MMC Transformer: Multiscale Multigrid Comparator Transformer for Few-Shot Video Segmentation

Supplementary Material

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Our supplemental material is organized into six major sections. Section 1 documents an associated supplementary video. Section 2 provides additional experimental design and implementation details that were not described in the main submission. Section 3 discusses additional results in terms of an ablation study on the number of memory entries used and the attention maps visualised along with a frequency analysis that motivates the multigrid formulation. Section 4 documents the attached code and how to reproduce YouTube-VIS FSVOS experiments. Finally, Sections 5 and 6 provide the licenses of the different assets and discusses the societal impact of our downstream task.

1 Supplemental video

We include an accompanying video as part of the supplementary materials. In this video, we show eight examples for few-shot video object segmentation on YouTube-VIS and three examples for few-shot common action localisation on A2D. Additionally, we show the attention maps output from the different memory entries for the finest scale across the entire input clip. The video is in MP4 format and is approximately eight minutes long. Layouts for each sampling pair are described in detail followed by the example video samples. The codec used for the realization of the provided video is H.264 (x264).

2 Experiment design details

In this section, we detail the experiment setup and implementation details that were not described in the main submission.

Evaluation. In the few-shot video object segmentation task we use the YouTube-VIS [3] dataset that is split into four folds, each with 30 training classes and 10 testing classes. In the few-shot common action localisation we use Common A2D [9] that has 43 classes split into 33, five and five classes for the training, validation and testing splits, resp. We choose as a baseline comparison algorithm a previous approach that made use of multiscale processing relying on correlation tensors between the support and query sets using features from different scales [5]; however, our baseline does not use a transformer based approach. We follow previous evaluations by reporting the mean intersection over union (mIoU) for one-way one-shot and five-shot evaluations [3, 9]. Evaluation is performed as the average over five runs for YouTube-VIS FS-VOS, similar to [3]. For the evaluation on Common A2D we report results from one run, following previous work [9].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Method	mIoU				
	1	2	3	4	Mean
$N = 2$	49.6	70.3	62.9	**65.1**	61.98
$N = 20$	**51.5**	**70.6**	**63.0**	64.6	**62.4**

Table 1: Ablation study on the number of memory entries in our multiscale memory learning mechanism. This study is performed on YouTube-VIS FS-VOS folds 1, 2, 3 and 4 with a five shot support set.

Implementation details. We follow standard few-shot segmentation and few-shot video segmentation practice of assigning the novel class objects that exist in training images to background [8]. During training on YouTube-VIS FS-VOS, we randomly sample five frames from the query set video and randomly sample one image from another video for the support set. In Common A2D, we use 25 frame trimmed clips subsampled uniformly into five frames for the support and query sets during training. During inference, in YouTube-VIS FS-VOS we infer over the entire video. While in Common A2D during inference, we use 25 frame clips in a temporal sliding window over the entire untrimmed query video. The features that are used to compute the correlation tensors in both tasks are extracted as follows. We extract features at the end of each bottleneck before the ReLU activation in the last three stages and concatenate the correlation tensors (with same spatial dimensions) along the channel dimension; this operation results in a three scale hypercorrelation pyramid, $P = 3$. Our MMC transformer is constructed as seven decoding layers with separate weights each corresponding to one communication exchange across scale with bidirectional exchange. We go through coarse-mid-fine-mid-coarse-mid-fine in that order, which results in the final enhanced feature maps at the finest scale. Finally, we train our models on a single NVIDIA RTX A6000 GPU using batch size 19.

3 Additional empirical results

In this section, we provide additional empirical results to understand the multiscale memory learning and confirm the motivation of our proposed multigrid formulation.

3.1 Memory ablation

We performed an ablation study on the number of memory entries, N, used in our multiscale memory learning technique. In the main submission, all the experiments, including the ablation on the *Stacked vs. Multigrid* formulation, were conducted using $N = 20$. In Table 1, we compare both $N = 2$ that denotes the novel class and background *vs.* using an overcomplete set with more than two, $N = 20$. It is seen that over all folds, and especially the first two folds, that using an overcomplete set improves the separation between the background and novel class leading to better segmentation accuracy. The overcomplete set captures different parts of the novel class and background and thereby improves the segmentation, as described in the following section.

3.2 Memory attention maps

In Fig. 1, we visualise the attention maps from our multiscale memory learning on the finest scale after the full bidirectional multigrid formulation is finalized on three randomly sampled frames from YouTube-VIS video. These visualizations allow us to gain better insights on how our multiscale memory learning enhances the output feature maps. Since we use $N = 20$ we show the attention maps of the different memory entries along with the original image and the predicted segmentation (highlighted in red). It is seen that our multiscale memory learning attends to different parts of the novel class and the background. It also shows the temporal consistency of the attended parts, for example memory entry 16 tends to attend to the novel class across all frames, while memory entry three attends to parts of the background. Thus, it generally confirms the benefit from our multiscale memory learning in enhancing the query feature maps and in separating the novel class from the background for better segmentation accuracy.

