# Knowledge Distillation from Large Vision Foundation Models for Efficient Dense Prediction

Master/Eng internship

## General information

— **Duration :** 6 months (standard stipend). To start between February and March 2026.
— **Institute :** Université Paris Cité, Laboratoire d'Informatique Paris Descartes (LIPADE), team Systèmes Intelligents de Perception
— **Location :** 45 rue des Saints-Pères, 75006, Paris
— **Supervision :** Ayoub Karine [1] , Camille Kurtz [2], Laurent Wendling [3]
— **Keywords :** computer vision, deep learning, foundation models, knowledge distillation, dense prediction
— **Application :** send an email with CV to ayoub.karine@u-paris.fr, camille.kurtz@u-paris.fr and laurent.wendling@u-paris.fr with subject "[KD-LVFM]" (the application through message in Linkedin will not be considered)

## Proposed topic

### Context

Recently, several Large Vision Foundation Models (LVFMs) have been proposed in the literature [1]. They are trained through a Self-Supervised Learning (SSL) paradigm on large-scale unlabeled datasets and evaluated on small labeled datasets (fine-tuning). These models have achieved state-of-the-art performance across a wide range of downstream computer vision tasks, including both non-dense tasks (e.g., image classification, image retrieval) and dense tasks (e.g., semantic segmentation, object detection). However, the growing size and computational demands of the LVFMs significantly constrain their applicability in resource-limited devices (e.g., drone, smarphone). For instance, CLIP (Contrastive Language–Image Pretraining, OpenAI) [2] comprises up to 0.4 billion parameters, DINOv3 (self-DIstillation with NO labels, MetaAI) [3] includes models with up to 7 billion parameters, and the SAM 2 (Segment Anything Model, Meta AI) [4] exceeds 224 million parameters. To reduce the computational demands of such massive architectures, this internship will focus on investigating knowledge distillation techniques.

### Work to be done

The knowledge distillation (KD) technique [5, 6, 7] transfers knowledge from a powerful teacher network to a smaller student model, enabling the student to achieve significantly improved performance with lower computational cost. In this process, the student is trained on the same dataset as the teacher, allowing it to directly leverage the teacher's learned representations. However, directly applying KD to LVFMs presents several challenges. First, the most performant LVFMs are developed by large tech companies, and their training datasets are often not publicly available. Second, these LVFMs typically employ Vision Transformer (ViT) architectures [8] as encoders, whereas convolutional neural networks (CNNs) are generally lighter and more computationally efficient, making them strong candidates for student models on edge devices. Third, there are significant discrepancies in capacity between LVFMs and smaller edge models. The latter two challenges are partially addressed by Lee *et al.* [9], who propose a method to customize the well-generalized features of LVFMs for a given student model. Despite promising results, this work does not thoroughly address the issues of unavailable source datasets and cross-architecture knowledge transfer. Additionally, only the image classification task is considered. In this internship, we aim to tackle these challenges by investigating state-of-the-art methods for cross-architecture KD [10], data-free KD [11] and adaptive KD [12]. As illustrated in figure 1, we will focus on two dense down-stream tasks : semantic segmentation and object detection. The different steps of the internship are :

▷ **Step 1 – Literature review on KD from foundation models**
▷ **Step 2 – Compare different methods of cross-architecture KD, data-free KD and adaptive KD :** The teacher will be a LVFM such as CLIP, DINOv3 and SAM2. The student encoder should be a CNN one like ResNet18.
▷ **Step 3 – Test the student model on different semantic segmentation and object detection datasets :** A comparison is to be done with classical KD methods dedicated to dense prediction.

1. https://www.ayoub-karine.com/
2. https://www.camille-kurtz.com/
3. https://helios2.mi.parisdescartes.fr/ lwendlin/

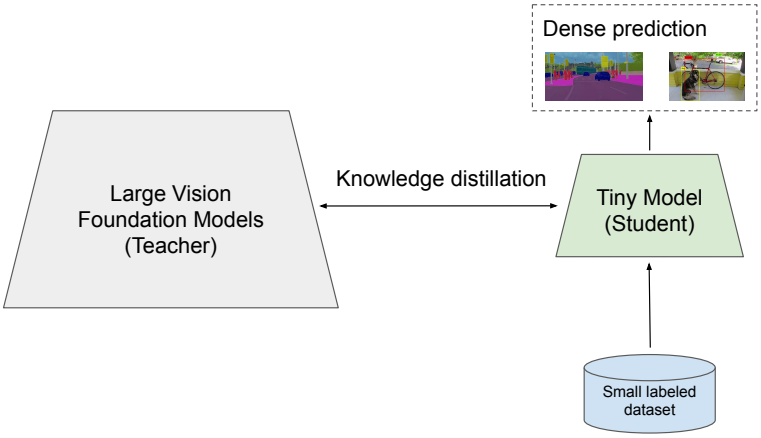


FIGURE 1 – Knowledge Distillation from Large Vision Foundation Models for Efficient Dense Prediction.

## Desired background for the candidate

We are looking for a Master 2 student or final year of MSc, or engineering school in computer science. The ideal candidate should have knowledge in deep learning, computer vision, Python programming and an interest in efficient deep learning.

## Bibliography


[1] Muhammad AWAIS et al. "Foundation models defining a new era in vision : a survey and outlook". In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[2] Alec RADFORD et al. "Learning transferable visual models from natural language supervision". In : *International conference on machine learning*. PmLR. 2021, p. 8748-8763.

[3] Oriane SIMÉONI et al. "Dinov3". In : *arXiv preprint arXiv :2508.10104* (2025).

[4] Nikhila RAVI et al. "Sam 2 : Segment anything in images and videos". In : *arXiv preprint arXiv :2408.00714* (2024).

[5] Geoffrey HINTON, Oriol VINYALS et Jeff DEAN. "Distilling the knowledge in a neural network". In : *arXiv preprint arXiv :1503.02531* (2015).

[6] Ayoub KARINE, Thibault NAPOLÉON et Maher JRIDI. "I2CKD : Intra- and inter-class knowledge distillation for semantic segmentation". In : *Neurocomputing* 649 (oct. 2025), p. 130791. DOI : 10.1016/j.neucom.2025.130791. URL : https://hal.science/hal-05144692.

[7] Ayoub KARINE, Thibault NAPOLÉON et Maher JRIDI. "Channel-spatial knowledge distillation for efficient semantic segmentation". In : *Pattern Recognition Letters* 180 (avr. 2024), p. 48-54. DOI : 10.1016/j.patrec.2024.02.027. URL : https://hal.science/hal-04488459.

[8] Alexey DOSOVITSKIY et al. "An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale". In : *International Conference on Learning Representations*. 2021. URL : https://openreview.net/forum?id=YicbFdNTTy.

[9] Jungsoo LEE et al. "Customkd : Customizing large vision foundation for edge model improvement via knowledge distillation". In : *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, p. 25176-25186.

[10] Weijia ZHANG et al. "Cross-Architecture Distillation Made Simple with Redundancy Suppression". In : *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, p. 23256-23266.

[11] Qianlong XIANG et al. "Dkdm : Data-free knowledge distillation for diffusion models with any architecture". In : *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, p. 2955-2965.

[12] Yichen ZHU et Yi WANG. "Student customized knowledge distillation : Bridging the gap between student and teacher". In : *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, p. 5057-5066.