CEA List
Laboratoire de Vision et Apprentissage pour l'analyse de scène
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Valentin Belissen
Phone +33 (0)1 69 08 33 63
E-mail valentin.belissen@cea.fr

THESIS 2022

Ref : LVA-22-T6

Learning of 3D representations independent of the image capture device

Presentation of the host laboratory

Based in Paris-Saclay campus, CEA-LIST is one of four technological research institutes of CEA TECH, the technological research direction of CEA. Dedicated to intelligent digital systems, it contributes to the competitiveness of companies via research and knowledge transfers.

The expertise and competences of the 800 research engineers and technicians at CEA-LIST help more than 200 companies in France and abroad every year on subjects categorized over 4 programs and 9 technological platforms. 21 start-ups have been created since 2003.

The Computer Vision and Machine Learning for scene understanding laboratory addresses computer vision subjects with a stronger emphasis on four axes:

- Recognition (detection or segmentation of objects and persons)
- Behavior analysis (action and gesture recognition, anomalous behavior of individuals or crowds)
- Smart annotation (large scale annotation of 2D and 3D data using semi-supervised methods)
- Perception and decision-making (Markovian decision processes, navigation)

Context

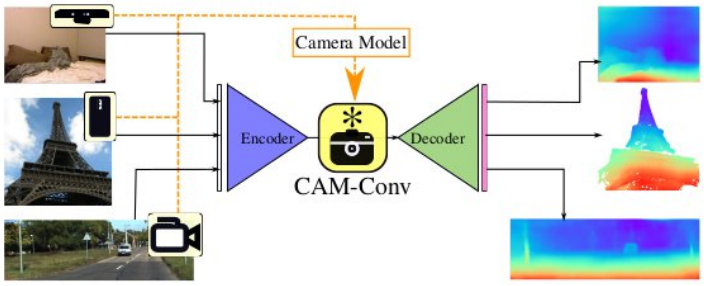
With the development of deep learning algorithms and architectures, classical image processing tasks such as object detection, semantic segmentation or depth estimation are now performed by automatically learning image-level representations. In these architectures, we find in particular the encoder-decoder type where more and more abstract representations are built by successive convolution operations on larger and larger parts of the images, or the Transformer type in which an attention mechanism allows to relate each part of an image to all the other parts.

As the learned representations are intrinsically linked to the appearance of the images, the performance of the trained models is likely to degrade in the case where inference data and training data differ signficantly. Domain adaptation is a field of research in its own right, with practical cases that are not obvious to deal with -- moving from images taken indoors to images taken outdoors, for example. However, we can also be interested in the particular case where the main domain shift comes from variations in the image capture device: change of focal length in particular. We can think of the case of autonomous vehicles, with several cameras of varying focal lengths capturing the same environment.

Work has been done in this case, with a particular application to depth estimation. While some works directly regress the camera parameters in the framework of a classical pinhole model [1, 3] or train a parallel network to estimate the complete camera model [6], other works [2, 4] have experimented with adding to the image representations maps specific to the supposedly known camera model. These maps are added at the output of the encoder, so that the decoder becomes more generic when dealing with images from cameras with variable focal lengths.

