5th Place Solution for YouTube-VOS Challenge 2022: Video Object Segmentation

Wangwang Yang*, Jinming Su*, Yiting Duan, Tingyi Guo and Junfeng Luo

Vision Intelligence Department (VID), Meituan

{yangwangwang, sujinming, duanyiting, guotingyi, luojunfeng}@meituan.com

Abstract

Video object segmentation (VOS) has made significant progress with the rise of deep learning. However, there still exist some thorny problems, for example, similar objects are easily confused and tiny objects are difficult to be found. To solve these problems and further improve the performance of VOS, we propose a simple yet effective solution for this task. In the solution, we first analyze the distribution of the Youtube-VOS dataset and supplement the dataset by introducing public static and video segmentation datasets. Then, we improve three network architectures with different characteristics and train several networks to learn the different characteristics of objects in videos. After that, we use a simple way to integrate all results to ensure that different models complement each other. Finally, subtle post-processing is carried out to ensure accurate video object segmentation with precise boundaries. Extensive experiments on Youtube-VOS dataset show that the proposed solution achieves the state-of-the-art performance with an 86.1% overall score on the YouTube-VOS 2022 test set, which is 5th place on the video object segmentation track of the Youtube-VOS Challenge 2022.

1. Introduction

Video object segmentation (VOS) [17, 23, 14, 26], as a dense prediction task, aims at segmenting particular object instances across one video. Based on VOS, Semi-supervised video object segmentation (Semi-supervised VOS) targets segmenting particular object instances throughout the entire video sequence given only the object mask in the first frame, which is very challenging and has attracted lots of attention. Recently, Semi-supervised VOS has made good progress and been widely applied to autonomous driving, video editing and other fields. In this paper, we focus on improving the performance of the Semi-supervised VOS (referred to as VOS for convenience below).

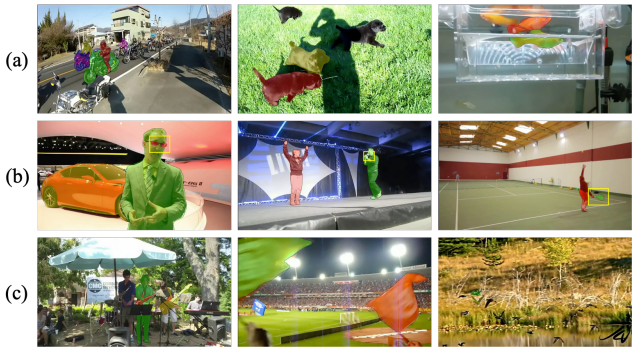

Figure 1. Challenges in video object segmentation. (a) Similar object are confused. (b) Tiny objects are difficult to detect. (c) Great differences in semantics and scenes bring great challenges.

In recent years, many VOS datasets have emerged, among which DAVIS[17] and Youtube-VOS [23] are the two most widely adopted. DAVIS 2017 is a multi-object benchmark containing 120 videos with dense annotation. Compared with DAVIS, Youtube-VOS is the latest large-scale benchmark for multi-object video object segmentation and is much bigger (about 40 times) than DAVIS 2017. In Youtube-VOS, camera jitter, background clutter, occlusion and other complicated situations are kept in the process of data collection and annotation, in order to restore the real scene and solve these complicated situations by means of algorithms. To address challenges in VOS, lots of learning-based methods have been proposed in recent years, achieving impressive performance. However, there still exist several challenges that hinder the development of VOS. First of all, there are many similar objects in the real application scenarios of VOS (as shown in Figure 1 (a)), where the accurate cross-frame tracking of these objects is very con-

* Equal contribution.

