Research in AI

Vlad Shakhuro











29 October 2026

Outline

1. Reading papers
2. Getting into research

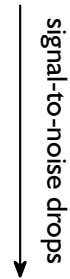
Where to find papers?

Literature review:

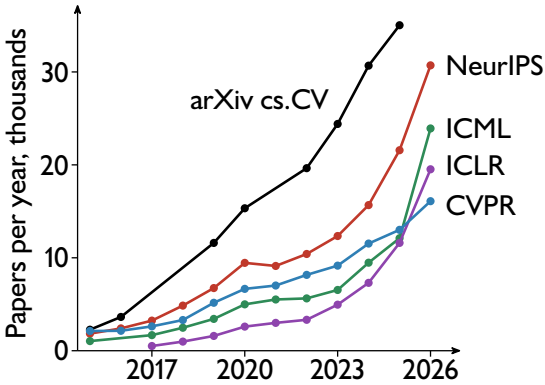
- Google Scholar, Semantic Scholar
- citation graphs: Connected Papers, ResearchRabbit
- deep research: Elicit, Undermind, ChatGPT, Gemini (check every source)

Fresh papers:

- Scholar alerts: favorite authors
- Scholar Inbox: learns your taste
- HF Daily Papers, alphaXiv
- X, Bluesky
- CVF Open Access, OpenReview (with reviews)
- raw arXiv



Which papers are worth reading?



Filters cut the flow 10–100×:

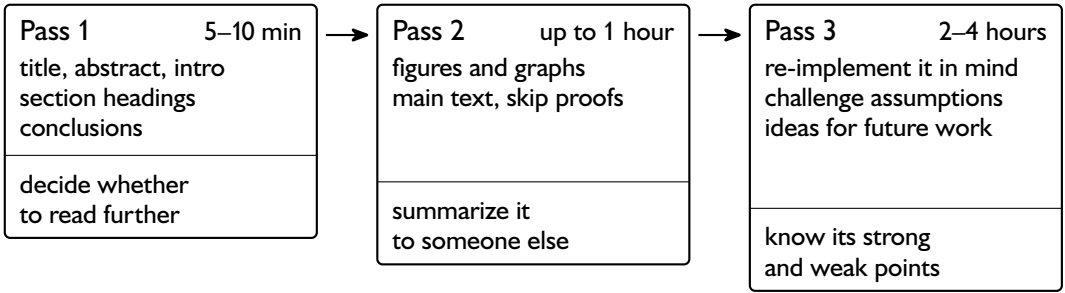
- A* venue (Core)

ML	NeurIPS, ICML, ICLR
CV	CVPR, ICCV, ECCV
NLP	ACL, EMNLP
other	SIGGRAPH, ICRA

- top lab or favorite author
- your area
- opens a new area (ViT, NeRF)
- large-scale benchmark
- crowded area? think twice

How to read a paper?

Read in up to three passes, each going deeper:



What will be your **output**?

- for yourself → write reviews, optionally publish online
- for lecture, seminar → slides and feedback from listeners

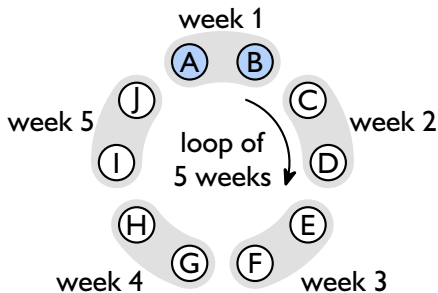
Keshav. How to read a paper. ACM SIGCOMM Computer Communication Review 2007

Organizing seminars

Shallow seminar: many papers, briefly

- weekly, 1 hour, fixed time
- 2 speakers, 2–3 papers each
- 10+ speakers in a looped queue
- shared paper pool, e.g. a Zotero group
- 1 organizer

Each speaker presents every 5 weeks
4–6 papers per week, 100+ per year



Writing reviews

Plan	Example review
1. Title, links	Cutie, arXiv:2310.12982, github.com/hkchengrex/Cutie
2. Short description	Segment objects in video given first-frame masks
3. Conference	CVPR 2024, highlight
4. Novelty	Reads memory per object, not only per pixel
5. Method	Object queries ↔ pixel features, masked attention
6. Evaluation	MOSE $\mathcal{J}\&\mathcal{F}$: 68.3 vs 59.6 XMem, 64.1 DeAOT
7. Pros, cons	+ robust to distractors; − fails on overlapping look-alikes
8. Production	Online, 36.4 FPS on V100, code public

