# Learning multiple cues for depth prediction from videos

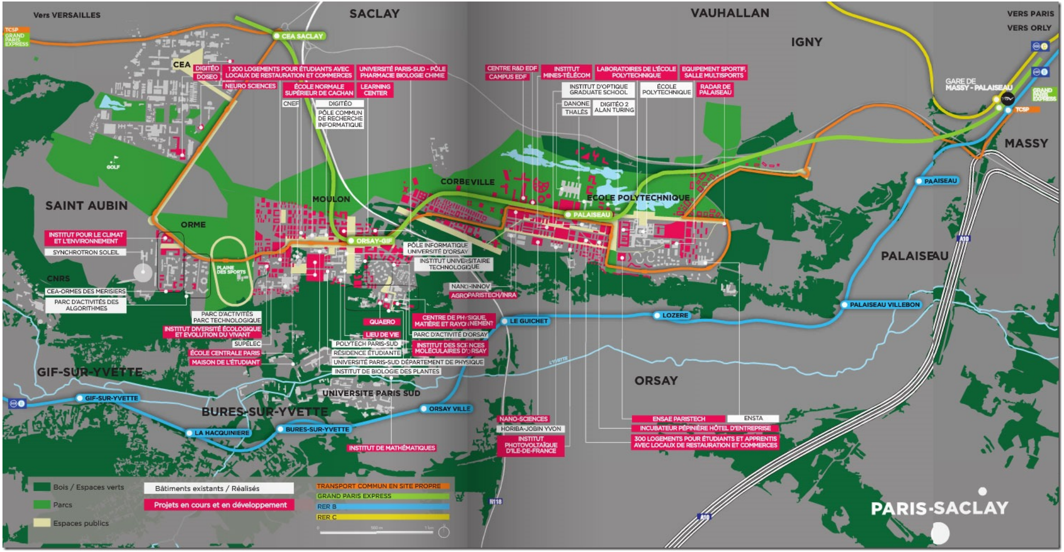
Antoine Manzanera

ENSTA Paris

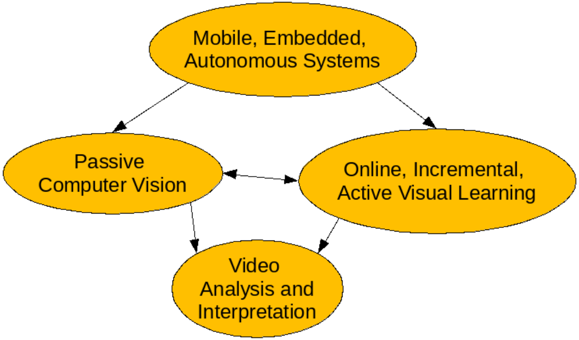
Keynote

Quito, December 2019

# ENSTA, Institut Polytechnique and Paris-Saclay



# Context and Objectives of our work

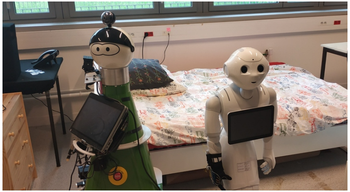


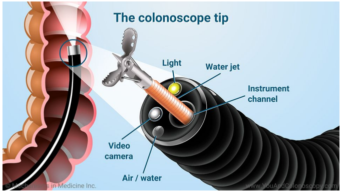

APPLICATIONS:

- Autonomous vehicles: Navigation, Road detection, Obstacle avoidance...
- Assistive robotics: gesture recognition, interaction...
- Defence and Safety: Videosurveillance,...
- Medicine: Aided diagnosis from videos...

# 3d Reconstruction from Videos

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



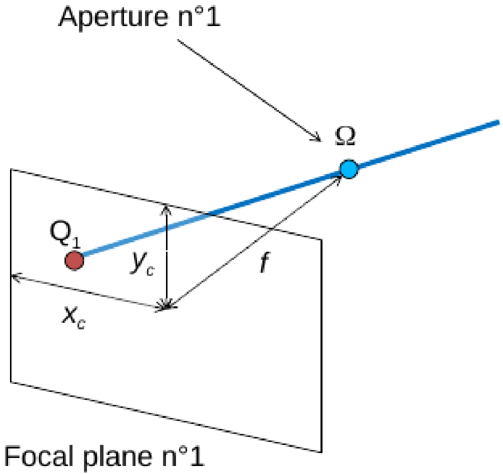



+ It is a cheap and flexible approach: One single passive camera, Adaptive baseline,...
− It strongly relies on scene structure (texture) and precise camera positioning.

