3d Reconstruction: Learning based methods

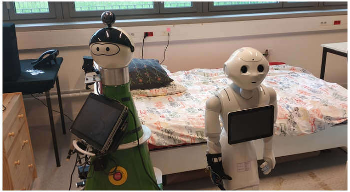
Antoine Manzanera

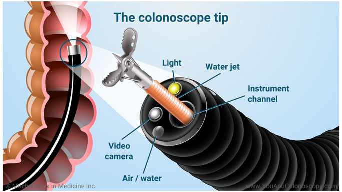
ENSTA Paris

ROB317 - 3d Computer Vision
September 2023

3d Reconstruction from Videos

Reconstructing the scene geometry from videos is useful in many applications: Robot navigation (obstacle detection), Metrology, 3d Cartography, Medicine...



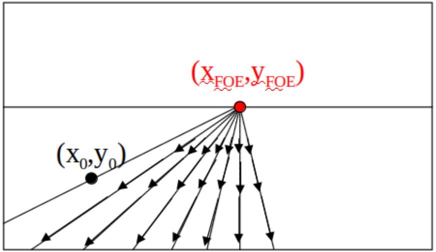


+ It is a cheap and flexible approach: One single passive camera, Adaptive baseline,...
− It strongly relies on scene structure (texture) and precise camera positioning.

Presentation Outline

Preamble: Limitations of analytical methods

- Estimation strongly relies on local structure (texture), then depth estimation on textureless areas depends on complicated regularization methods.
- Depth calculation depends on the apparent displacement (speed) of a point with respect to the epipole (i.e. the Focus of Expansion FoE, that indicates the translation direction of the camera). Such calculation turns undetermined when the point gets close to the FoE.



Presentation Outline

DNN for 3d reconstruction

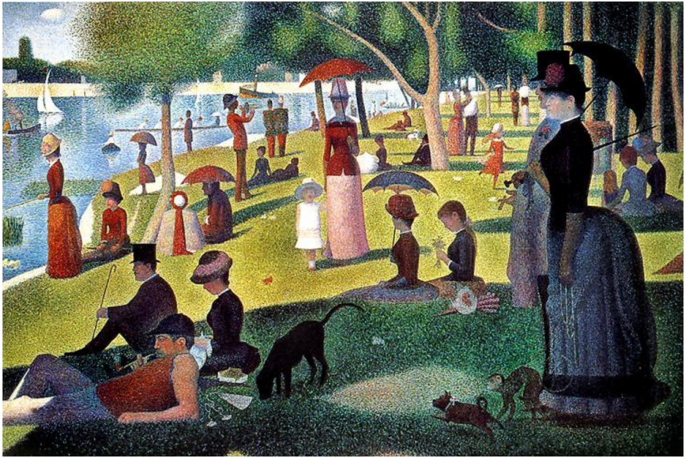
- Like Optical Flow, Depth can benefit from Deep Networks dense prediction capabilities.
- Training can be easily done on *synthetic* or *real RGB-d* data, and loss function is also relatively straightforward.
- One determining benefit of DNN is their ability to exploit potentially *all the depth indices:* parallax, perspective, size and texture gradients, shading,...

Monocular Depth Cues? Occlusions!

Giotto - Pentecoste
(*circa* 1305)

Monocular Depth Cues? Object sizes!

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



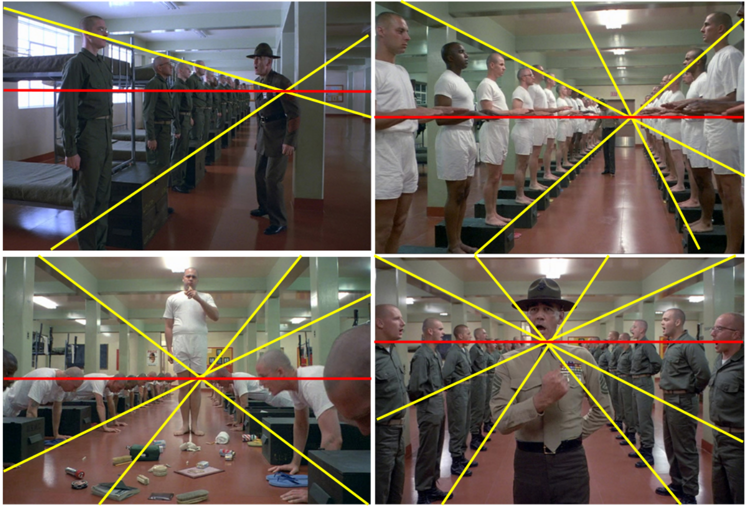
Monocular Depth Cues? Object sizes, Perspective, and Texture Gradients!

Gustave Caillebotte - Rue de Paris, temps de pluie (1877)

Monocular Depth Cues? Perspective, Horizon and Vanishing Points!

Gustave Caillebotte - Rue de Paris, temps de pluie (1877)

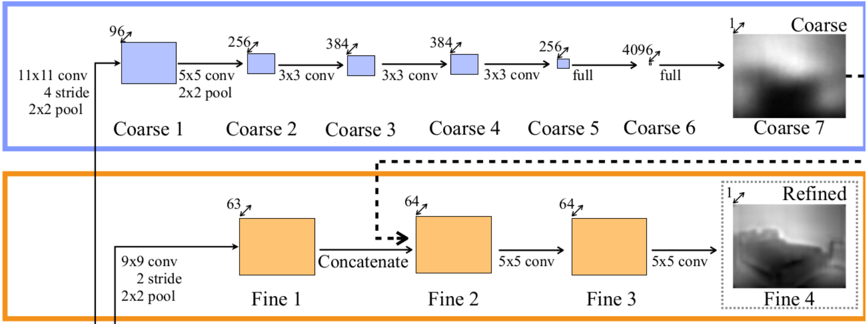
Monocular Depth Cues? Horizon and Camera Pose!