Keep reviews searchable: Zotero notes, Obsidian, a team wiki

Cheng et al. Putting the Object Back into Video Object Segmentation. CVPR 2024

Reading a paper



Putting the Object Back into Video Object Segmentation

Ho Kei Cheng[1] Seoung Wug Oh[2] Brian Price[2] Joon-Young Lee[2] Alexander Schwing[1]
[1]University of Illinois Urbana-Champaign [2]Adobe Research
{hkchengrex,aschwing}@illinois.edu, {seoh,bprice,jolee}@adobe.com

Abstract

We present Cutie, a video object segmentation (VOS) network with object-level memory reading, which learns to represent information from memory back into the video object segmentation result. Recent works on VOS employ bottom-up pixel-level memory reading which struggles due to matching noise, especially in the presence of distractors, resulting in a linear dependency between memory size and noise. Cutie performs top-down object-level memory reading by adapting a small set of object queries. Via these, it interacts with the bottom-up pixel features iteratively with a query-based object transformer (qt, hence Cutie). The object queries act as high-level summaries of the target object while high-resolution pixel features are retained for accurate segmentation. Together with foreground-background object-level memory reading, Cutie cleanly separates the semantics of the foreground object from the background. On the challenging MOSE dataset, Cutie improves by 8.7 J&F over XMem with a similar running time and improves by 4.2 J&F over DeAOT while running three times faster. Code is available at: hkchengrex.github.io/Cutie

1. Introduction

Video Object Segmentation (VOS), specifically the "semi-supervised" setting, requires tracking and segmenting objects in a video given the segmentation mask of the first frame(s). This task methods are broadly applicable in robotics [1], video editing [2], reducing costs in data annotation [3], and can also be combined with Segment Anything Models (SAMs) [4] for universal video segmentation (e.g., Tracking Anything [5–7]).

Recent VOS approaches employ a memory-based framework [8–17] which stores segmentation information from past segmented frames (either given as input or segmented by the model), and any new query frame "reads" from this memory to retrieve features for segmentation. Importantly, these approaches mainly use *pixel-level matching* for memory reading, either with one [8] or multiple matching layers [16], and generate the segmentation bottom-up from the pixel memory readout. Pixel-level matching maps every query pixel independently to a linear combination of memory pixels (Fig. 1, top). Consequently, the resulting pixel-level matching lacks high-level consistency and is prone to matching noise, especially in the presence of distractors. This leads to lower performance in challenging scenarios with occlusions and frequent distractors. Concretely, the performance of recent methods [8, 16, 17] drops by over 10 points in $\mathcal{J}\&\mathcal{F}$ *better* when evaluating on the recently proposed challenging MOSE [12] dataset rather than the simpler DAVIS 2017 [18] dataset.

We think this unsatisfactory result in challenging scenarios is caused by the lack of object-level reasoning. To address this, we propose *object-level memory reading*, which reads from memory by querying object-level information (Fig. 1, bottom). Inspired by recent query-based object detection/segmentation [14–16] that represent objects as "object queries," we implement our object-level memory reading with an object transformer. This object transformer uses a small set of role-assigned object queries to (1) iteratively probe and calibrate a *loose* map (initialized by a pixel-level

arXiv:2310.12982v2 [cs.CV] 11 Apr 2024

Method paper



MOSE: A New Dataset for Video Object Segmentation in Complex Scenes

Henghui Ding[1] Chang Liu[2] Shuting He[2] Xudong Jiang[1] Philip H.S. Torr[3] Song Bai[4]
[1]Nanyang Technological University [2]Zhejiang University [3]University of Oxford [4]ByteDance
https://henghuiding.github.io/MOSE/

Figure 1. Examples of video clips from the newly collected MOSE (coMplex video Object SEgmentation) dataset. The selected target objects are marked in colors. ■. The most notable feature of MOSE is complex scenes, including the disappearance-reappearance of objects, small/incomplete objects, heavy occlusions, crowded environments, *etc*. For example, the target player in the 2nd row turns around when reappearing to the camera, and thus his appearance changes greatly. The 3rd row shows a crowded scene with many similar objects moving here. More videos in MOSE contain crowded and occluded objects with the target object seldom being the salient one. The goal of MOSE dataset is to provide a platform that promotes the development of more comprehensive and robust video object segmentation algorithms.