# Presentation Outline

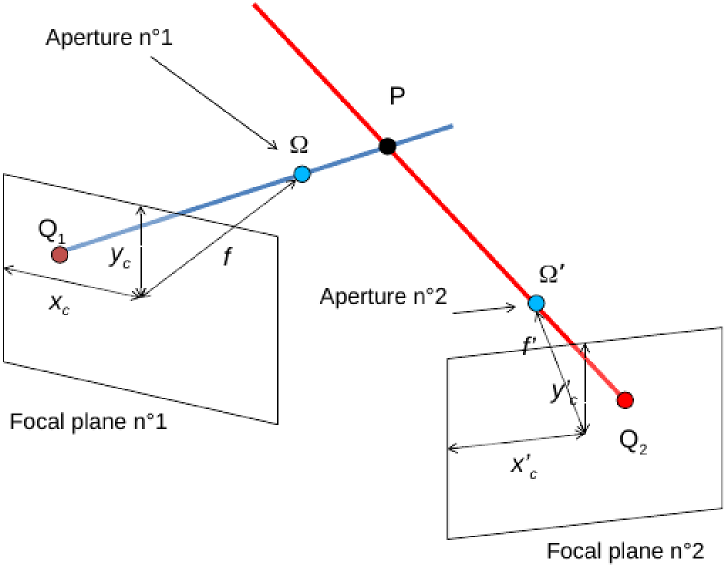
- 3d from images: Analytical Methods
- Natural Cues for Depth Inference
- Supervised learning based methods
- Unsupervised learning based Methods

# Principles of Analytical Methods




The geometry of the camera (intrinsic parameters) identifies the projection line of any point in the focal plane.

# Principles of Analytical Methods

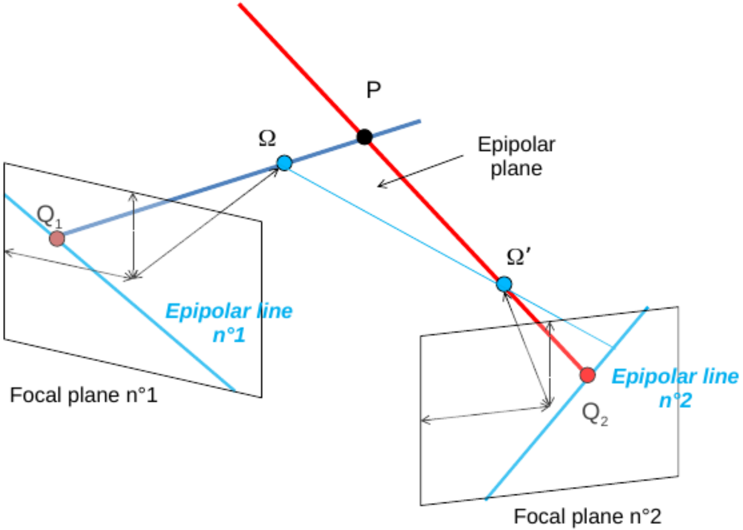



Another position of the camera (extrinsic parameters) allows to recover the 3d position of a point projected on the two focal planes:

$$\Omega P = \Omega\Omega' \frac{\sin \hat{\Omega}'}{\sin \hat{P}}$$

$$\Omega' P = \Omega\Omega' \frac{\sin \hat{\Omega}}{\sin \hat{P}}$$

# Principles of Analytical Methods




The epipolar constraints may reduce the search area for matching points. It is expressed by the fundamental matrix **F** in the projective geometry framework: $Q_1 \mathbf{F} Q_2 = 0$.

- $Q_1 \mathbf{F}$: epipolar line n.2.
- $\mathbf{F} Q_2$: epipolar line n.1.