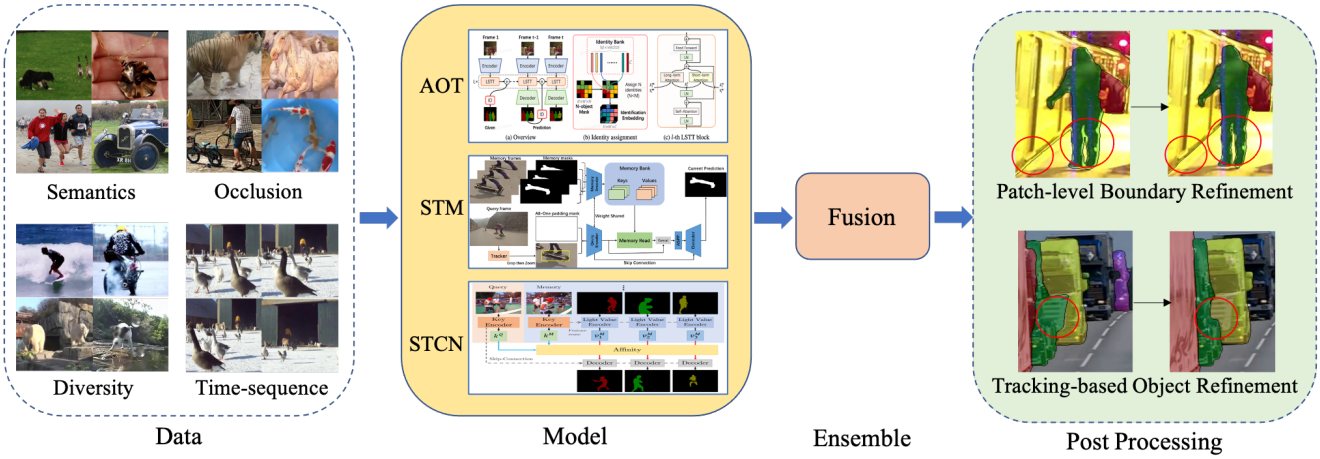

Figure 2. Overview of the proposed solution.

fusing, which leads to the different objects being wrongly matched as the same one. Secondly, tiny objects are difficult to detect, especially in the process of moving across frames and the size of objects fluctuates greatly (as depicted in Figure 1 (b)), which makes it difficult for the algorithm to accurately detect and track these objects. In addition, many scenarios are very different, containing different objects and behaviors (as displayed in Figure 1 (c)), which lead to many scenes and semantics not being included in the training dataset, thus bringing great challenges to the generalization of the algorithm. Actually, the above points are prominent problems, and there are many other difficulties to be solved in the task of VOS, which together make VOS still a challenging task.

To deal with unresolved difficulties in VOS, many methods made great efforts. Space-Time Memory network(STM)[14] introduces memory networks to learn relevant information from all available sources. In STM, the past frames with object masks form an external memory, and the current frame as the query is segmented using the mask information in the memory. Specifically, the query and the memory are densely matched in the feature space, covering all the space-time pixel locations in a feed-forward fashion. By this way, it is verified that STM is able to handle the challenges such as appearance changes and occlusions effectively. In addition, Space-Time Correspondence Network (STCN) [3] uses direct image-to-image correspondence for efficiency and more robust similarity measures in matching process, which greatly improves the efficiency and performance of STM. Recently, An associating objects with transformers algorithm (AOT) [25] is proposed to deal with the challenging multi-object scenarios. In detail, AOT employs an identification mechanism to associate multiple targets into the same high-dimensional embedding space, thus simultaneously processing multiple objects' matching and segmentation decoding as efficiently as processing a single object. Within these methods, they can track and segment most specified objects across one video, but similar objects, tiny objects and objects in complex scenes are still difficult to track and segment.

Inspired by these existing methods, we propose a simple yet effective solution for VOS, as shown in Figure 2. In order to deal with existing difficulties, we first analyze the Youtube-VOS dataset and other video segmentation related datasets (*e.g.*, OVIS[18] and VSPW[12]). We find that other datasets can supplement the diversified scenes with similar objects and occlusion situation from the data aspect. And in the model aspect, we choose three different basic networks (*i.e.*, AOT[28], STCN[3] and FAMNet[24](An improved STM[14])), which have different structures and can learn the information of objects from different aspects, so as to achieve complementary promotion. Next, a simple fusion method is used to integrate different predictions from several variants based on the above three basics. Finally, a series of exquisite post-processing strategies are introduced to improve the prediction results and ensure accurate video object segmentation with precise boundaries. Extensive experiments on the Youtube-VOS dataset show that the proposed solution achieves state-of-the-art performance.

