# View Reviews

**Paper ID**
8972

**Paper Title**
Multiscale Memory Comparator Transformer for Few-Shot Video Segmentation

**Reviewer #1**

## Questions

**1. [Reviewer Guidelines] By taking this review assignment and checking on "I agree" below, I acknowledge that I have read and understood the reviewer guidelines (https://iccv2023.thecvf.com/reviewer.guidelines-362000-2-16-20.php).**
Agreement accepted

**2. [Large Language Model (LLM) Ethics] ICCV'23 does not allow the use of Large Language Models or online chatbots such as ChatGPT in any part of the reviewing process. There are two main reasons: - Reviewers must provide comments that faithfully represent their original opinions on the papers being reviewed. It is unethical to resort to Large Language Models (e.g., an offline system) to automatically generate reviewing comments that do not originate from the reviewer's own opinions. - Online chatbots such as ChatGPT collect conversation history to improve their models. Therefore their use in any part of the reviewing process would violate the ICCV confidentiality policy (https://iccv2023.thecvf.com/ethics.for.reviewing.papers-362100-2-16-21.php). Herewith I confirm that I have not used an online chatbot such as ChatGPT in preparing the review. This review reflects my own opinions, and no parts were generated by an automatic system.**
I agree

**4. [Summary] Describe the key ideas, experiments, and their significance (preferably in 5-7 sentences).**
The authors introduce a multiscale comparison framework for few-shot video object segmentation (FS-VOS) and automatic video object segmentation (AVOS). Specifically, a multiscale memory decoder is proposed for support-query comparison. Two types of information exchange across scales are also investigated to find optimal model design. The proposed method outperforms existing SOTA methods on FS-VOS and AVOS benchmarks.

**5. [Strengths] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these aspects of the paper are valuable. Short, unjustified review is NOT OK.**
[s1] Employing multiscale comparison scheme for support-query comparison is quite intuitive and interesting. The quantitative evaluation also backups its effectiveness.

[s2] Evaluation on two different tasks (FS-VOS and AVOS) makes the proposed method technically solid.

[s3] The proposed method outperforms existing SOTA methods on public benchmarks.

**6. [Weaknesses] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these are weak aspects of the paper. Short, unjustified review is NOT OK. For example: a comment saying that this has been done before must be accompanied by specific reference(s) and an explanation of how they overlap; a comment saying that the experiments are**

**insufficient to validate the claims should be accompanied by explanations of what exactly is missing; not achieving state-of-the-art, or not surpassing prior methods, is not sufficiently a weakness, unless this comment is justified; a comment of “lack of novelty” should be carefully justified, and novelty could involve new insights/understandings or new discoveries/observations, not just new technical methods.**
[w1] The proposed components are not effective. Although the main contribution of this paper is "MMemory" that helps FS-VOS performance, there is no improvement at all compared to the baseline model according to Table 4. The performance gain comes from other tricks such as using larger memory size or additional refinement step. Can multiscale memory decoding scheme really reduce confusion between the background and novel class?

[w2] Evaluation on AVOS (Table 2) is not a fair comparison. First, unsupervised video object segmentation (UVOS) and camouflaged object detection (COD) are funtamentally different tasks. UVOS aims at finding the most distinctive object in a scene, while objects to detect are not distinctive in COD. Therefore, directly evaluating "existing VOS methods [1,2,3,4,5] trained on VOS datasets" on COD datasets is not valid. Second, unlike these VOS methods, the proposed method is trained on a larger training samples. Have you tried learning the proposed method using the same training dataset? Third, the compared VOS methods are not SOTA anymore. Some recent works [6,7,8,9] should also be reported.

[1] Learning unsupervised video object segmentation through visual attention, CVPR 2019
[2] See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks, CVPR 2019
[3] Zero-Shot Video Object Segmentation via Attentive Graph Neural Networks, ICCV 2019
[4] Motion-Attentive Transition for Zero-Shot Video Object Segmentation, AAAI 2020
[5] Reciprocal Transformations for Unsupervised Video Object Segmentation, CVPR 2021
[6] Iteratively Selecting an Easy Reference Frame Makes Unsupervised Video Object Segmentation Easier, AAAI 2022
[7] Hierarchical Feature Alignment Network for Unsupervised Video Object Segmentation, ECCV 2022
[8] Treating Motion as Option to Reduce Motion Dependency in Unsupervised Video Object Segmentation, WACV 2023
[9] Unsupervised Video Object Segmentation via Prototype Memory Network, WACV 2023

**7. [Paper rating]**
4. Weak reject

**8. [Recommendation confidence]**
Very confident: I am an expert on this topic.

**9. [Justification of rating] Provide detailed justification of your rating. It should involve how you weigh the strengths and weaknesses of the paper.**
I am not sure about the effectiveness of the proposed method. Evaluation on AVOS is not appealing as well.