# Epipolar Flow Estimation

Input: Image pair

## Keypoint-based sparse flow estimation

- Detector: Blockwise FAST
- Descriptor: 11x11 pixel patch
- Error filtering based on local coherence

## Fundamental matrix estimation

- With the 8-points algorithm + RANSAC

## Dense optical flow estimation

- Search domain reduced to the epipolar lines
- Propagation of the seed flow vectors coming from the sparse flow estimation

## Error filtering

- Make erroneous pixels diverge from epipolar lines
- Filter them according to the epipolar line distance and to local coherence

## Small holes filling

- Simple linear interpolation of the disparity to fill small holes caused by error filtering

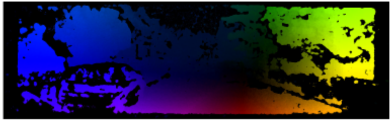
**[Garrigues 17]**

# Epipolar Flow Estimation

**[Garrigues 17]**:

- Real-Time semi-dense optical flow and relative depth estimation.
- Was ranked #1 on Kitti 2012 Optical Flow dataset (on sparse optical flow category).

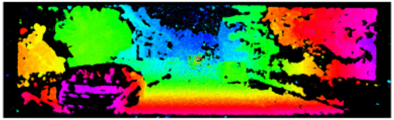
Output 1: optical flow



Output 2: disparity map

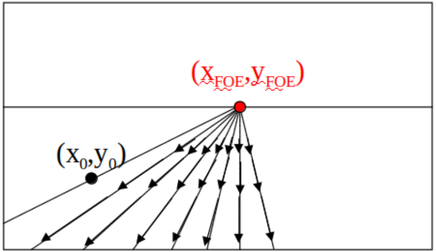
Output 3: relative depth map

(if the camera projection matrix is available)



# 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 ratio between the apparent speed of a point and its distance to 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.



# DNN for 3d reconstruction

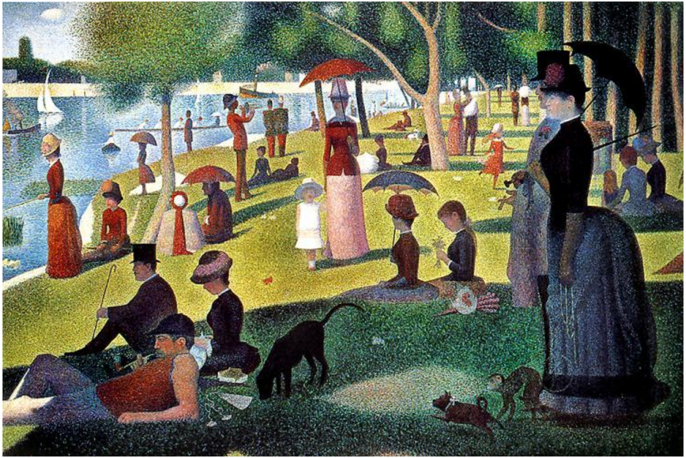
- Like for Optical Flow, Depth can benefit from Deep Networks dense prediction capabilities.
- Training can be easily done on *synthetic* or *real RGB-d* data, and loss function is also relatively straightforward.
- One determining benefit of DNN is their ability to exploit potentially *all the depth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



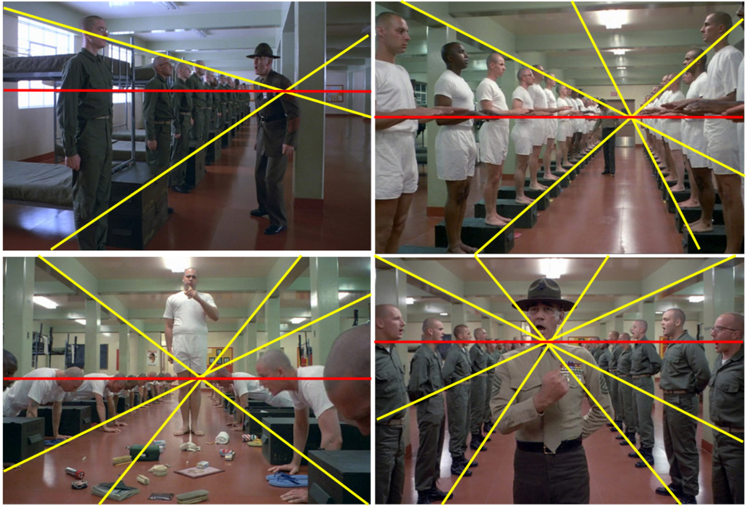
# Monocular Depth Cues? Object sizes, Perspective, and Texture Gradients!