The main contributions of this paper include: 1) We analyze the characteristics of the Youtube-VOS dataset, and supplement the dataset with static and video segmentation datasets. 2) We improve three basic network architectures and train several variants to learn the different aspects of information for objects in videos. 3) We introduce a series of subtle fusion and post-processing strategies to ensure accurate video object segmentation with precise boundaries. 4) The proposed solution achieves the state-of-the-art performance with an 86.1% overall score on the YouTube-VOS 2022 test set, which is the 5th on the video object segmen-

tation track of the Youtube-VOS Challenge 2022.

2. Method

To address difficulties in VOS, we propose a simple yet effective solution, as shown in Figure 2. Details of the proposed solution are described as follows.

2.1. Data Matters

Learning-based VOS methods are highly dependent on data. Although YouTube-VOS is the largest video object segmentation dataset, it is still unable to utilize the potential of the current state-of-the-art methods sufficiently.

Through the analysis of the Youtube-VOS dataset, we find that there are four key points to pay attention to at the data level, which are summarized as semantics, occlusions, diversity and time-sequence. Most state-of-the-art methods in the VOS task adopt a two-stage training strategy. In the first stage, video clips synthesized from the static images are used for pre-training. Then the real video data is used for final training in the second stage. Large-scale static image datasets come from fields like instance segmentation and salient object detection, and therefore have more semantics and diversity. By pre-training on them, the VOS model can extract robust feature embedding for pixel-level spatiotemporal feature matching, and improve the ability to identify and discriminate against diverse targets. We try to introduce ImageNet-Pixel[29], a more diverse image dataset in the pre-training stage, but it does not bring obvious benefits. We believe that this is because the current model structure and separated two-stage training method cannot fully utilize the information in the static image datasets. On the other hand, video data in the real world have additional temporal information compared to static image data, and the data form in the testing stage is video, so it is more straightforward to bring more video segmentation datasets to improve performance. Benefit from the release of several new datasets in the video segmentation field recently, such as YoutubeVIS which has more objects in each video, OVIS [18] which occlusion scenarios are significant, and VSPW [12]) which have dense annotations and high-quality resolution, we introduce them into the second training stage, thus significantly improving the performance of models.

After the above data supplement, the ability of the model in the aspect of semantics extraction, occlusions recognition and other aspects has been enhanced.

2.2. Strong Baseline Models

We adopt three kinds of architectures as our baseline models including AOT[28], FAMNet[24], and STCN[3], which have high performance in the VOS field recently. The detailed implementation can be found in their original papers.

Benefit from an effective identification mechanism and long short-term transformer module, AOT can achieve high performance for both seen and unseen objects in the test phase. Meanwhile, FAMNet and STCN can produce better results for unseen objects because of the simplicity and robustness of their core pixel-wise feature matching module. By combining methods with different network designs, we can get several sets of results at different aspects of information for video objects and obtain more gains in the model ensemble stage.

2.3. Nontrivial Attempts

LSTT block V2: Although the network structure of the AOT[28] is delicate, it still has room for improvement. Following AOST[25], we improve Long Short-Term Transformer (LSTT) block, the core module in AOT model, and obtain the LSTT block V2 which has better performance. Specifically, LSTT block utilize the attention mechanism to perform pixel-level feature matching between the current frame and memory frames. The formulas of common attention-based matching mechanism, attention-based matching mechanism of the LSTT block and attention-based matching mechanism of the LSTT block V2 are,

$$\begin{aligned} Att(Q,K,V) &= Corr(Q,K)V \\ &= softmax(\frac{QK^{tr}}{\sqrt{C}})V, \end{aligned} \tag{1}$$

$$Att(Q,K,V+E), \tag{2}$$

$$Att(Q,K\odot\sigma(W_l^G E),V+W_l^{ID}E). \tag{3}$$

By combining the target identification embedding E with the value embedding of the memory frames V in Eq. 1, AOT can propagate the information of multiple targets to the current frame simultaneously (Eq. 2). Compared with the original LSTT block, the LSTT block V2 is more effective. There are two main differences between them. The first one is in the value part of the attention mechanism. LSTT block V2 projects E by a linear layer W_l whose weights are various in different LSTT layers. Such modification increases the degree of freedom of the model. The second one is in the key part of the attention mechanism. LSTT block V2 generates a single channel map to adjust the key embedding of the memory frames K using E so that the target information of memory frames can be used in the matching process between Q and K too.