**10. [Additional comments] Minor suggestions, questions, corrections, etc. that can help the authors improve the paper, if any.**
[a1] Overall writings are hard to follow.

[a2] Figure 3 (which is the main visualization of the proposed algorithm) is really hard to understand.

**11. [Dataset contributions] Does a paper claim a dataset release as one of its scientific contributions?**
No

**12. [Post-Rebuttal Recommendation] Give your final rating for this paper. Don't worry about poster vs oral. Consider the input from all reviewers, the authors' feedback, and any discussion. (Will be visible to authors**

**after author notification)**
5. Weak Reject

**13. [Post-Rebuttal Justification] Justify your post-rebuttal assessment. Acknowledge any rebuttal and be specific about the final factors for and against acceptance that matter to you. (Will be visible to authors after author notification)**
According to Table 4, it seems that "MMemory itself" brings negligible performance gain on FSVS task. Although it may boost AVOS performance, I think that cannot be a valid contribution, as MMemory is designed for FSVS (as in title).

Regarding evaluation on AVOS, I think comparison with other methods is quite unfair. The proposed memory uses YouTube-VOS dataset, which is one of the largest video segmentation datasets, as training dataset. However, only MATNet uses YouTube-VOS for network training in Table 2. In addition, the reported performance is based on a recent backbone network (Swin-Video). If ResNet backbone is used as other methods reported in Table 2, the performance significantly degrades (Table 1 and Table 2 in Supp.). Furthermore, I have checked four VOS papers that I listed on my review, and all of them are turned out to be published before ICCV submission date.

I think this manuscript proposes some interesting ideas, but they are not validated correctly and do not bring clear advantages. Therefore I recommend a WR.

**Reviewer #2**

---

## Questions

**1. [Reviewer Guidelines] By taking this review assignment and checking on "I agree" below, I acknowledge that I have read and understood the reviewer guidelines (https://iccv2023.thecvf.com/reviewer.guidelines-362000-2-16-20.php).**
Agreement accepted

**2. [Large Language Model (LLM) Ethics] ICCV'23 does not allow the use of Large Language Models or online chatbots such as ChatGPT in any part of the reviewing process. There are two main reasons: - Reviewers must provide comments that faithfully represent their original opinions on the papers being reviewed. It is unethical to resort to Large Language Models (e.g., an offline system) to automatically generate reviewing comments that do not originate from the reviewer's own opinions. - Online chatbots such as ChatGPT collect conversation history to improve their models. Therefore their use in any part of the reviewing process would violate the ICCV confidentiality policy (https://iccv2023.thecvf.com/ethics.for.reviewing.papers-362100-2-16-21.php). Herewith I confirm that I have not used an online chatbot such as ChatGPT in preparing the review. This review reflects my own opinions, and no parts were generated by an automatic system.**
I agree

**4. [Summary] Describe the key ideas, experiments, and their significance (preferably in 5-7 sentences).**
This paper proposes a meta-learned multiscale memory comparator Transformer for few-shot video segmentation task. Compared to prior works that leverage a single feature layer for support and query feature comparison, this work explores multiple forms of information exchange across scales aiming to preserve the detailed feature maps and further reduce confusion between background and novel class. The proposed method outperforms few-shot video objective segmentation (FS-VOS) baselines and the adapted version on fully supervised automatic VOS achieves good results as well.

**5. [Strengths] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these aspects of the paper are valuable. Short, unjustified review is NOT OK.**

- This paper is well-written and the method is easy to follow.
- The results are strong and surpass the prior state-of-the-art.