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

# Monocular Depth Cues? Perspective, Horizon and Vanishing Points!

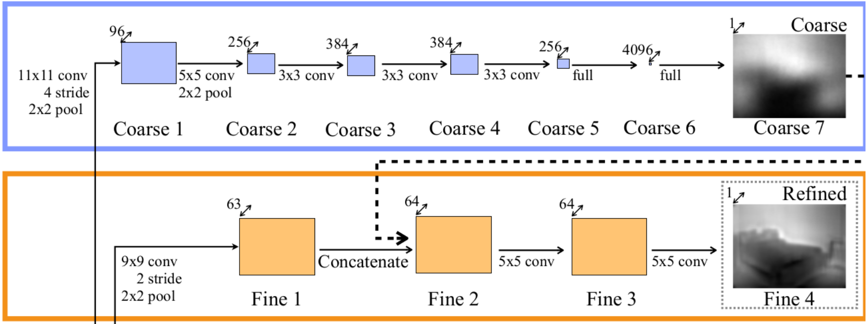
Gustave Caillebotte - Rue de Paris, temps de pluie (1877)

# Monocular Depth Cues? Horizon and Camera Pose!



*Stanley Kubrick – Full Metal Jacket (1987)*

# Depth inference from single view!

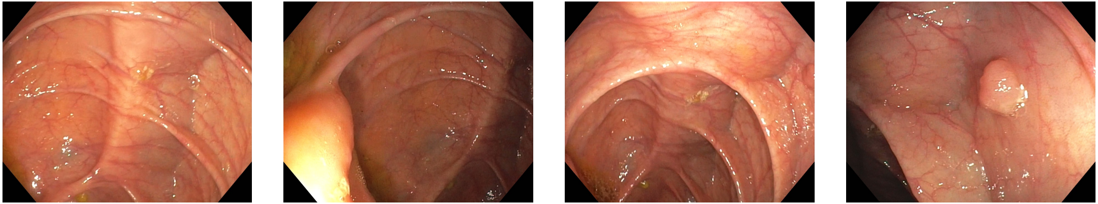




| | | Coarse | | | | | Fine |
|---|---|---|---|---|---|---|---|
| Layer | input | 1 | 2,3,4 | 5 | 6 | 7 | 1,2,3,4 |
| Size (NYUDepth) | 304x228 | 37x27 | 18x13 | 8x6 | 1x1 | 74x55 | 74x55 |
| Size (KITTI) | 576x172 | 71x20 | 35x9 | 17x4 | 1x1 | 142x27 | 142x27 |
| Ratio to input | /1 | /8 | /16 | /32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

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

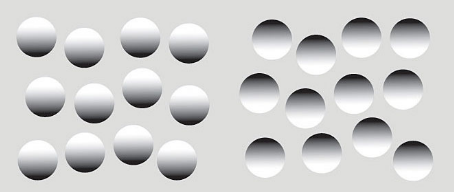
# One very particular context...