Abstract

*Video object segmentation (VOS) aims at segmenting a particular object throughout the entire video clip sequence. The state-of-the-art VOS methods have achieved excellent performance (*e.g.*, 90+% J&F) on existing datasets. However, since the target objects in these existing datasets are usually relatively salient, dominant, and isolated, VOS under complex scenes has rarely been studied. To track VOS and make it more applicable in the real world, we collect a new VOS dataset called coMplex video Object SEgmentation (MOSE) to study the tracking and segmenting objects in complex environments. MOSE contains 2,149 video clips and 5,200 objects from 36 categories, with 431,725 high-quality object segmentation masks. The most notable feature of MOSE dataset is complex scenes with crowded and occluded objects. The target objects in the VOS benchmark are very hard to be fully segmented in some frames. To analyze the proposed MOSE dataset, we benchmark 18 existing VOS methods under 4 different settings on the proposed MOSE dataset and conduct comprehensive comparisons. The experiments show that current VOS algorithms cannot well perceive objects in complex scenes. For example, under the semi-supervised VOS setting, the highest $\mathcal{J}\&\mathcal{F}$ by existing state-of-the-art VOS methods is only 59.4% on MOSE, much lower than their ~90% $\mathcal{J}\&\mathcal{F}$ performance on DAVIS. The results reveal that although existing methods have achieved great success in existing benchmarks, there are unresolved challenges under complex scenes and more efforts are desired to explore these challenges in the future. The proposed MOSE dataset has been released at https://henghuiding.github.io/MOSE/.*

1. Introduction

Video object segmentation (VOS) [1, 2, 7] aims at segmenting a particular object, *e.g.*, the dominant objects or the objects indicated by users, throughout the entire video clip sequence. It is a crucial and fundamental task in many computer vision tasks and has a crucial role in many practical applications that involve video analysis and understanding, *e.g.*, self-driving vehicles, augmented reality, video editing, *etc*. There are many different settings for VOS, for example, semi-supervised VOS [1, 2, 7] that gives the frame mask of the target object, unsupervised VOS [3, 7] that automati-

†Henghui.ding@gmail.com

arXiv:2302.01872v1 [cs.CV] 3 Feb 2023

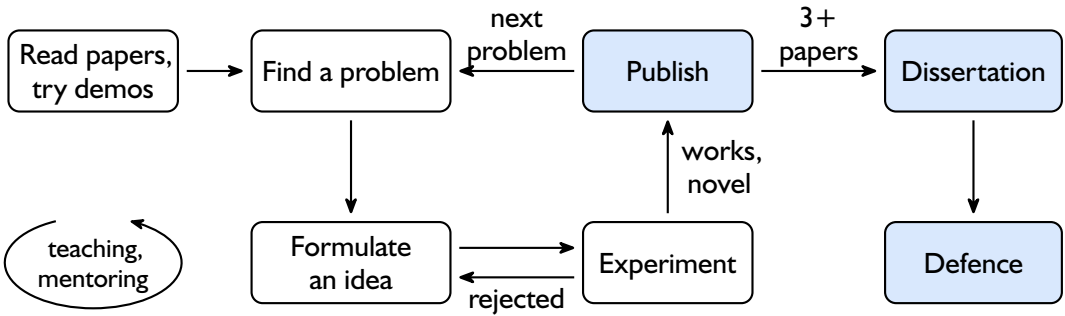
Dataset paper

Cheng et al. Putting the Object Back into Video Object Segmentation. CVPR 2024
Ding et al. MOSE: A New Dataset for Video Object Segmentation in Complex Scenes. ICCV 2023

Outline

1. Reading papers
2. Getting into research

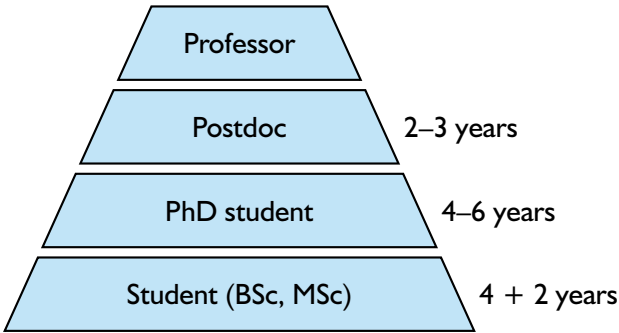
What is a PhD?



Typical length: aspirantura in Russia 4 years, PhD in the US 5–6 years

Academia vs industry

Academia hierarchy:



Industry vs academia:

- more resources
- less freedom
- less teaching
- higher salary
- goal: product, not papers

Criteria for a good research lab

Criterion	How to check
Quality of publications	Members' profiles in Google Scholar A* papers and demos at specialized events
Supportive supervisor and group	Talk to current students Where alumni went How often meetings happen
Resources	GPUs per student, AI subscriptions (ChatGPT, Claude) Funding: grants, industry projects Co-authors from other labs on papers

Conclusion

We reviewed two topics:

- papers: how to find and read them
- how research is organized in academia