**6. [Weaknesses] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these are weak aspects of the paper. Short, unjustified review is NOT OK. For example: a comment saying that this has been done before must be accompanied by specific reference(s) and an explanation of how they overlap; a comment saying that the experiments are insufficient to validate the claims should be accompanied by explanations of what exactly is missing; not achieving state-of-the-art, or not surpassing prior methods, is not sufficiently a weakness, unless this comment is justified; a comment of “lack of novelty” should be carefully justified, and novelty could involve new insights/understandings or new discoveries/observations, not just new technical methods.**

- The motivation from FS-VOS to AVOS is unclear. The target of the paper is to deal with few-shot video segmentation by using the multiscale memory comparator (MMC) between support-query features, but the setting of AVOS has no support-query set and therefore MMC is not applicable.
- The claim "the potential existence of erroneous correlation at different scales" in L737-L738 needs to be supported by references or evidences (e.g. visualization). If this is a false assumption, it would make no sense for the conclusion made by the proposed bidirectional information exchange.
- The motivation of the adaptive K-shot scheme is not detailed in L557-L559.
- Ablation on the number of memory entries indicates a trivial improvement from N=2 to 20 in Table 5. What is the possible reason and what is the impact of number N? More detailed ablation with N between 2 to 20 would be helpful.

**7. [Paper rating]**

2. Weak accept

**8. [Recommendation confidence]**

Somewhat confident: I do not directly work on this topic, but my expertise and experience are sufficient to evaluate this paper.

**9. [Justification of rating] Provide detailed justification of your rating. It should involve how you weigh the strengths and weaknesses of the paper.**

Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations

**10. [Additional comments] Minor suggestions, questions, corrections, etc. that can help the authors improve the paper, if any.**

See Weaknesses

**11. [Dataset contributions] Does a paper claim a dataset release as one of its scientific contributions?**

No

**12. [Post-Rebuttal Recommendation] Give your final rating for this paper. Don't worry about poster vs oral. Consider the input from all reviewers, the authors' feedback, and any discussion. (Will be visible to authors after author notification)**

4. Borderline Reject

**13. [Post-Rebuttal Justification] Justify your post-rebuttal assessment. Acknowledge any rebuttal and be specific about the final factors for and against acceptance that matter to you. (Will be visible to authors after author notification)**

My major concern is about the AVOS part. However, the author feedback did not convince me. Instead, it makes me

more confused about the effectiveness of the proposed method.
This work aims to deal with FS-VOS but the authors mentioned "AVOS is a simper setup without these challenges" so they used it. If this is the case, the results of AVOS can not provide support for the effectiveness of MMC because it involves no usage of the multiscale comparator.
Besides, "the current FS-VOS task is under-explored with only one" is a weak reason for AVOS.

Even if this work introduces some interesting ideas and conducted extensive experiments, the experimental results are not convincing enough to verify the effectiveness of the proposed method. So I decreased my rating.

**Reviewer #3**

## Questions

**1. [Reviewer Guidelines] By taking this review assignment and checking on "I agree" below, I acknowledge that I have read and understood the reviewer guidelines (https://iccv2023.thecvf.com/reviewer.guidelines-362000-2-16-20.php).**
Agreement accepted

**2. [Large Language Model (LLM) Ethics] ICCV'23 does not allow the use of Large Language Models or online chatbots such as ChatGPT in any part of the reviewing process. There are two main reasons: - Reviewers must provide comments that faithfully represent their original opinions on the papers being reviewed. It is unethical to resort to Large Language Models (e.g., an offline system) to automatically generate reviewing comments that do not originate from the reviewer's own opinions. - Online chatbots such as ChatGPT collect conversation history to improve their models. Therefore their use in any part of the reviewing process would violate the ICCV confidentiality policy (https://iccv2023.thecvf.com/ethics.for.reviewing.papers-362100-2-16-21.php). Herewith I confirm that I have not used an online chatbot such as ChatGPT in preparing the review. This review reflects my own opinions, and no parts were generated by an automatic system.**
I agree

**4. [Summary] Describe the key ideas, experiments, and their significance (preferably in 5-7 sentences).**
This paper presents a new method coalled a meta-learned Multi-scale Memory Comparator (MMC) for few-shot video segmentation. This task is highly similar to conventional few-shot segmentation except for the input is video. As the name tells, the proposed method attempts to preserve the detailed multi-scale information by the proposed memory memory transformer. The proposed method is primarily evaluated on few-shot VOS while some results for fully supervised setting are also reported. This method attains SOTA performance for both.