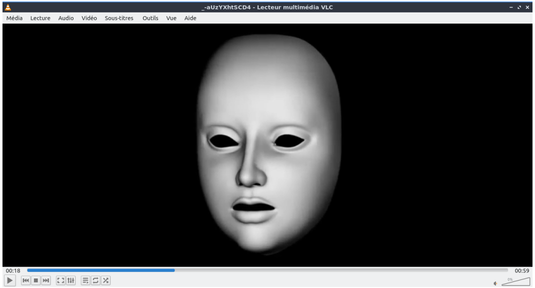


Colonoscopy images **[Ruano 19]**

# Monocular Depth Cues? Shading!

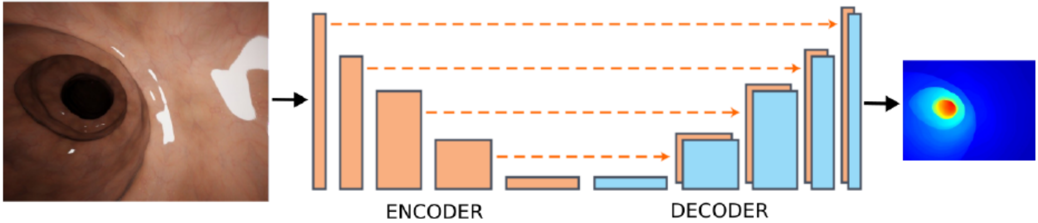
Self shadowing is a strong but ambiguous depth cue (light source position *vs* concavity). Without shape prior, the concavity is determined by a prior of top lighting (right image).





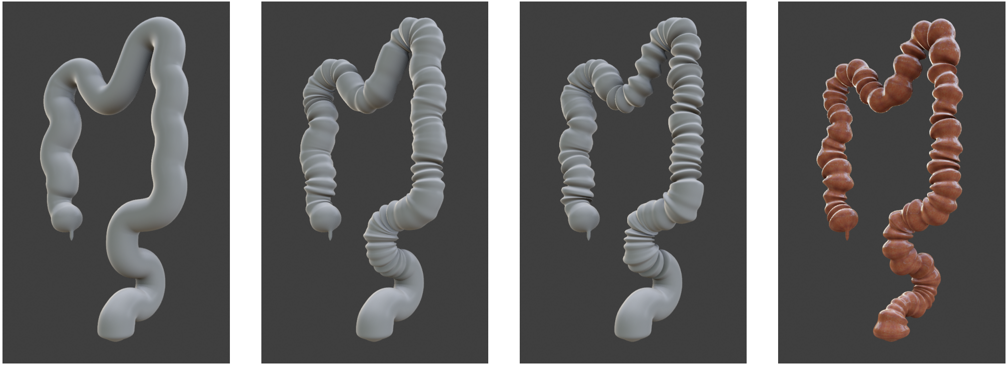

When the shape prior is strong (face then convex), the concavity prior dominates the lighting prior (top-down effect, animation on the left).

# Learning Shape from Shading for Automated Colonoscopy



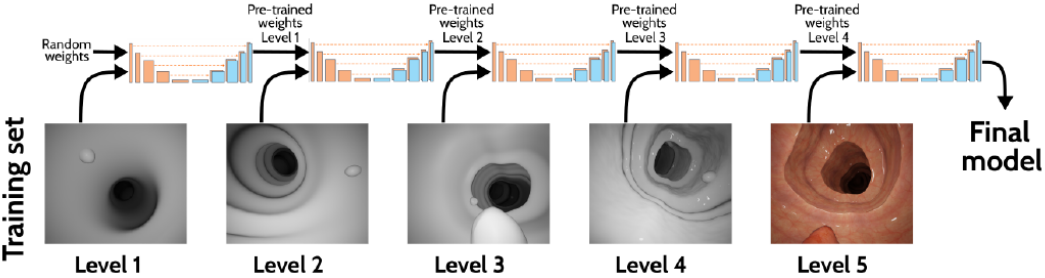

Images from synthetic videos are used to train a CNN using a loss function based on the ground truth depthmap **[Ruano 19]**

