Bachelor Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Tracking in the Age of SAM, Dino, Monodepth and CLIP

Ruslan Rozumnyi

Supervisor: Prof. Ing. Jiří Matas, Ph.D.
Field of study: Open Informatics
Subfield: Artificial Intelligence and Computer Science
January 2025

Acknowledgements

I would like to express my deepest gratitude to my supervisor, prof. Ing. Ji°í Matas, Ph.D., for the time and patience he dedicated to me throughout this thesis. This work would not have been possible without his excellent guidance.

I am also thankful to all my VRG collaborators for creating a supportive and warm environment, especially Nikos Ypsilantis, Yash Patel, and my colleagues from the G-9 office, including Mat¥j, Tomá². Special thanks go to Denys Rozumnyi for his invaluable suggestions in writing the thesis as well as for all his unwavering support.

Also, I am profoundly grateful to my parents and entire family for their love and encouragement. Last but not least, I would like to thank Alexandra for her support.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses. I have used Grammarly, Writefull, and ChatGPT for grammar correction, paraphrasing and reformulation of my own text.

Prohla²uji, ºe jsem p°edloºenou práci vypracoval samostatn¥ a ºe jsem uvedl ve²keré pouºité informa£ní zdroje v souladu s Metodickým pokynem o dodrºování etických princip· p°i p°íprav¥ vysoko²kolských záv¥re£ných prací. Pouºil jsem Grammarly, Writefull a ChatGPT pro opravu gramatiky, parafrázování a p°eformulování svého vlastního textu.

In Prague, 7 January 2025

.... Ruslan Rozumnyi

Abstract

This work presents several approaches to enhance the Segment Anything Model 2 (SAM2), a foundation model for image and video segmentation, particularly in challenging tracking scenarios. The limitations of SAM2 are outlined, followed by a comprehensive analysis of its architecture and components.

Key findings reveal that SAM2 struggles with accuracy and robustness when tracking small or textureless objects. To address this, a cropping-based approach is proposed, which focuses on maintaining a window around the object throughout the video, yielding notable improvements on the Visual Object Tracking (VOT) challenge datasets.

Additionally, we demonstrate that incorporating optical flow into the mask selection process of the SAM2 mask decoder shows improvement in scenarios involving slow and smooth object motion. However, under general conditions, the integrated optical flow lacked accuracy.

Finally, two architectural modifications are explored: increasing the memory stride in the memory bank and excluding empty mask predictions from the memory bank. These adjustments consistently enhance tracking robustness and overall model performance across various tracking scenarios.

Keywords: tracking, segmentation, foundation model, cropping, optical flow, segment anything model

Supervisor: Prof. Ing. Jiří Matas, Ph.D.

Abstrakt

Tato práce představuje několik přístupů ke zlepšení modelu Segment Anything Model 2 (SAM2), základního modelu pro segmentaci obrazu a videa, zejména v náročných scénářích sledování. Nejprve jsou identifikována omezení modelu SAM2, po nichž následuje podrobná analýza jeho architektury a komponent.

Klíčová zjištění ukazují, že model SAM2 má problémy s přesností a robustností při sledování malých objektů nebo objektů s chudě definovaným povrchem. K řešení tohoto problému jsme navrhli přístup založený na vyřezávání, který se zaměřuje na udržení okna kolem objektu během celého videa, což přináší významná zlepšení na datových sadách Visual Object Tracking (VOT).

Dále demonstrujeme, že zapojení optického toku do procesu výběru masky v dekodéru SAM2 přináší zlepšení ve scénářích zahrnujících pomalý a plynulý pohyb objektů. Nicméně za obecných podmínek integrovaný optický tok postrádal dostatečnou přesnost.

Na závěr jsou zkoumány dvě úpravy v architektuře SAM2: zvýšení memory stride v memory bance a vyloučení prázdných masek z memory banky. Tyto úpravy významně zlepšují robustnost sledování a celkový výkon modelu v různých scénářích sledování.

Klíčová slova: sledování, segmentace, základní model, vyřezávání, optický tok, segment anything model

Překlad názvu: Sledování objektu s pomocí SAM, Dino, Monodepth nebo CLIP reprezentace

Contents

Figures

Tables