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
Tél +33 (0)1 69 08 33 63
E-mail valentin.belissen@cea.fr

THESE 2022 Réf : LVA-22-T6

Apprentissage de représentations 3D indépendantes du dispositif de capture d'image

Présentation du laboratoire d'accueil

Basé à Paris-Saclay, le CEA List est l'un des quatre instituts de recherche technologique de CEA Tech, direction de la recherche technologique du CEA. Dédié aux systèmes numériques intelligents, il contribue au développement de la compétitivité des entreprises par le développement et le transfert de technologies.

L'expertise et les compétences développées par les 800 ingénieurs-chercheurs et techniciens du CEA List permettent à l'Institut d'accompagner chaque année plus de 200 entreprises françaises et étrangères sur des projets de recherche appliquée s'appuyant sur 4 programmes et 9 plateformes technologiques. 21 start-ups ont été créées depuis 2003.

Labellisé Institut Carnot depuis 2006, le CEA List est aujourd'hui l'institut Carnot Technologies Numériques

Le Laboratoire de Vision et Apprentissage pour l'analyse de scène (LVA) mène ses recherches dans le domaine de la Vision par Ordinateur (Computer Vision) selon quatre axes principaux :

- La reconnaissance visuelle (détection et/ou segmentation d'objets, de personnes, de patterns; détection d'anomalies; caractérisation)
- L'analyse du comportement (reconnaissance de gestes, d'actions, d'activités, de comportements anormaux ou spécifiques pour des individus, un groupe, une foule)
- L'annotation intelligente (annotation à grande échelle de données visuelles 2D/3D de manière semi-automatique)
- La perception et la décision (processus de décision markovien, navigation)

Contexte

Avec le développement d'algorithmes et d'architectures d'apprentissage profond, les tâches classiques de traitement d'image comme la détection d'objet, la segmentation sémantique ou l'estimation de profondeur sont désormais réalisées en apprenant de manière automatique des représentations au niveau image. Dans ces architectures, on trouve notamment le type encodeur-décodeur où des représentations de plus en plus abstraites se construisent par des opérations de convolution successives sur des parties de plus en plus grandes des images, ou encore le type Transformer dans lequel un mécanisme d'attention permet de mettre en rapport chaque partie d'une image avec toutes les autres parties.

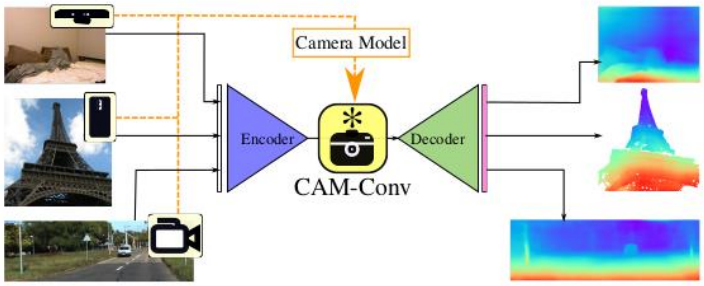
Les représentations apprises étant intrinsèquement liées à l'apparence des images, la performance des modèles entraînés est susceptible de se dégrader significativement dès lors que celle-ci n'est pas identique entre les données utilisées en entraînement et en inférence. L'adaptation de domaine est un champ de recherche à part entière, avec des cas pratiques non évidents à traiter -- passer d'images prises en intérieur à des images prises en extérieur par exemple. Cependant, on peut aussi s'intéresser au cas particulier où le principal changement de domaine vient de variations dans le dispositif de capture d'image : changement de type de focale en particulier. On peut penser au cas de véhicules autonomes, avec plusieurs caméras de focales variables captant le même environnement.

Des travaux ont été réalisés dans ce cas de figure, avec pour application particulière l'estimation de profondeur. Tandis que certains régressent directement les paramètres de caméra dans le cadre d'un modèle classique pinhole [1, 3] ou bien entrainent un réseau parallèle à estimer le modèle de caméra complet [6], d'autres travaux [2, 4] ont expérimenté d'adjoindre aux représentations propres aux images