# Curriculum Learning Shape from Shading for Automated Colonoscopy



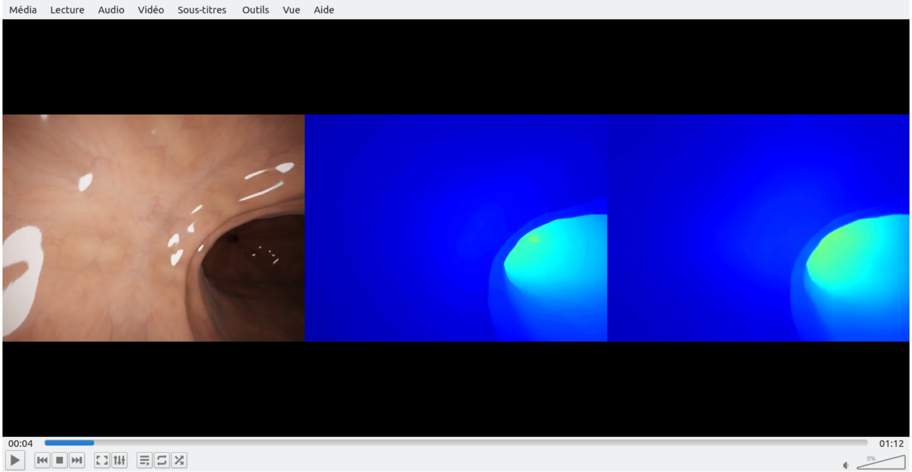
Synthetic exploration videos are created from a hierarchy of synthetic colons of increasing complexity **[Ruano 19]**

# Curriculum Learning Shape from Shading for Automated Colonoscopy



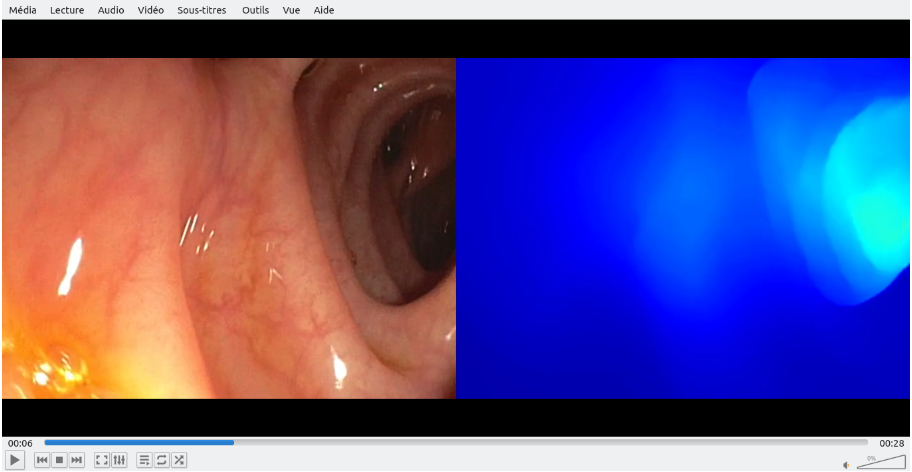

The training is performed with progressive complexity **[Ruano 19]**

# SfSNet on Synthetic Videos




ShapeFromShadingNet on Synthetic Test Videos [Ruano 19]

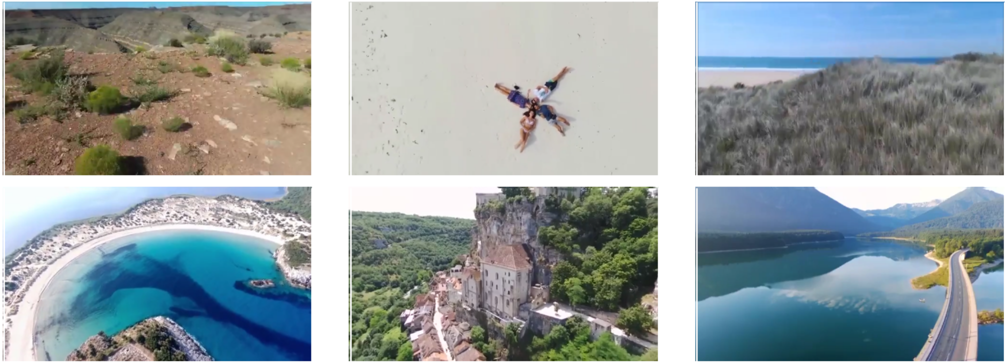
# SfSNet on Real Videos



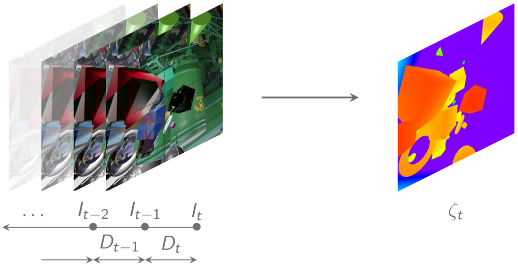

ShapeFromShadingNet on Real Videos **[Ruano 19]**. Single images seem to be sufficient in such particular context!