Stanley Kubrick – Full Metal Jacket (1987)

Presentation Outline

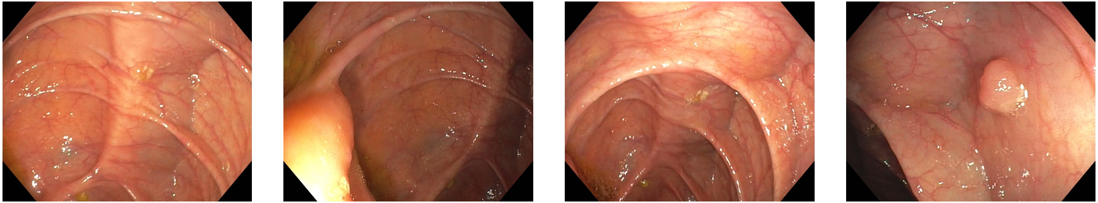
Depth inference from single view!



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

CNN based Depth estimation from single view **[Eigen 14]** works well on a particular context!

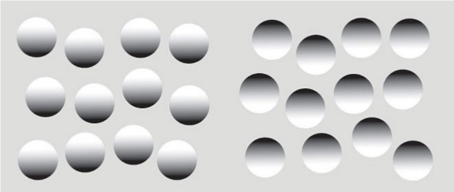
One very particular context...

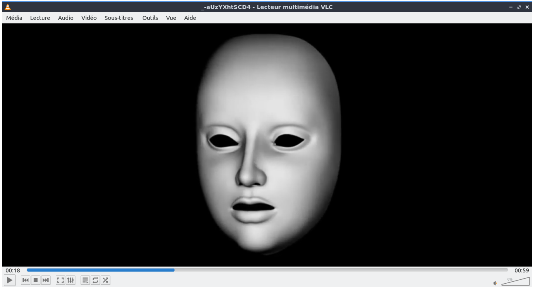


Colonoscopy images **[Ruano 19]**

Monocular Depth Cues? Shading!

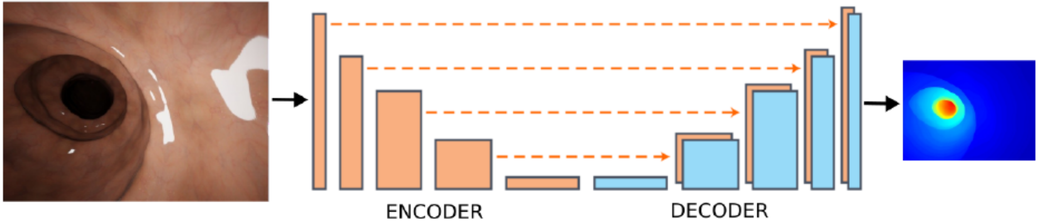
Self shadowing is a strong but ambiguous depth cue (light source position *vs* concavity). Without shape prior, the concavity is determined by a prior of top lighting (right image).





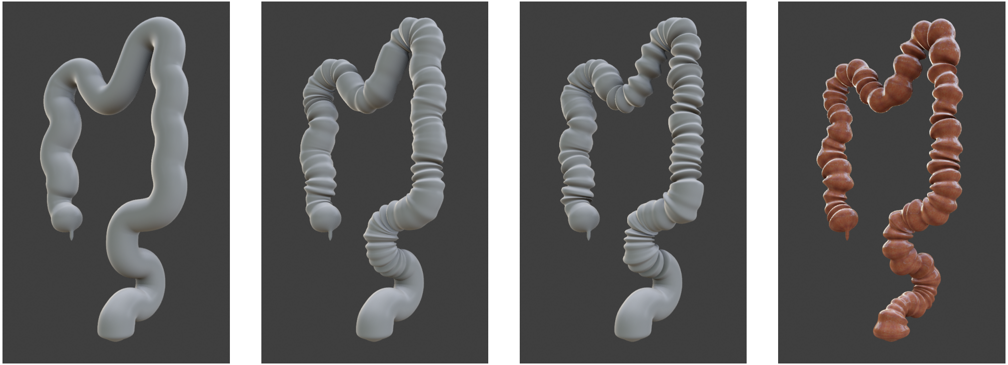
When the shape prior is strong (face then convex), the concavity prior dominates the lighting prior (top-down effect, animation on the left).