Turning off strong augmentation: In order to further improve the performance of our models, a trick frequently used in the object detection field is performed. Specifically, we turn off data augmentation operations other than random cropping for the last few epochs when running the second stage of training. Meanwhile, we only use YouTube-VOS as our training data. In this way, the data distribution in the

final training stage is more consistent with the data distribution in the testing stage.

Attaching top-k filtering to memory read: The long-term memory bank in AOT is similar to the memory bank in STM-like methods. So the number of memory frames used in the training stage and testing stage is inconsistent. Besides, as the video's length increases in the testing stage, the size of long-term memory bank also grows dynamically. So we attempt to add top-k filtering operation[2] to the Long-Term Attention module in AOT, to alleviate the problem of information redundancy and remove noises in long-term memory. But this attempt doesn't always work in all of our models.

Model ensemble: Considering that the performance of different models is diverse, we adopt offline ensembling to fuse these models' predictions for getting higher precision frame by frame. Specifically, we have tried two fusion methods. First, we average predictions of all models directly, to help the models complement each other and reduce error prediction. The second interesting idea is keypoint voting, we use feature matching[5][15] to correlate the target in the previous frames and current frame, so as to judge the quality of the prediction of different models in the current frame and weight them by keypoint voting, which reduces some wrong predictions.

Patch-level boundary refinement and tracking-based small object refinement: In addition to the above attempts, we use some post-processing strategies to refine the predicted results. We adopt boundary patch refinement (BPR)[21] to improve the boundary quality of object segmentation. BPR is a conceptually simple yet effective post-processing refinement framework. After that our predictions have significant improvements near object boundaries, as shown in Figure 2. Besides, the input of most existing state-of-the-art methods is the whole video frame, and the resolution of the feature map in the pixel-level feature matching process is further reduced because of the down-sampling operation. Both of them cause poor results for small objects. Therefore, we adopt the crop-then-zoom strategy in FAMNet. Firstly, we integrated box-level object position information provided by the tracker[1] and the preliminary segmentation result of the object provided by the VOS model to get the approximate positions of partial small objects within the dataset in every frame. Then the original frames are cropped and resized to a larger size. Finally, a secondary segmentation is performed on the clip to obtain more accurate segmentation results for small objects.

3. Experiments

3.1. Training Details

To comprehensively improve the accuracy, four different frameworks, including AOT[28], improved AOT[25], STCN[3], and FAMNet[24] are used in our experiments. For AOT, multiple networks such as Swin[10], EfficientNet[20], and ResNext[22] are used as the encoder to obtain better accuracy. Noted that the parameters of BN layers and the first two blocks in the encoder are frozen in view of stabilizing the training. Following the official settings of AOT, we also take a two-stage training strategy. In the pre-training stage, several static image datasets including COCO[9], ECSSD[19], MSRA10K[4], PASCAL-S[7], PASCAL-VOC[6] are used for preliminarily semantic learning. During the main training, video datasets including Youtube-VOS[23], DAVIS 2017[17], YouTubeVIS[8], OVIS[18], and VSPW[12] are used to enhance the generalization and robustness of the model.

During the training of AOT, images are cropped into patches with the fixed size of 465×465, and multiple image and video augmentations are randomly applied following [13, 27] to enrich the diversity of data. The experiments are performed on Pytorch by using 4 Tesla A100 GPUs. We minimize the sum of bootstrapped cross-entropy loss and soft Jaccard loss by adopting AdamW[11] optimizer. In the pre-training, the initial learning rate is set to 4e-4 with a weight decay of 0.03 for 100,000 steps, and the batch size is set to 32 for acceleration. In the main training, with the initial learning rate of 2e-4, weight decay of 0.07, and batch size of 16, the training step is extended to 130,000 due to the expansion of data. Noted that all learning rates will decay to 2e-5 by using a polynomial manner as [27]. To ensure the stability of training and enhance the robustness of the model, we also adopt the Exponential Moving Average (EMA)[16] to average the parameters of the model for better performance. For the training of STCN and FAMNet, all training process follows their official implementations.

3.2. Inference and Evaluation

For evaluating the single model of AOT, to reduce the sensitivity of the model to various object scales, online flip and multi-scale testing is applied to obtain better accuracy. Specifically, the predictions generated from videos with three scales of $1.2 \times$ 480p resolution, $1.3 \times$ 480p resolution, and $1.4 \times$ 480p resolution are ensembled frame by frame during the inference. In addition, to balance the number of objects that appear in the quite long or quite short video sequence, a dynamic memory frame sampling strategy is introduced in our experiments. For STCN and FAMNet, we also apply the flip and multi-scale testing in our experiments.

