Laboratoire de Vision et d'Apprentissage pour l'analyse de scène

CEA Saclay 91191 Gif-sur-Yvette France
https://kalisteo.cea.fr

Date de parution : 04/07/2021
Référence : LVA-2021-Th5

Contact :
Camille Dupont camille.dupont@cea.fr
Hejer Ammar hejer.ammar@cea.fr
Quoc Cuong Pham quoc-cuong.pham@cea.fr

Offre de thèse de doctorat

Détection non-supervisée d'objets mobiles dans des données vidéos monoculaires

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.

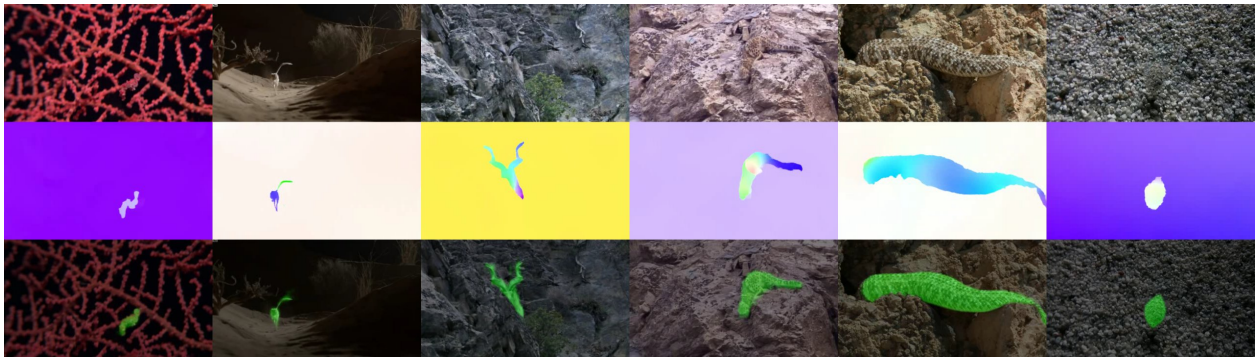
Le Laboratoire de Vision et Apprentissage pour l'analyse de scène (LVA) mène ses recherches dans le domaine de la Vision par Ordinateur (Computer Vision) selon quatre axes principaux :

- La reconnaissance visuelle
- L'analyse du comportement
- L'annotation intelligente et la caractérisation de données
- Les modèles de perception et de décision.

Contexte

Les techniques traditionnelles de détection d'objet par apprentissage supervisé nécessitent une quantité importante de données d'apprentissage. Ce constat motive depuis plusieurs années la communauté scientifique à investiguer des méthodes d'apprentissage avec la plus faible supervision possible. Dans ce contexte, on remarque une popularité croissante de l'apprentissage auto-supervisé qui génère de manière automatique les annotations d'apprentissage en exploitant les relations entre les signaux d'entrée. Parmi les tâches phares pouvant tirer parti de l'apprentissage, on note surtout des tâches bas-niveaux peu sémantisées telles que l'estimation de flot optique, l'estimation de la profondeur à partir de vidéo monoculaires ou des tâches de pré-apprentissage [1].

Résultats d'estimation de flot optique, profondeur et segmentation à partir de deux images avec la méthode EffiScene [3]



Résultats de prédiction de segmentation à partir d'un flot optique avec la méthode [2]

Objectif de la thèse

L'objectif de cette thèse est d'aller encore plus loin dans l'exploitation de cette famille d'apprentissage dans un contexte de détection d'objets. Plus particulièrement, le travail attendu dans la thèse est de concevoir un modèle de détection d'objet entraîné à partir de données vidéos monoculaires qui pourra ensuite être utilisé dans l'image ou la vidéo pour détecter l'ensemble des classes d'objets présentes dans les données d'entraînement.

De récents travaux montrent des résultats intéressants de prédiction de segmentation d'objet à partir de flot optique [2]. On observe cependant que cette dernière approche se focalise seulement sur les objets en mouvements et rate complètement les objets statiques bien que pouvant appartenir à la même classe d'objet. Un autre challenge se pose quand la caméra est en mouvement (par exemple sur un véhicule). D'autres méthodes incluent l'apprentissage d'autre sous-taches comme la carte de profondeur ou bien la localisation 3D des objets qui peuvent augmenter la compréhension du modèle de la sémantique et la géométrie des objets et ainsi corriger les limitations de la focalisation sur les objets en mouvements [3]. Enfin, les « transformers » (ViT) ont récemment montré de bonnes capacités à extraire des cartes d'attention pertinentes pour localiser les objets saillants de l'image [8].

Le doctorant pourra dans un premier temps étudier la capacité de généralisation et la robustesse des méthodes de segmentation auto-supervisées aux objets non-mobiles. Dans un second temps, il pourra proposer des améliorations en investiguant par exemple l'utilisation de méthodes de labels manquants « Missing Labels » [4] et de classification non-supervisée d'objets présents dans les données d'apprentissage [5]. Pour prouver l'efficacité de la méthode développée, il devra mener des expérimentations sur des benchmarks de détection ou segmentation d'objets comme KITTI et se comparer avec des méthodes de détection supervisées et non-supervisées [6,7].

Références

[1] Lorre, G., Rabarisoa, J., Orcesi, A., Ainouz, S. and Canu, S., (2020). Temporal contrastive pretraining for video action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 662-670).
[2] Yang, C., Lamdouar, H., Lu, E., Zisserman, A. and Xie, W., (2021). Self-supervised Video Object Segmentation by Motion Grouping. arXiv preprint arXiv:2104.07658.
[3] Jiao, Y., Tran, T.D. and Shi, G., (2021). EffiScene: Efficient Per-Pixel Rigidity Inference for Unsupervised Joint Learning of Optical Flow, Depth, Camera Pose and Motion Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5538-5547).
[4] Xu, M., Bai, Y., Ghanem, B., Liu, B., Gao, Y., Guo, N., Ye, X., Wan, F., You, H. and Fan, D., (2019), January. Missing Labels in Object Detection. In *CVPR Workshops* (Vol. 3).
[5] Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M. and Van Gool, L., (2020), August. Scan: Learning to classify images without labels. In *European Conference on Computer Vision* (pp. 268-285). Springer, Cham.
[6] Hayat, N., Hayat, M., Rahman, S., Khan, S., Zamir, S.W. and Khan, F.S., (2020). Synthesizing the unseen for zero-shot object detection. In *Proceedings of the Asian Conference on Computer Vision*.

[7] Rahman, S., Khan, S. and Barnes, N., (2020), April. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11932-11939).
[8] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. and Joulin, A., (2021). Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.
[9] George, A. and Marcel, S., (2020). On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing. *arXiv preprint arXiv:2011.08019*.

Profil du(de la) candidat(e) :

- Master recherche ou ingénieur avec une expérience en vision et en machine learning
- Compétences : vision par ordinateur, apprentissage automatique, notamment *deep learning*
- Programmation : Python, C++, frameworks de Deep Learning (Tensorflow, PyTorch)

Si vous vous reconnaissez dans ces compétences et ce profil et si développer vos compétences sur la thématique de l'IA, au sein d'un institut ambitieux et reconnu, au cœur de l'environnement dynamique du plateau de Saclay s'inscrit dans votre projet professionnel, merci de transmettre CV + lettre de motivation à :

camille.dupont@cea.fr / hejer.ammar@cea.fr / quoc-cuong.pham@cea.fr