Learning Shape from Shading for Automated Colonoscopy



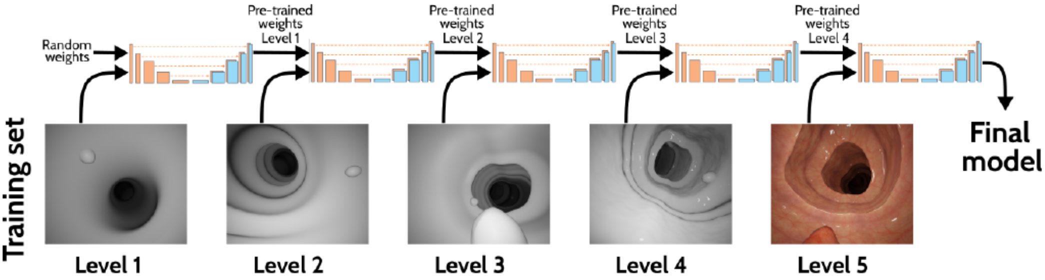
Images from synthetic videos are used to train a CNN using a loss function based on the ground truth depthmap **[Ruano 23]**

Curriculum Learning Shape from Shading for Automated Colonoscopy



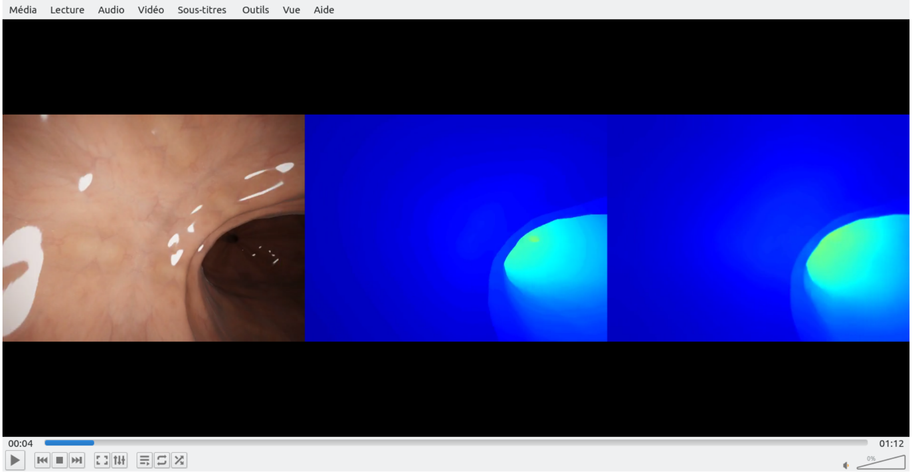
Synthetic exploration videos are created from a hierarchy of synthetic colons of increasing complexity **[Ruano 23]**

Curriculum Learning Shape from Shading for Automated Colonoscopy



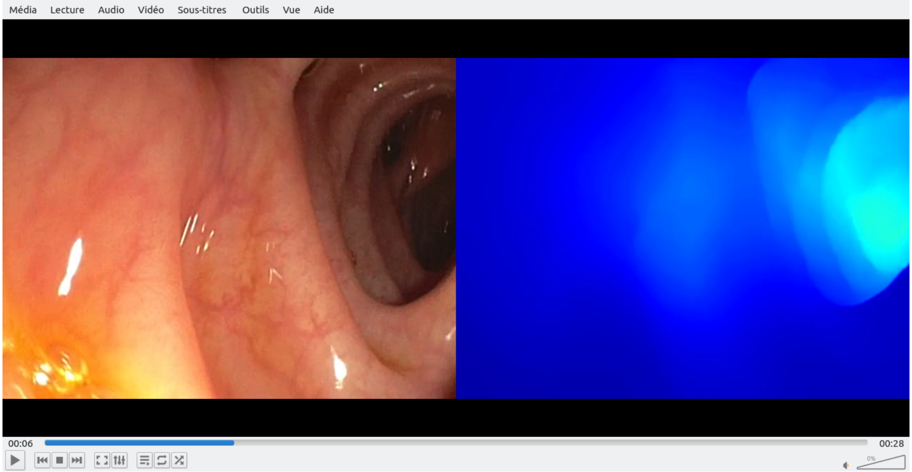
The training is performed with progressive complexity **[Ruano 23]**

SfSNet on Synthetic Videos



ShapeFromShadingNet on Synthetic Test Videos **[Ruano 23]**

SfSNet on Real Videos



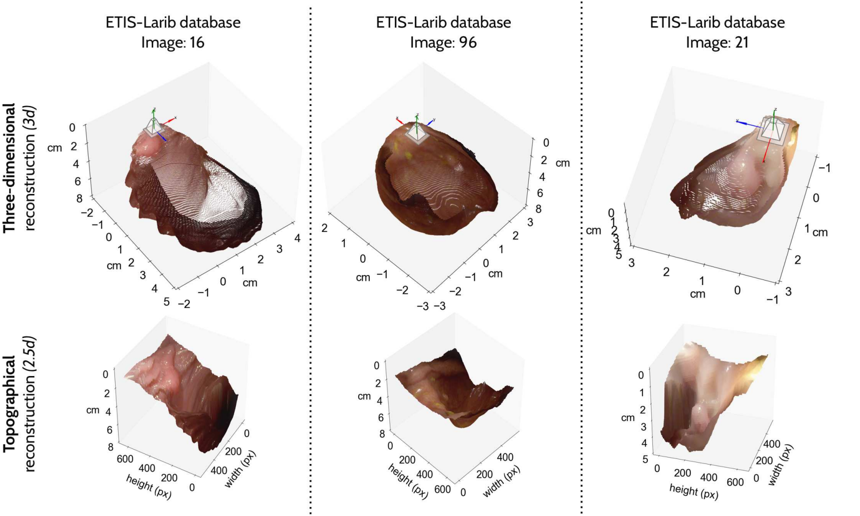
ShapeFromShadingNet on Real Videos [Ruano 23]. Single images seem to be sufficient in such particular context!