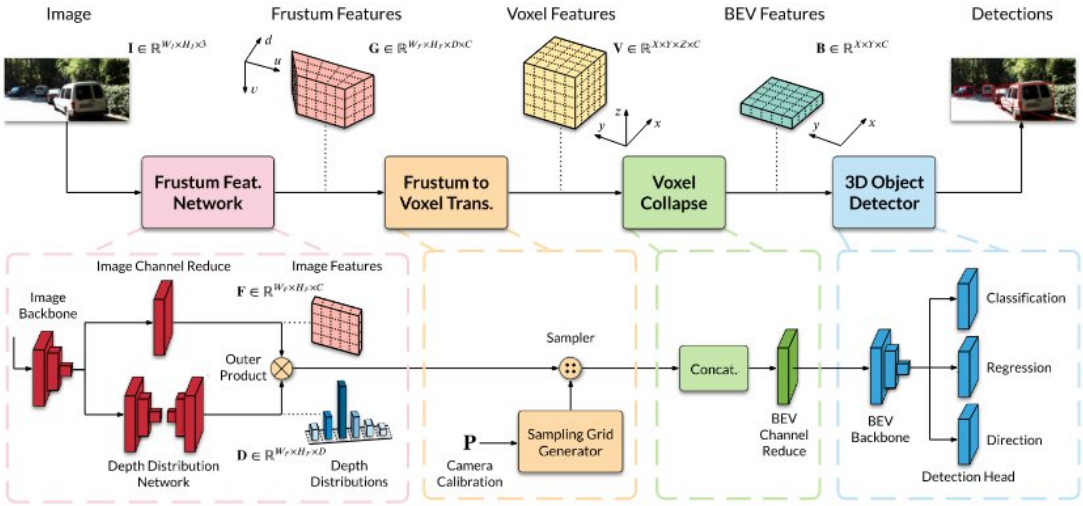
CEA List
Laboratoire de Vision et Apprentissage pour l'analyse de scène
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Valentin Belissen
Phone +33 (0)1 69 08 33 63
E-mail valentin.belissen@cea.fr



Depth estimation on images of varying focal lengths, by adding maps related to the camera model to the representations from the encoder [2].

While the representations learned in these works remain focused on the image plane, other works [5] have shown that using an intermediate representation in three dimensions -- i.e., distinct from the image plane -- could be interesting for an object detection task. These works build representations in the form of a grid of voxels, itself deduced from an estimation of the depth distribution in the image.



Detection of 3D objects, through a spatial intermediate representation, in the form of voxels [5]

Objective of the project

The thesis work will be centered around the learning of representations that are as free as possible from the type of image capturing device, thus allowing a significant genericity.

In a first step, the candidate will try to apply the methods developed in [1-4, 6] to other tasks than depth estimation, such as semantic segmentation or object detection. An evaluation of the potential of these methods will be carried out, in particular with respect to the domain adaptation related to the training of a model on a dataset and its testing on other types of data.

In a second step, the candidate will be able to propose innovative methods by bringing together the work of [1-4, 6], where representations less dependent on the camera model are learned, but still in the image plane, and the work of [5], where representations linked to the camera model are learned, but passing by a spatial representation much more relevant for a large number of perception tasks. A thorough exploration of the different architectures adapted to these methods will be conducted,

CEA List
Laboratoire de Vision et Apprentissage pour l'analyse de scène
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Valentin Belissen
Phone +33 (0)1 69 08 33 63
E-mail valentin.belissen@cea.fr

together with the implementation of adapted data sets and evaluation processes.

Finally, the candidate will be able to study the application of such spatial representations to multi-camera datasets, where spatial fusion directly at the representation level could be interesting.

References:

[1] Sai Shyam Chanduri, Zeeshan Khan Suri, Igor Vozniak, and Christian Müller. CamLessMonoDepth : Monocular Depth Estimation with Unknown Camera Parameters. October 2021.

[2] Jose M. Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, and Javier Civera. CAM-Convs : Camera-Aware Multi-Scale Convolutions for Single-View Depth. April 2019.

[3] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from Videos in the Wild : Unsupervised Monocular Depth Learning from Unknown Cameras. The IEEE International Conference on Computer Vision (ICCV), 2019, pp. 8977-8986 , April 2019.

[4] Varun Ravi Kumar, Senthil Yogamani, Hazem Rashed, Ganesh Sitsu, Christian Witt, Isabelle Leang, Stefan Milz, and Patrick Mäder. OmniDet : Surround View Cameras based Multi-task Visual Perception Network for Autonomous Driving. IEEE Robotics and Automation Letters , 6(2) :2830 2837, 2021.

[5] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical Depth Distribution Network for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pages 8555 8564, 2021.

[6] Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-motion. August 2020

Level required:	Engineer, M.Sc.
Duration:	3 years
Compensation:	1800 € to 2000 €.

Required skills:

- Computer vision
- Machine learning (deep learning)
- Pattern recognition
- C/C++, Python
- - Mastery of a deep learning framework (in particular Tensorflow or PyTorch) is a plus.