# What about UAV's context?

These scenes are all taken from the same drone !



# Non photorealistic synthesis for learning SfM



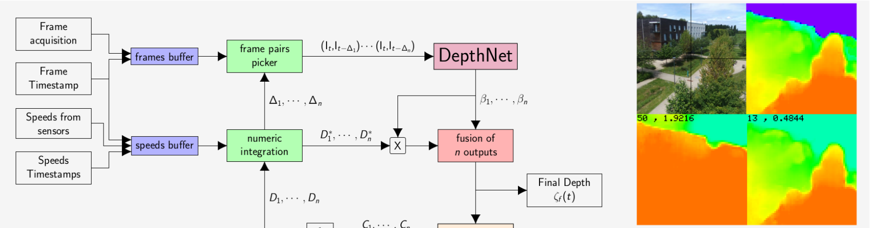

Supervised learning of depth from synthetic sequences

**[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

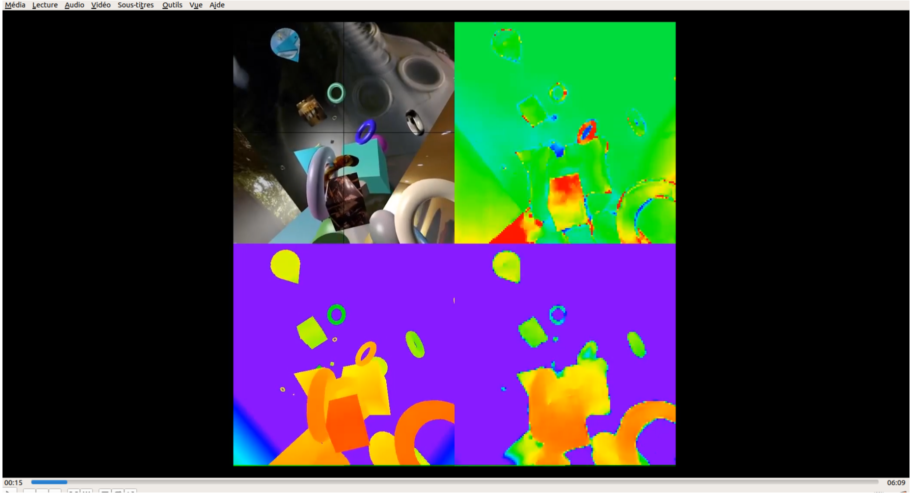
- At the inference time, the depth which is relative to the trained speed, is scaled with respect to the actual velocity.
- Adaptable precision is achieved by dynamically adapting the image pairs (baselines) to the depth distribution.




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

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

# Supervised DepthNet

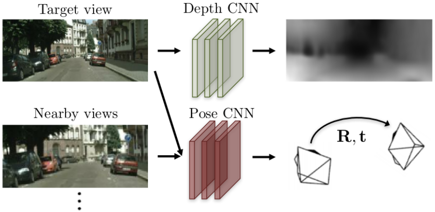



Supervised DepthNet results **[Pinard 17a]**

# 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!

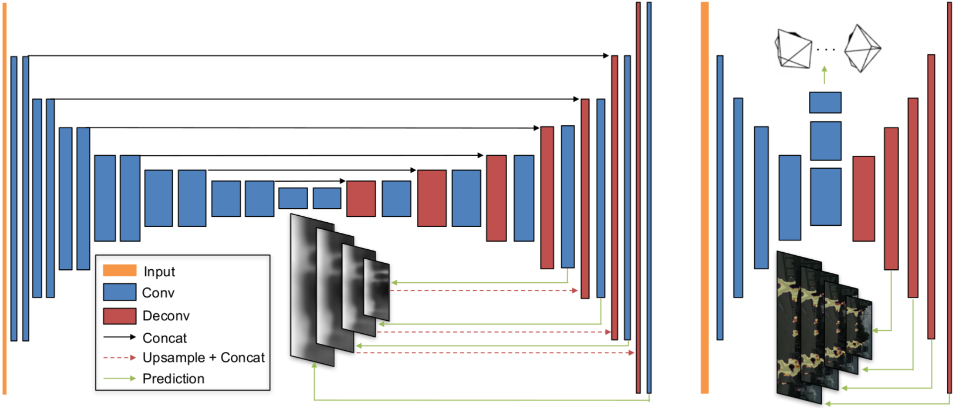
(a) Training: unlabeled video clips.