3d reconstruction from depth maps

Back-projection from the depth map Z:

$$M = Z(m)\mathbf{K}^{-1}m$$

[Ruano 23]



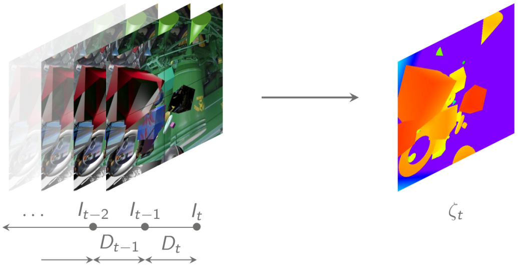
What about UAV's context?

These scenes are all taken from the same drone !






Non photorealistic synthesis for learning SfM



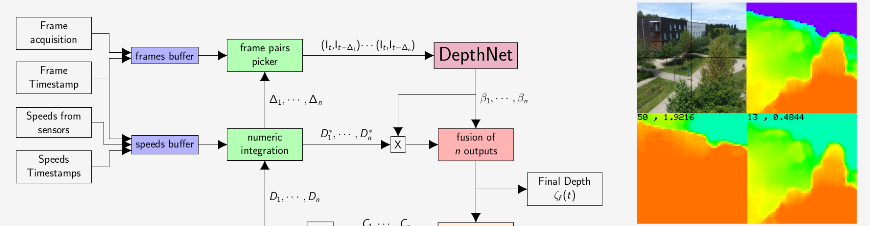
Supervised learning of depth from synthetic sequences

[Pinard 17a]

- Network is based on FlowNet_S
- Unrealistic scenes $\leftrightarrow$ Abstraction of the context
- Focus on geometry / motion, not on appearance /context
- Trained on rotationless movement, at a constant speed

Baseline adaptation using multiple image pairs

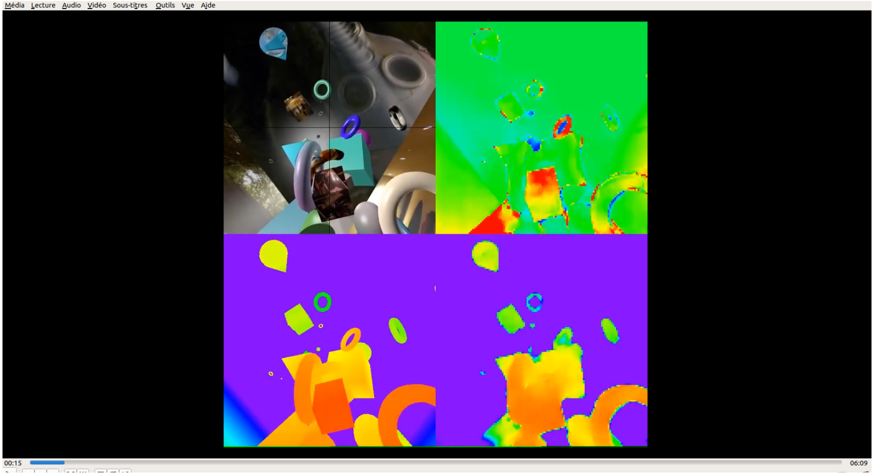
- At the inference time, the depth which is relative to the trained speed, is scaled with respect to the actual velocity.
- Adaptable precision is achieved by dynamically adapting the image pairs (baselines) to the depth distribution.



Real confitions result for a 2-planes depth inference ($n = 2$)

Adaptation of the baselines to the depth distribution **[Pinard 17b]**

Supervised DepthNet



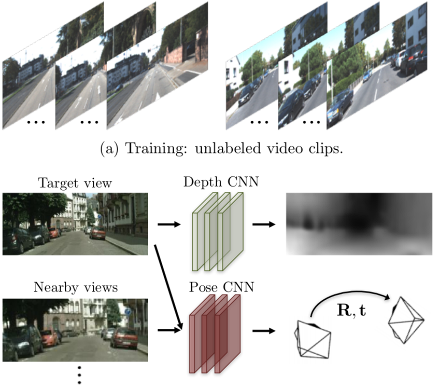
Supervised DepthNet results [Pinard 17a]: See

https://perso.ensta-paris.fr/~manzaner/Download/ECMR2017/DepthNetResults.mp4

Presentation Outline

Unsupervised depth estimation CNN

- Re-training on real/operative context is still essential.
- But data are rarely annotated.
- Self-supervised learning is then necessary.
- *Photometric loss function* can be used, that compares a pair of registered images, knowing the depth and the camera pose.
- Camera pose then needs to be known, or predicted!

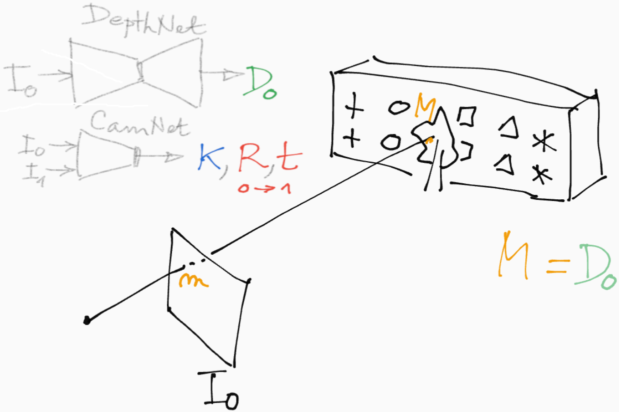


(a) Training: unlabeled video clips.

(b) Testing: single-view depth and multi-view pose estimation.