**5. [Strengths] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and data contribution (if relevant). Explain clearly why these aspects of the paper are valuable. Short, unjustified review is NOT OK.**
1. Strong results. All the results clearly outperform other works.

2. Ablation studies are thorough and carefully conducted.

3. I am impressed by the details and additional experiments the authors included in the supplementary material. Although I haven't read all of them, I checked the section 4.3 in ablation study, and I highly appreciate these careful analysis

**6. [Weaknesses] Consider the aspects of key ideas, experimental or theoretical validation, writing quality, and**

**data contribution (if relevant). Explain clearly why these are weak aspects of the paper. Short, unjustified review is NOT OK. For example: a comment saying that this has been done before must be accompanied by specific reference(s) and an explanation of how they overlap; a comment saying that the experiments are insufficient to validate the claims should be accompanied by explanations of what exactly is missing; not achieving state-of-the-art, or not surpassing prior methods, is not sufficiently a weakness, unless this comment is justified; a comment of “lack of novelty” should be carefully justified, and novelty could involve new insights/understandings or new discoveries/observations, not just new technical methods.**
1. First, this is not a weakness but more like a quesiton. So if I understood correctly, from L96 to L122, the paper explains its motivation. In the beginning of this paragraph, it says the key question is what features should be compared? shallow or deep? -> To address this, the paper says multi-scale comparator is proposed.

However, in the next sentences, the paper points out that [20,5] proposed multiscale architecture to tackle the task, while they do not consider the time dimension, which is unique to video. So what is actually proposed to consider time dimension explicitly in this method? The only consideration seems to be simply the change of dimensions of inputs from for example H x W x C to T x H x W x C. As the name of the proposed module tells (MMC), it seems that the temporal dimension is simply handled by feeding the multiple images? So if the time dimension is only considered by the inputs, I am not fully convinced why this method tackles the few-shot "VOS". It seems what this method does is just few-shot segmentation.

From the last summary of the contributions in L146-L161, I can't find any contribution that actually considers the temporal axis, which is one of the most important additional component that differs from image-based tasks.

2. In table 1, it seems that few-shot image-based segmentation methods are used as competitors except [3]. [40,29,17] are all image based methods, but their performance is quite low compared to current SOTA in their task. It is good that the proposed method exceeds [3] by large margin (I'm not sure if there exists any more recent works directly tackling FS-VOS. But i assume there should be because [3] is released in 2021, which is 2 years ago. This is a significant amount of time in computer vision field). However, I am questioning whether the proposed method is actually compared to the 'right' competitors.

3. I am not familiar with the datasets FS-VOS commonly use for evaluation. But from the qualitative comparisons, it seems like query clips mostly include only single object, which seems relatively easy. Is there more challenging-looking like visualization?

**7. [Paper rating]**
3. Borderline

**8. [Recommendation confidence]**
Somewhat confident: I do not directly work on this topic, but my expertise and experience are sufficient to evaluate this paper.

**9. [Justification of rating] Provide detailed justification of your rating. It should involve how you weigh the strengths and weaknesses of the paper.**
Currently, I am slightly leaning towards reject as I have a small concern with the methods the authors chose to compare with. If this is clarified, I am willing to increase my rating.

**10. [Additional comments] Minor suggestions, questions, corrections, etc. that can help the authors improve the paper, if any.**
See weakness above

**11. [Dataset contributions] Does a paper claim a dataset release as one of its scientific contributions?**
No

**12. [Post-Rebuttal Recommendation] Give your final rating for this paper. Don't worry about poster vs oral. Consider the input from all reviewers, the authors' feedback, and any discussion. (Will be visible to authors after author notification)**
4. Borderline Reject

**13. [Post-Rebuttal Justification] Justify your post-rebuttal assessment. Acknowledge any rebuttal and be specific about the final factors for and against acceptance that matter to you. (Will be visible to authors after author notification)**
My major concern is that barely any module is designed or proposed for FS"VS".
I personally believe this work barely contributes anything to "Video" field, which I find as a weakness.
After reading R1's comment, I was able to find that their proposed MM is not much effective for FSVS, which further confirms my point.