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

[Zhou 17]

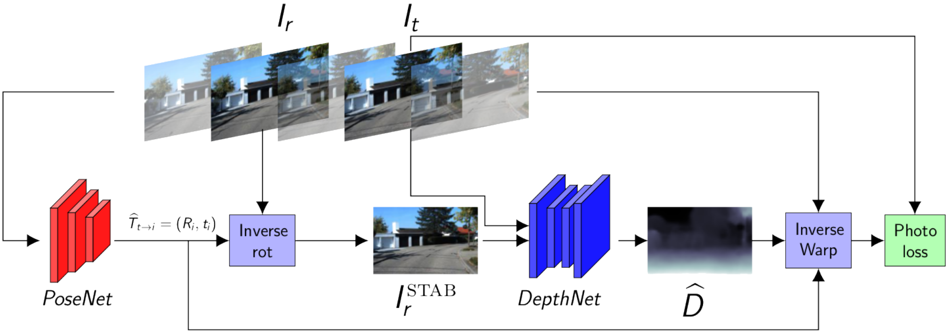
# Unsupervised depth estimation CNN




(a) Single-view depth network (b) Pose/explainability network

**[Zhou 17]**

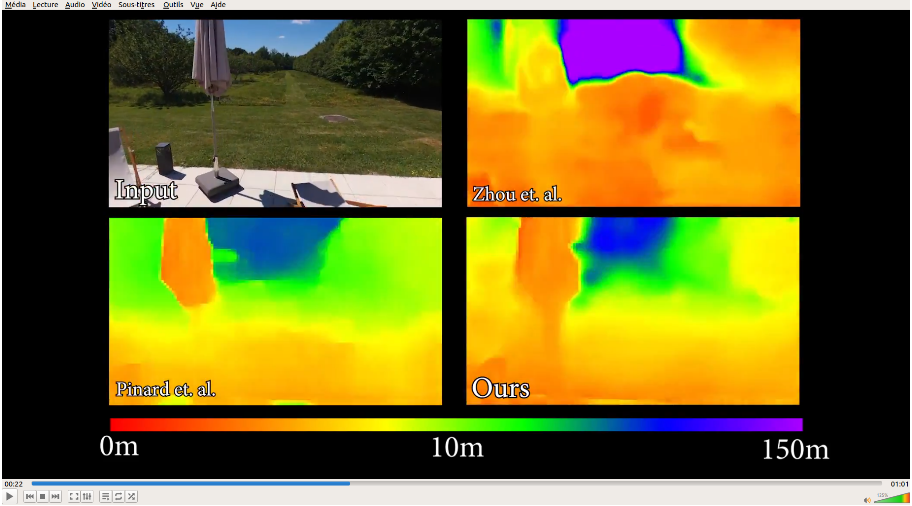
# Unsupervised DepthNet




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

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

# Unsupervised DepthNet




Unsupervised DepthNet real fly demo **[Pinard 18]**

# Conclusion and Perspectives

- 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...

# Contributors for this talk

- **Matthieu Garrigues**: PhD student 2012-2016
- **Clément Pinard**: PhD student (CIFRE ANRT Parrot) 2016-2019
- **Josué Ruano Balseca**: PhD student (w. UNAL Bogotá) 2018-

# References (1)


**[Garrigues 17]** M. Garrigues and A. Manzanera
Fast Semi Dense Epipolar Flow Estimation
IEEE Winter Conf. on Applications of Computer Vision (WACV). Sta Rosa, CA, pp.1-8, 2017

**[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

**[Ruano 19]** J. Ruano Balseca and A. Manzanera and E. Romero Castro
Curriculum-based strategy for learning shape-from-shading from colonoscopy synthetic database
Research Report, 2019

# 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

**[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.

**[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