[Zhou 17]

Photometric Loss (1): Back-projection from first image

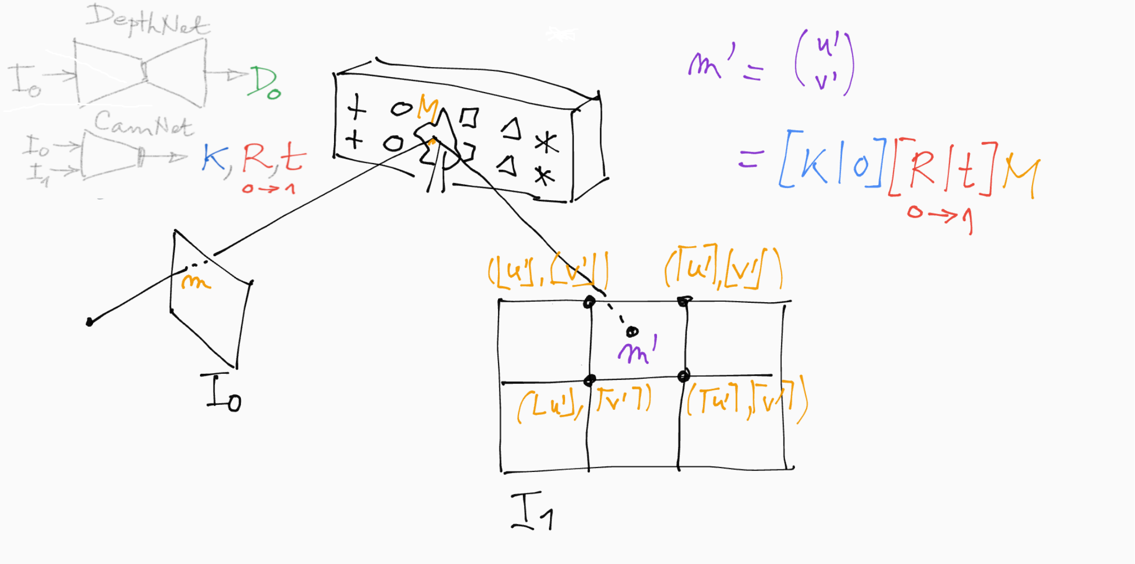


$$m = \begin{pmatrix} u \\ v \end{pmatrix}$$

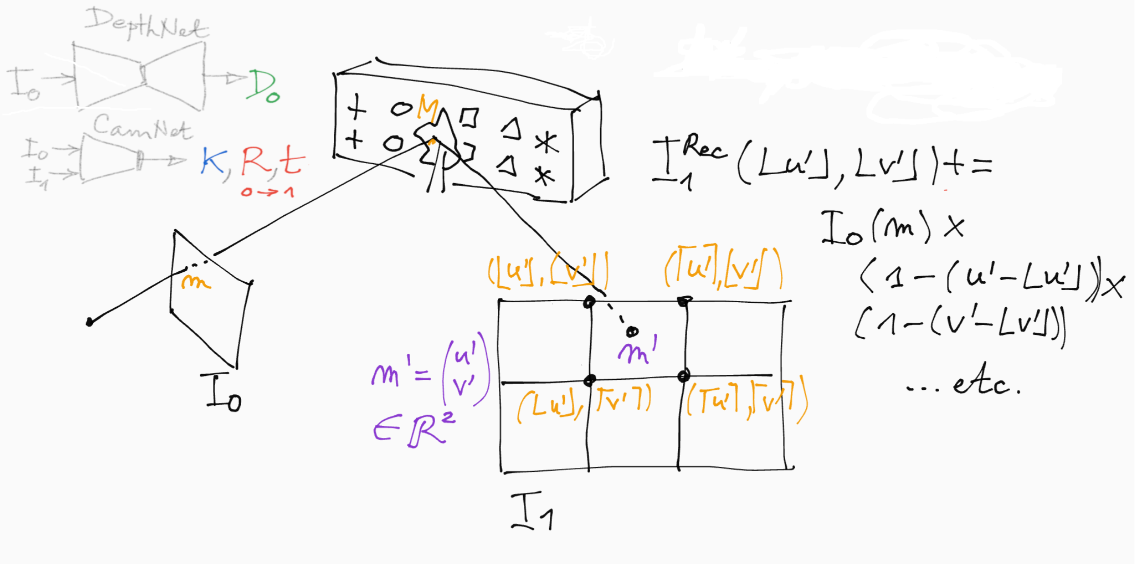
$$M = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