Additionally, Fig. 2 shows the output attention maps from the mid levels during the bidirectional information exchange for one randomly sampled frame. Notably, the bidirectional information exchange with separate decoder layers entails that each layer captures a different aspect of the input

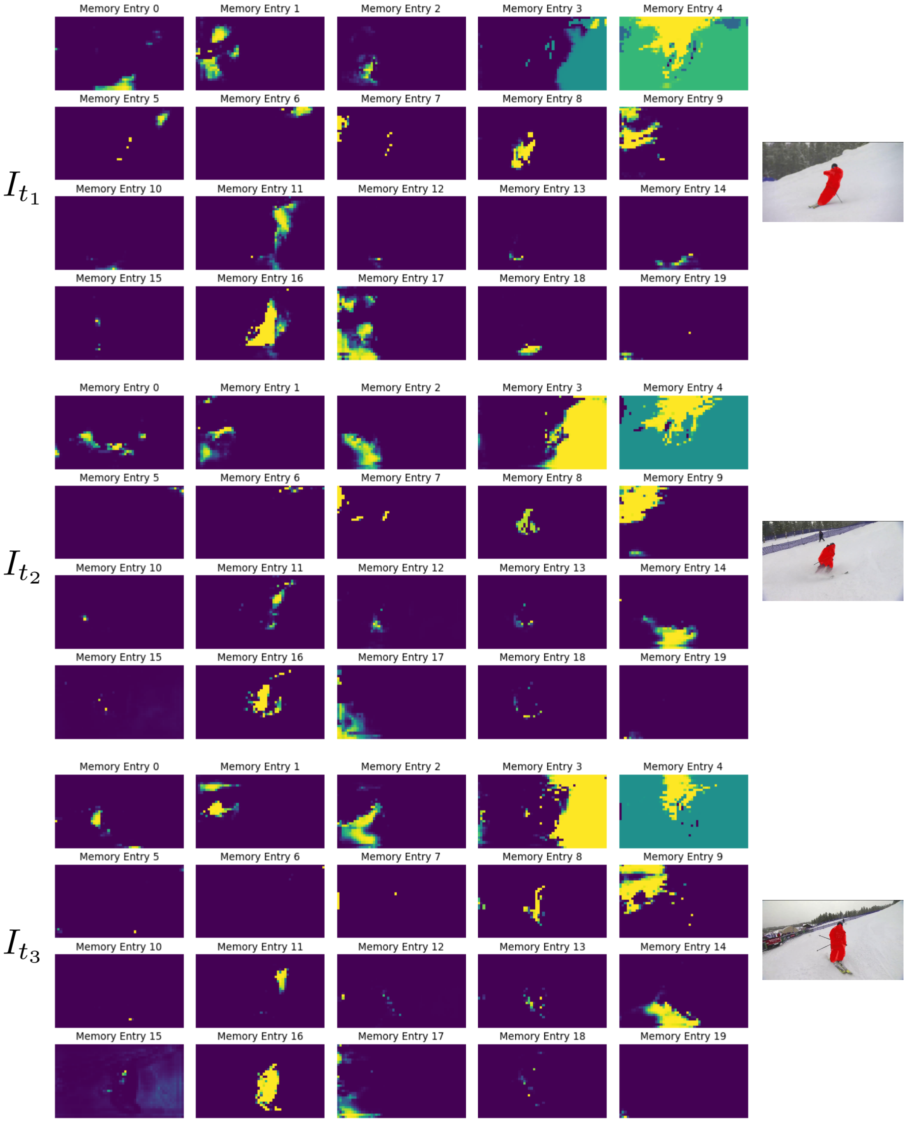

Figure 1: Visualisation of learned memory attention maps. **Mid:** Visualisation of the output attention maps in the finest scale after the full bidirectional multigrid formulation for the memory entries in the learned memory with $N = 20$ for three randomly sampled frames, $I_{t_1}, I_{t_2}, I_{t_3}$. Attention is visualised as a heatmap where yellow indicates higher attention. **Right:** Input image with the segmentation prediction highlighted in red.

even within the same scale. This fact suggests that the different decoder layers with separate learnable weights will capitalize on different components of the input imagery. Here, we focus on what is being captured in the mid level with each exchange of information in the coarse-mid-fine-mid-coarse-mid-fine scheme (*i.e.* three times communicating with the mid scale). It is seen that with each exchange of information focus is on a separate component of the input, where the early exchange, p_1, focuses on the hard negatives around the novel class that can be confused with the background (*e.g.* memory entries five to seven). The second exchange, p_2, captures better the novel classes (*e.g.* memory entry 18), and finally the last encounter, p_3, captures the highly confident parts of the novel class (*e.g.* memory entry 15).