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

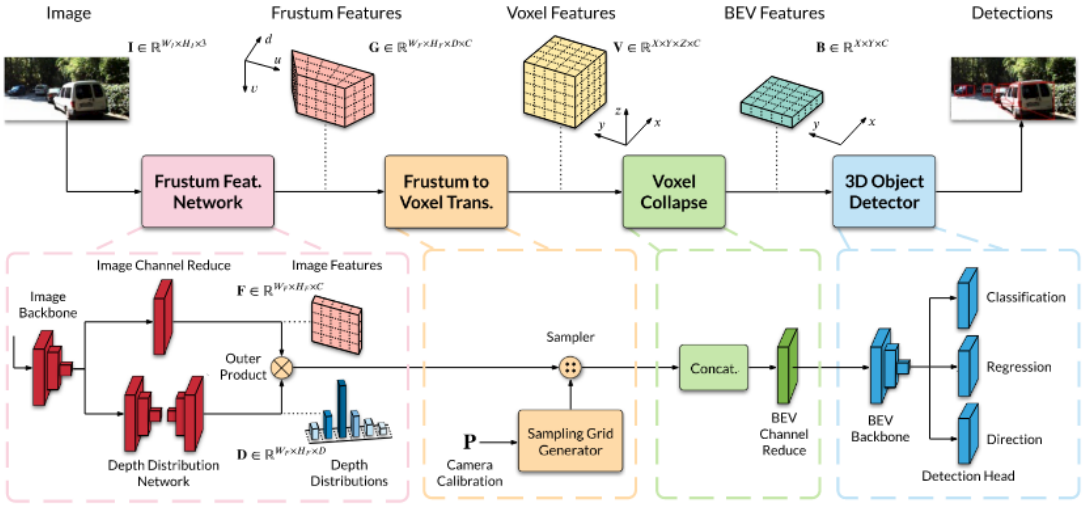
Contact Valentin Belissen
Tél +33 (0)1 69 08 33 63
E-mail valentin.belissen@cea.fr

des cartes spécifiques au modèle de caméra supposé connu. Ces cartes sont ajoutées en sortie de l'encodeur, de sorte que le décodeur devient plus générique face à des images issues de caméras à focale variable.



Estimation de profondeur sur des images de focales variables, en adjoignant aux représentations issues de l'encodeur des cartes liées au modèle de caméra [2]

Tandis que les représentations apprises dans ces travaux restent malgré tout focalisées sur le plan image, d'autres travaux [5] ont montré que passer par une représentation intermédiaire en trois dimensions -- c'est-à-dire distincte du seul plan image -- pouvait être intéressant pour une tâche de détection d'objets. Ces travaux construisent en effet des représentations sous la forme d'une grille de voxels, elle-même déduite d'une estimation de la distribution de profondeur dans l'image.



Détection d'objets 3D, passant par une représentation intermédiaire spatiale, sous forme de voxels [5]

Objectif de la thèse

Le travail de thèse sera centré autour de l'apprentissage de représentations qui s'affranchissent autant que possible du type de dispositif de capture d'image, permettant ainsi une généricité importante.

Dans un premier temps, le ou la candidat.e pourra tenter d'appliquer les méthodes développées dans [1-4, 6] à d'autres tâches que l'estimation de profondeur, comme la segmentation sémantique ou la détection d'objets. Une évaluation du potentiel de ces méthodes sera menée, notamment vis-à-vis de l'adaptation de domaine liée à l'entraînement d'un modèle sur un ensemble de données et son test sur d'autres types de données.

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
Tél +33 (0)1 69 08 33 63
E-mail valentin.belissen@cea.fr

Dans un second temps, le ou la candidat.e pourra proposer des méthodes innovantes en faisant se rejoindre les travaux de [1-4, 6], où des représentations moins dépendantes du modèle de caméra sont apprises, mais toujours dans le plan image, aux travaux de [5], où des représentations liées au modèle de caméra sont apprises, mais en passant par une représentation spatiale beaucoup plus pertinente pour un grand nombre de tâches de perception. Une exploration approfondie des différentes architectures adaptées à ces méthodes sera menée, conjointement à la mise en place d'ensemble de données et de processus d'évaluation adaptés.

Enfin, le ou la candidat.e pourra étudier l'application de telles représentations spatiales à des jeux de données multi-caméras, où la fusion spatiale directement au niveau des représentations pourrait être intéressante.

Références:

[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

Niveau demandé :	Ingénieur, Master 2
Durée :	3 ans
Rémunération :	entre 1800 € et 2000 €.

Compétences requises :

- Vision par ordinateur
- Apprentissage automatique (deep learning)
- Reconnaissance de formes
- C/C++, Python
- La maîtrise d'un framework d'apprentissage profond (en particulier Tensorflow ou PyTorch) est un plus.