$$M = D_0(m) \times K^{-1} m$$

Photometric Loss (2): Re-projection onto second image



Photometric Loss (3): Interpolation within second image

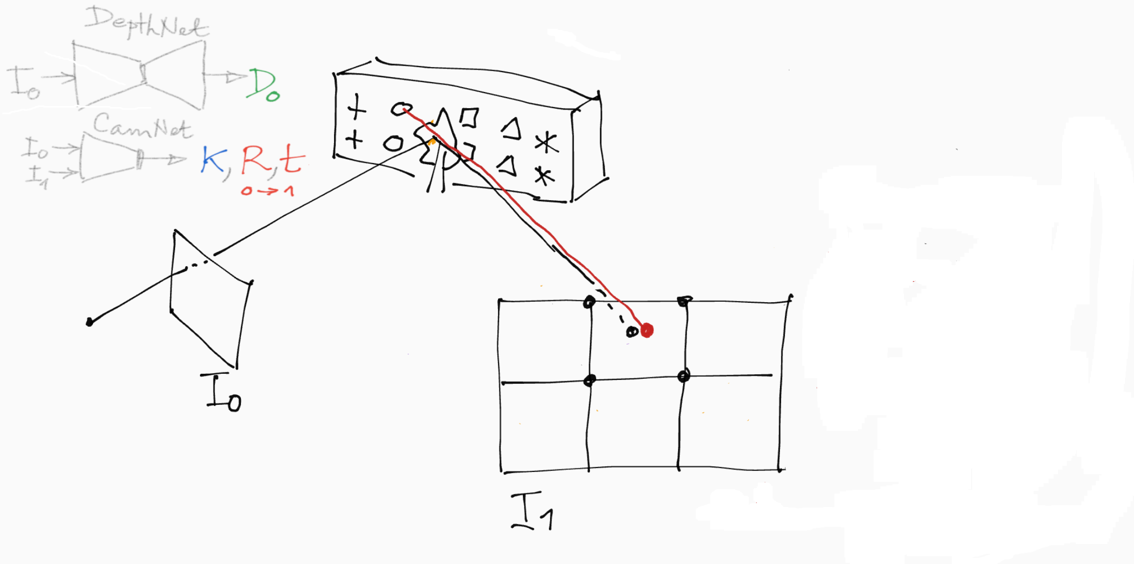


Photometric Loss: Summary and formula

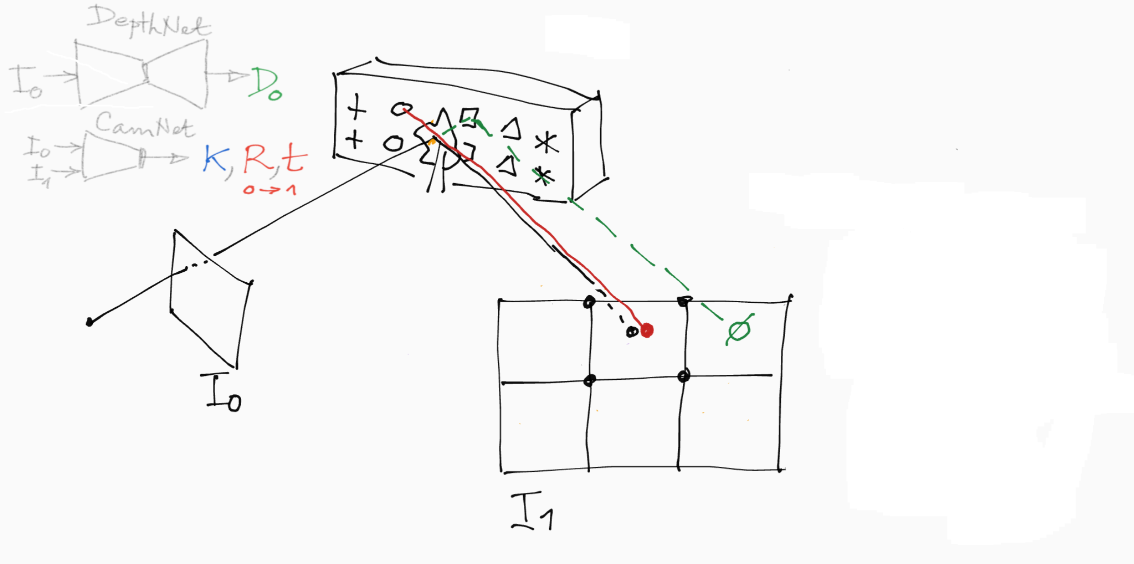
The photometric loss provides a self-supervision signal by comparing the observed image with the reconstructed image from the previous view is the depth map and odometry have been well predicted:

