supervised co-training of depth, odometry and optical flow predictions.
Winter Conference on Applications of Computer Vision (WACV), 2023.

[Yang 2018] Z. Yang and P. Wang and Y. Wang and W. Xu and R. Nevatia
LEGO: Learning Edge With Geometry All at Once by Watching Videos
Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[Hariat 2025] M. Hariat and A. Manzanera and D. Filliat
Self-supervised depth estimation using distance transform over pre-semantic contours
Computer Vision and Pattern Recognition (CVPR), 2025.