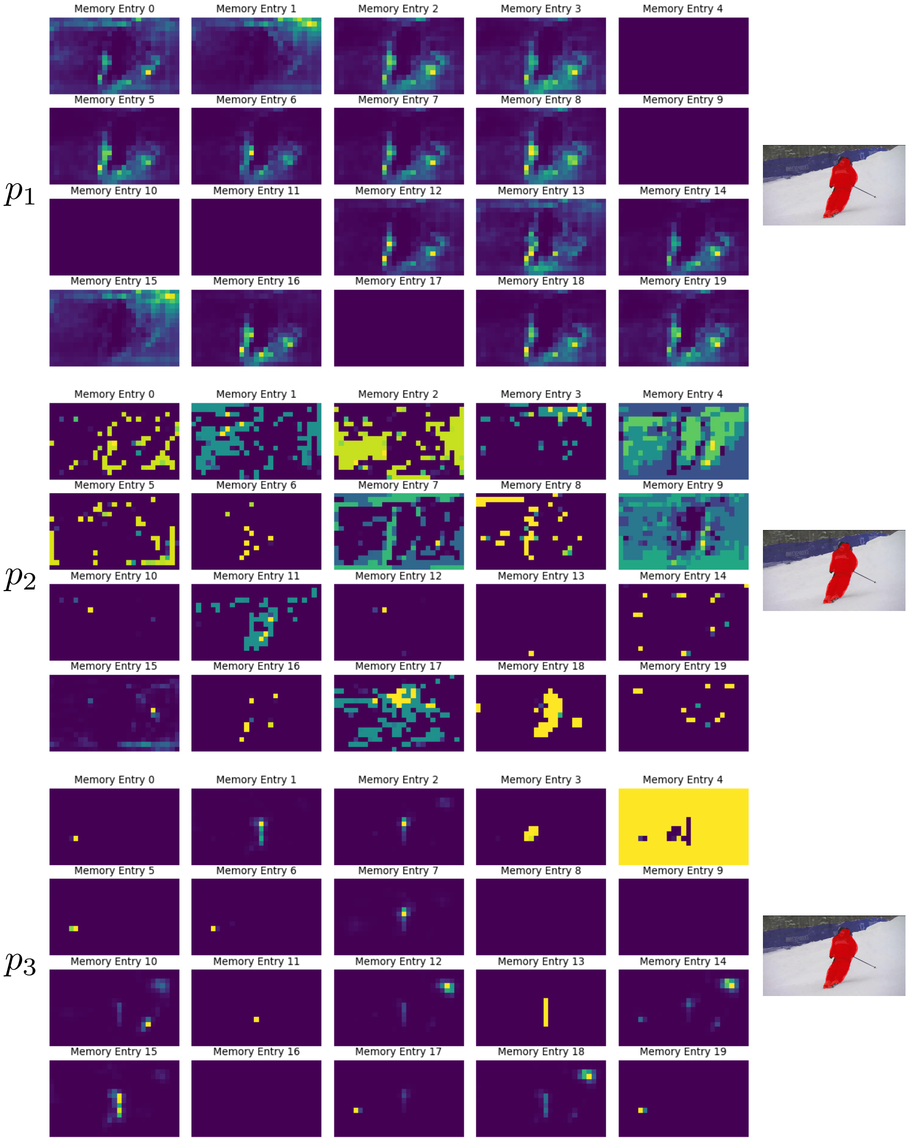

Figure 2: Visualisation of the output attention maps from our learned memory for the mid scale for each information exchange, $p_{1}, _{2}, _{3}$ in the bidirectional communication. Attention is visualised as a heatmap, where yellow indicates higher attention. **Right:** Input image with the segmentation prediction highlighted in red.