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

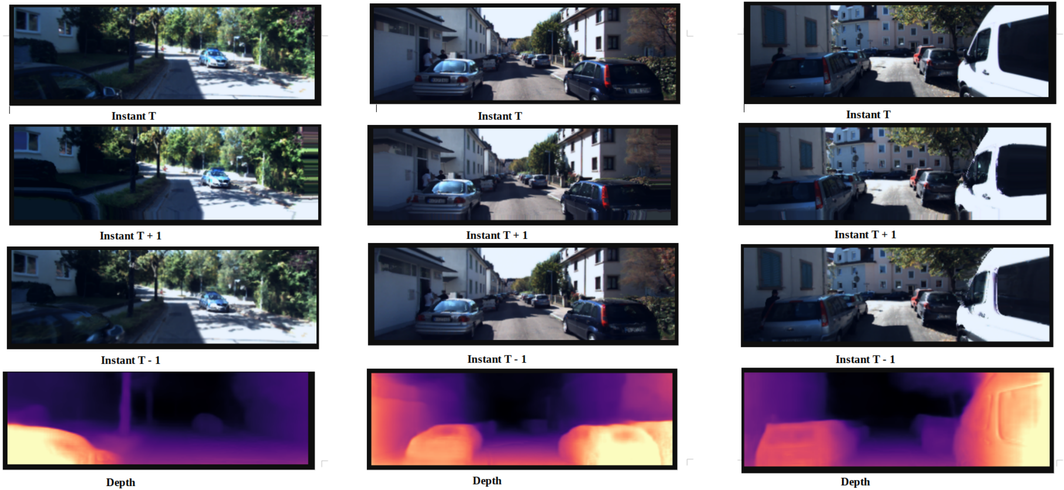
Photometric Loss: Occlusion issue



Photometric Loss: Un-occlusion issue

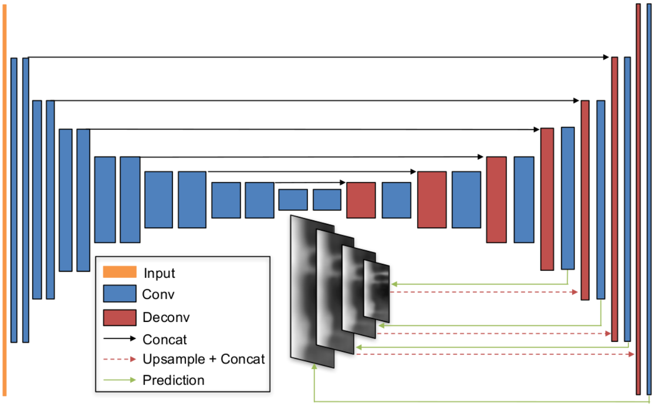


Examples of reprojected images

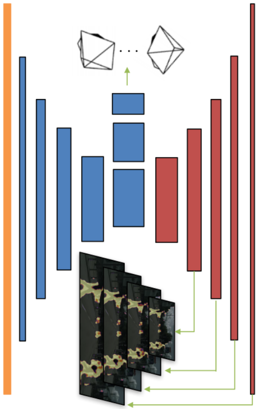


[PhD Marwane Hariat]

Unsupervised depth estimation CNN



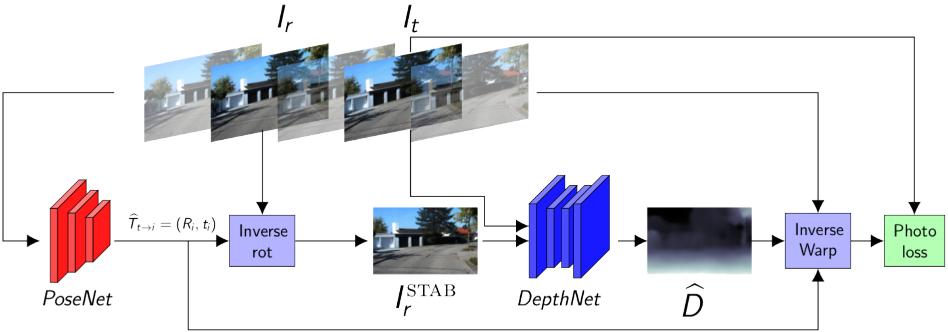
(a) Single-view depth network



(b) Pose/explainability network

[Zhou 17]

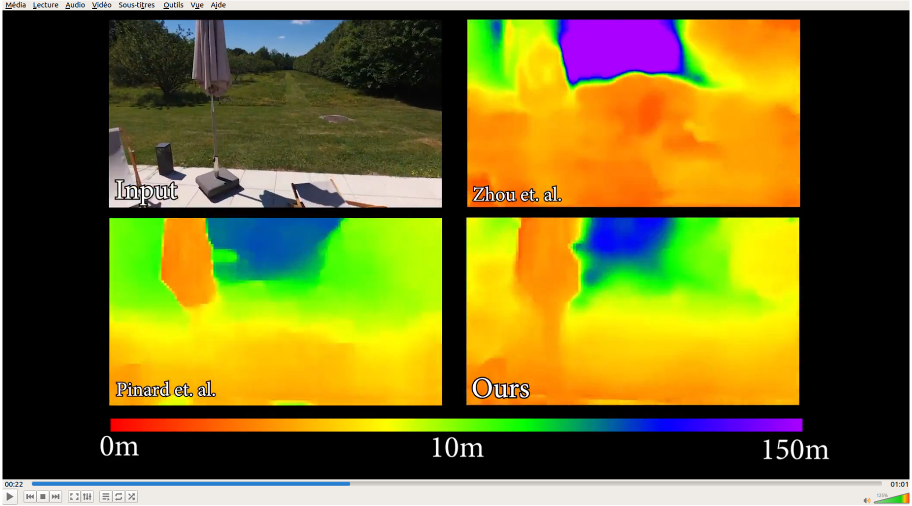
Unsupervised DepthNet



$$\forall i,\ t_i^{\mathrm{NORM}} = t_i \frac{T_0}{\epsilon + \|t_r\|}$$

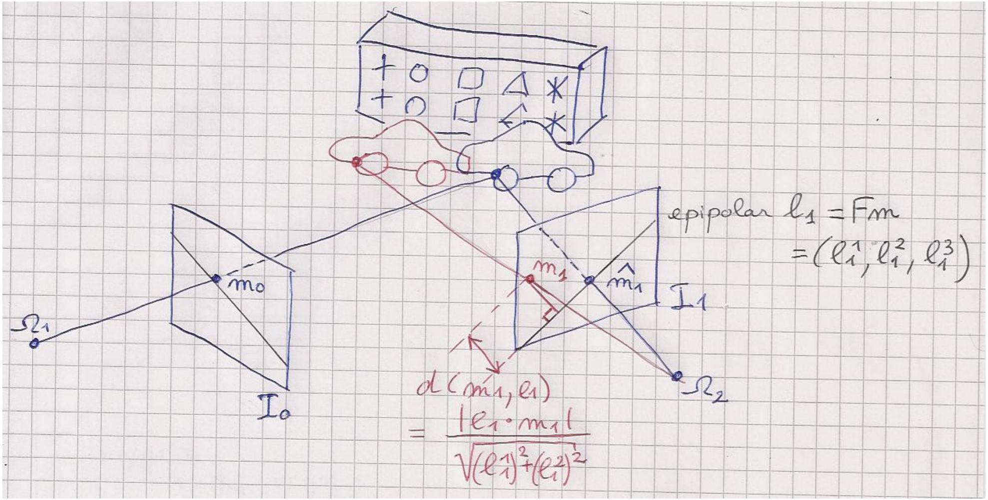
Unsupervised re-learning of Structure from Motion with adaptive baseline **[Pinard 18]**

