

Projet Fil Rouge VitCow

Charlotte TIBI Damien Glaizal Mafalda FRERE Marie MEIER
Marius CHARLES

Sous la direction de
Joseph Allyndree

Agroparistech
Institut national des sciences et industries du vivant et de l'environnement

November 12, 2025

Plans

1. Introduction

Introduction

Introduction

Le projet choisit vise au développement d'un modèle d'auto-apprentissage de comportements de vaches sur de vidéos. Pour débutter nous allons étudiers les ressources fournies.

Sources fournies

Sources:

- A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 04, 2021, arXiv: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.
- J. J. Sun et al., “Video Foundation Models for Animal Behavior Analysis,” Jul. 31, 2024. doi: 10.1101/2024.07.30.605655.
- N. Ravi et al., “SAM 2: Segment Anything in Images and Videos”, [Online]. Available: <https://ai.meta.com/sam2>
- C. Zhou et al., “EdgeTAM: On-Device Track Anything Model,” Jan. 14, 2025, arXiv: arXiv:2501.07256. doi: 10.48550/arXiv.2501.07256.
- Z. Tong, Y. Song, J. Wang, and L. Wang, “VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training,” Oct. 19, 2022, arXiv: arXiv:2203.12602. doi: 10.48550/arXiv.2203.12602.

[1] A. Dosovitskiy et al.

Points clés:

- Application des Transformers (généralement utilisés pour les NLP) à des images. Découpage des images en Patches puis embedding des patches, ces derniers sont ensuite traités comme des tokens.
- De tels models entraînés sur de très grands jeux de données ont de très bonnes performances (supérieurs ou égales à celles des CNN classiques).

[2] J. J. Sun et al.

Points clés:

- Les méthodes déjà existantes sont spécifiques à une tâche et/ou à une espèce.
- Le but de la publication est d'évaluer l'intérêt du développement d'un "video foundation model".
- En classification, Retrieval et localisation un tel modèle augmente les performances et facilite le fine tuning.

Points clés:

- Sam 2: modèle généraliste de segmentation promptable avec ajout d'une mémoire temporelle.
- Les prompts sont faits sous forme de "clic sur un pixel" ou du dessin d'un cadre autour de l'élément à segmenter.
- Le modèle retient les éléments et les positions des éléments segmentés sur la frame précédente.
- L'architecture est un transformer.

[4] C. Zhou et al.

Points clés:

- EdgeTam: modèle plus léger que Sam 2. Il peut tourner sur un smartphone ou tout autres appareils embarqué.
- Pour limiter le cout computationnel et la mémoire nécessaire seulement certains pixels clés sont gardés en mémoire et la représentation de ces points est différente.
- Il y aussi un processus de distillation de Sam 2 vers EdgeTam

Points clés:

- C'est une méthode de pré-entraînement auto supervisé pour la vidéo.
- Le modèle apprend des représentations générales qui permettent un apprentissage plus rapide et avec moins de données labélisées par la suite.
- Pour l'entraînement: masquage d'une grande partie des patches spatio-temporels et le modèle apprend à reconstruire les parties manquantes.

Merci de votre écoute!

charlotte.tibi@agroparistech.fr






damien.glaizal@agroparistech.fr

mafalda.frere@agroparistech.fr

marie.meier@agroparistech.fr

marius.charles@agroparistech.fr

References

-  A. Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, Jun. 04, 2021, arXiv: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.
-  J. J. Sun et al., *Video Foundation Models for Animal Behavior Analysis*, Jul. 31, 2024. doi: 10.1101/2024.07.30.605655.
-  N. Ravi et al., *SAM 2: Segment Anything in Images and Videos*, Jul. 31, 2024. doi: 10.1101/2024.07.30.605655. <https://ai.meta.com/sam2>
-  C. Zhou et al., *EdgeTAM: On-Device Track Anything Model*, Jan. 14, 2025, arXiv: arXiv:2501.07256. doi: 10.48550/arXiv.2501.07256.
-  Z. Tong, Y. Song, J. Wang, and L. Wang, *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*, Oct. 19, 2022, arXiv: arXiv:2203.12602. doi: 10.48550/arXiv.2203.12602.