3.3 Frequency analysis

In the main submission, we hypothesize that one of the reasons behind our multigrid formulation benefit is related to the fact that different feature abstraction levels capture different frequency components, *cf*. [4]. The bidirectional multigrid formulation allows for a better information exchange across scales and hence better captures the different frequency components required for delineating the object boundaries. Inspired by previous work that analyzed transformers from a frequency domain perspective [1], we conduct a frequency analysis where we perform low pass filtering with a Gaussian kernel on the input images with different filter sizes and standard deviations, as shown in Fig. 3. We then evaluate our final predictions similar to the original setup on the input filtered images and average per filter size over the four folds on YouTube-VIS. In Fig. 3 across the early three filter sizes our approach can better cope with frequency band-passed input compared to the baseline. These

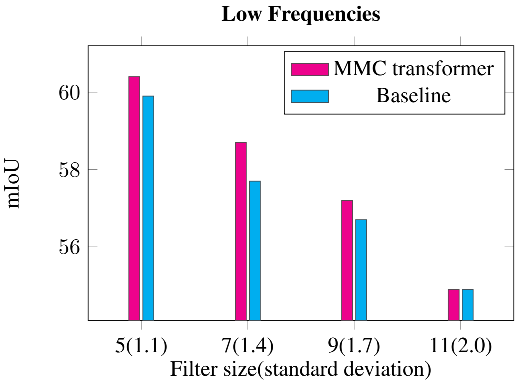

Figure 3: Evaluation of mIoU averaged over the four folds of our MMC transformer *vs*. the multiscale hypercorrelation squeeze network baseline [5] with input images that have gone through low pass filtering with different filter sizes and standard deviations.

results show that MMC transformer is more robust than the baseline to restricted frequency content, where here the restricted frequency content comes about through low pass filtering.

4 Code

We include our code as part of the supplementary material with a README file that can reproduce our experiments in the main submission, along with our trained weights to reproduce inference.

5 Assets and licenses

We use the YouTube-VIS[1] dataset which is under a Creative Commons Attribution 4.0 License that allows non commercial research use. We also used the A2D[2] dataset where the license states that the dataset may not be republished in any form without the written consent of the authors.

6 Societal impact

Few-shot video object segmentation, where the query set to be segmented is a video, is a crucial task that can help reduce the annotation cost required to label large-scale video datasets. It can serve a variety of applications in autonomous systems [2] and medical image processing [6] which require the model to learn from few labelled examples for novel classes that are beyond the closed set of training classes with abundant labels. It can also help bridge the gap between developing and developed countries, where the former lacks the resources necessary to annotate large-scale labelled datasets that are required in a variety of tasks that serves the community such as, the use of satellite imagery in agricultural monitoring and crop management [7]. We believe our work in general provides positive impact in empowering developing countries to establish labelled datasets that satisfy the needs of their own communities rather than following public benchmarks.

However, as with many artificial intelligence algorithms, video object segmentation can have negative societal impacts, *e.g*. through application to automatic target detection in military and surveillance systems. There are emerging movements to limit such applications, *e.g*. pledges on the part of researchers to band use of artificial intelligence in weaponry systems. We have participated in signing that pledge and are supporters of its enforcement through international laws. Nonetheless, we strongly believe these misuses are available in both few-shot and non few-shot methods and are not tied to the specific few-shot case. On the contrary, we argue that empowering developing countries towards decolonizing artificial intelligence can help go beyond centered power that currently lies within developed countries and big tech companies and is guided solely by their interests.

---

[1] `https://youtube-vos.org/dataset/`

[2] `https://web.eecs.umich.edu/~jjcorso/r/a2d/`

References

[1] Jiawang Bai, Li Yuan, Shu-Tao Xia, Shuicheng Yan, Zhifeng Li, and Wei Liu. Improving vision transformers by revisiting high-frequency components. *arXiv preprint arXiv:2204.00993*, 2022.

[2] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15333–15342, 2021.

[3] Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, and Shengfeng He. Delving deep into many-to-many attention for few-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14040–14049, 2021.

[4] Isma Hadji and Richard P Wildes. Why convolutional networks learn oriented bandpass filters: Theory and empirical support. *arXiv preprint arXiv:2011.14665*, 2020.

[5] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6941–6952, 2021.

[6] Tobias Ross, Annika Reinke, Peter M Full, Martin Wagner, Hannes Kenngott, Martin Apitz, Hellena Hempe, Diana Mindroc Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Robust medical instrument segmentation challenge 2019. *arXiv preprint arXiv:2003.10299*, 2020.

[7] Joel Segarra, Maria Luisa Buchaillot, Jose Luis Araus, and Shawn C Kefauver. Remote sensing for precision agriculture: Sentinel-2 improved features and applications. *Agronomy*, 10(5):641, 2020.

[8] Amirreza Shaban, Zhen Bansal, Shrayand Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *Proceedings of the British Machine Vision Conference*, pages 167.1–167.13, 2017.

[9] Pengwan Yang, Pascal Mettes, and Cees GM Snoek. Few-shot transformation of common actions into time and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16031–16040, 2021.