Unsupervised DepthNet

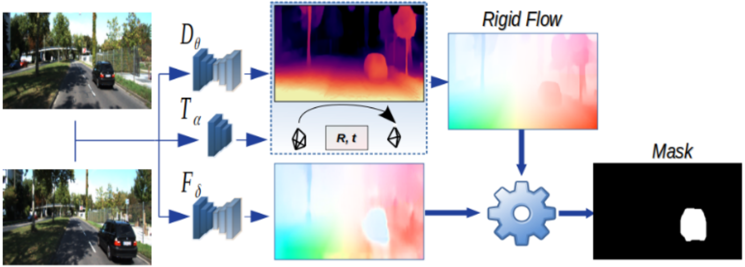


Unsupervised DepthNet real fly demo **[Pinard 18]**: See `https://www.youtube.com/watch?v=ZDgWAWTwU7U`

Photometric Loss: Moving objects issue



CoopNet: Joint training of Optical Flow, Odometry and Depth

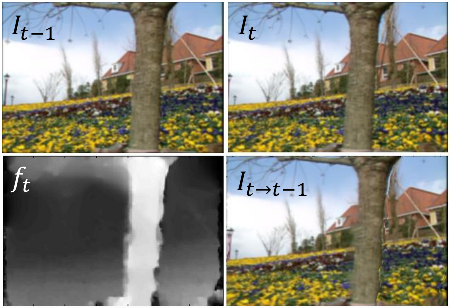


CoopNet **[Hariat 23]**

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:

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

Comparison with photometric loss for Optical Flow



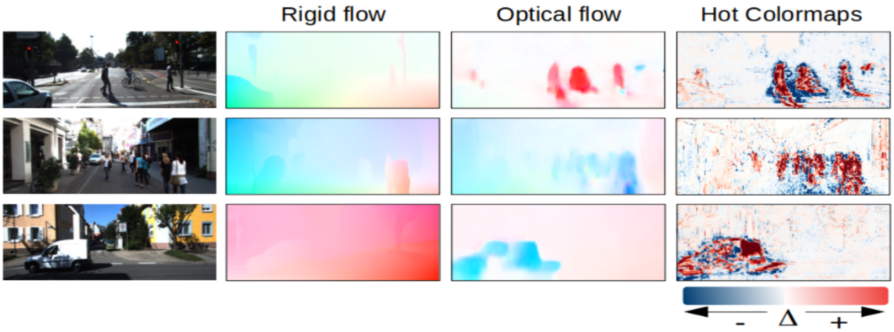
Self-supervised Optical Flow is also based on photometric loss, that measures the difference between an image and its prediction based on the optical flow:

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

with:

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

CoopNet: Joint training of Optical Flow, Odometry and Depth



CoopNet [Hariat 23]

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

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

Conclusion on Learning-based methods

- Learning optical flow and depth from videos has many advantages:
  - Globally addressing the context
  - Multi-cues depth inference
  - Natural regularization of ill-posed problem
- The main issues to adress are the hard dependence to the learned context, and the difficulties inherent to online learning. The current work perspectives are:
  - Domain adaptation: ground robotics, medical robotics,...
  - Incremental and online learning...
  - Explainability and Reliability...

Contributors for this lecture

- **Clément Pinard**: PhD student (CIFRE ANRT Parrot) 2016-2019
- **Josué Ruano Balseca**: PhD student (w. UNAL Bogotá) 2018-
- **Marwane Hariat**: PhD student 2021-

References (1)

[Eigen 14] D. Eigen and C. Puhrsch and R. Fergus
Depth map prediction from a single image using a multi-scale deep network
Advances in neural information processing systems (NIPS), pp.2366–2374, 2014

[Zhou 17] 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.

[Ruano 23] J. Ruano Balseca and M. Gómez and E. Romero and A. Manzanera
Leveraging a realistic synthetic database to learn Shape-from-Shading for estimating the colon depth in colonoscopy images
Submitted, 2023

References (2)

[Pinard 17a] C. Pinard and L. Chevalley and A. Manzanera and D. Filliat
End-to-end depth from motion with stabilized monocular videos
Int. Conf. on Unmanned Aerial Vehicles in Geomatics (UAV-g) Bonn, pp. 67-74, 2017

[Pinard 17b] C. Pinard and L. Chevalley and A. Manzanera and D. Filliat
Multi range Real-time depth inference from a monocular stabilized footage using a Fully Convolutional Neural Network
European Conference on Mobile Robotics (ECMR), Palaiseau, 2017

[Pinard 18] 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

[Hariat 23] M. Hariat and A. Manzanera and D.Filliat
Rebalancing gradient to improve self-supervised co-training of depth, odometry and optical flow predictions
IEEE Winter Conf. on Applications of Computer Vision (WACV). Waikoloa, 2023