# PhD position — Occlusion Handling in Multi-Object Tracking

**Pierre Perrault, Vincent Despiegel, Stephane Gentric, Liming Chen**
**pierre.perrault@idemia.com — liming.chen@ec-lyon.fr**

## About Idemia and Ecole Centrale de Lyon

Formed through the merger of Oberthur Technologies (OT) and Safran Identity & Security (Morpho) in 2017, Idemia is a world leader in identity technologies, specialized in biometrics, identification and authentication, data and video analysis. Ecole Centrale de Lyon is a top ranked French elite engineer grande école which has developed an excellence in research of engineering overs its history of more than 160 years since its foundation.

## Context

The Tracking & Scene Understanding research team at Idemia, in collaboration with Ec Lyon, has opened a 3-year PhD position (Cifre). This position involves dividing the research responsibilities equally between Idemia's Courbevoie location in La Defense and EC Lyon (the chosen PhD student will have the chance to work collaboratively at both institutes). The context for the PhD position at Idemia is centered on road safety and public security applications, with a specific emphasis on tracking pedestrians and cars.

## PhD topic

Occlusions (i.e., when objects are partially or completely obscured by other objects) remain a significant barrier to high performance in scene understanding tasks. This doctoral research project aims to improve multi-object (e.g., pedestrians and vehicles) tracking (MOT) models to make them robust to occlusions. Occlusions are challenging because:

- $(i)$ Public dataset annotations typically prioritize visible data, which is easier for humans to annotate. This bias in annotation leads to a scarcity of labeled data that handle occlusions effectively.
- $(ii)$ Even when non visible parts of objects are fully annotated, models struggle to directly link hidden elements with visual patterns, and have to rely heavily on contextual cues from the spatio-temporal surrounding of the element, which often requires significantly more training data. The same phenomenon arises in 3D detection/tracking, as it typically necessitates looking beyond pixel-based visual patterns.

To address the above mentioned difficulties, the use of very large datasets with non-human supervision (or limiting it to a few examples) in training is a promising approach. One research direction could be to exploit the implicit signals present in the spatio-temporal context of many unlabeled videos, using self-supervised learning. Another direction could be to use synthetic data generated by simulation engines, which can benefit from having perfect labels (thereby benefiting the aforementioned 3D tasks as well). Both offer the advantage of being relatively unlimited in dataset size, the first focusing on the quality/realism of the data and the second focusing on the quality of the labels. By combining the two, the hope is to be able to leverage the large size and high quality of both the data and labels, thereby enhancing the overall training process and ultimately improving the performance of the scene understanding algorithms in difficult and dense scenarios.

## Profile

Candidates must have completed or be in the final stages of defending their MSc degree. They should possess a strong foundation in computer vision, which encompasses 3D processing, as well as machine learning, specifically deep learning, along with proficient coding skills in PyTorch.

## Applying

To apply, candidates are required to email Pierre Perrault at pierre.perrault@idemia.com and Liming Chen at liming.chen@ec-lyon.fr. The email should include the following:

- A cover letter demonstrating their interest and suitability for the thesis topic.
- Their CV.
- A transcript of their MSc grades.
- Some references or recommendation letters.

Applications will be reviewed on a rolling basis. The anticipated starting date is fall 2023.

## Literature

In recent works, self-supervised learning has shown promising results for MOT. (*1*) proposed a self-supervised method for MOT using unlabeled video data. They introduced a novel self-supervisory signal called cross-input consistency, which trains an RNN model to produce consistent tracks across two distinct inputs. In (*2*), instead of directly supervising invisible object locations, a self-supervised objective based on temporal coherence of memory is introduced. This objective optimizes a Markov walk along a space-time graph of memories, allowing the model to store occluded objects and predict their motion for better localization. In (*3*), the authors address the propagation and association tasks in MOT by introducing a pixel-guided approach. The proposed method aims to unify different aspects of MOT, such as appearance modeling, motion modeling, and object associations, into a joint-detection and tracking framework. While self-supervised trackers have the potential to leverage large amounts of raw data, they often underperform compared to supervised methods in terms of re-identification. (*4*) proposes a training objective that focuses on learning consistent re-identification features over a sequence of frames.

### References

1. Bastani, F., He, S., Madden, S. (2021). Self-Supervised Multi-Object Tracking with Cross-input Consistency.
2. Tokmakov et al. Object Permanence Emerges in a Random Walk along Memory.
3. Boragule et al., Pixel-Guided Association for Multi-Object Tracking
4. Lang et al., Self-Supervised Multi-Object Tracking From Consistency Across Timescales