Considering that models with different structures have unique predictive advantages, we adopt an offline model ensemble strategy to further improve the performance of results. Specifically, soft prediction scores produced by models which have different frameworks and different backbone networks are simply averaged as the final result. Noted that

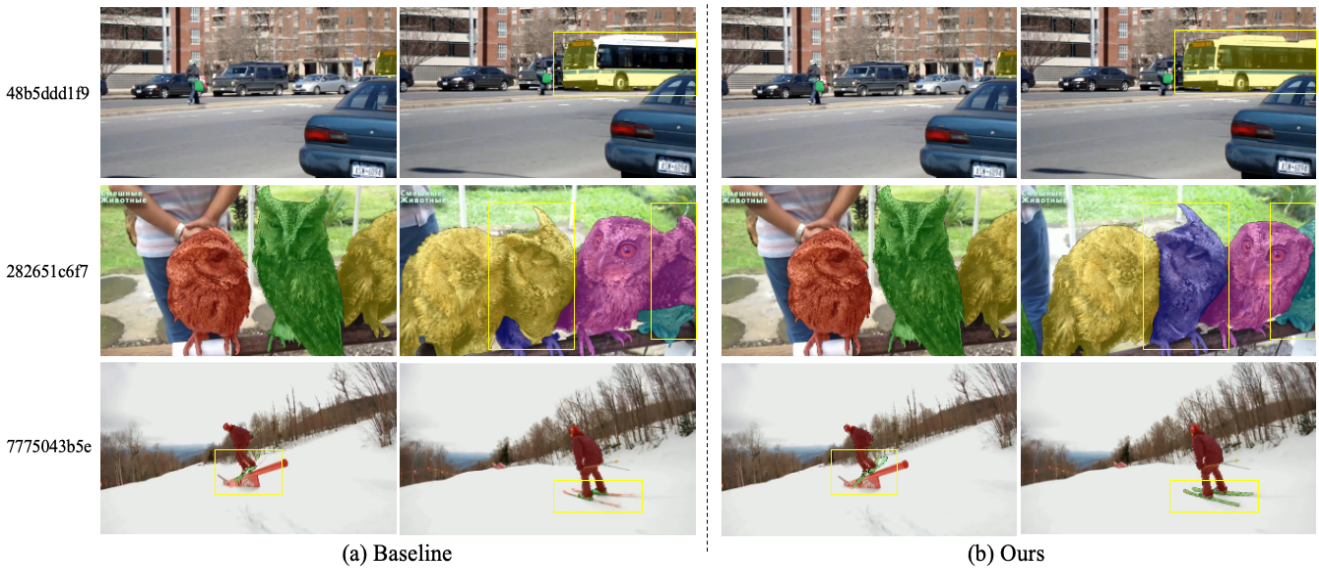

Figure 3. Representative visual examples from the baseline model and the proposed solution.

Team Name	Overall	$\mathcal{J}_{seen}$	$\mathcal{J}_{unseen}$	$\mathcal{F}_{seen}$	$\mathcal{F}_{unseen}$
Thursday_Group	0.872(1)	0.855(1)	0.817(3)	0.914(1)	0.903(1)
ux	0.867(2)	0.844(3)	0.819(1)	0.903(2)	0.903(2)
zjmagicworld	0.862(3)	0.841(4)	0.816(4)	0.895(4)	0.896(4)
whc	0.862(4)	0.840(5)	0.818(2)	0.894(5)	0.896(5)
gogo	0.861(5)	0.847(2)	0.808(7)	0.901(3)	0.890(6)
sz	0.857(6)	0.831(6)	0.815(5)	0.886(7)	0.896(3)
PinxueGuo	0.856(7)	0.832(7)	0.812(6)	0.887(6)	0.892(7)

Table 1. Ranking results in the YouTube-VOS 2022 test set. We mark our results in blue.

we also have tried other strategies like max weighting and key-point voting, and the average operation gains the best performance. All the results are evaluated on the official YouTube-VOS evaluation servers.

3.3. Results

Through test-time augmentation, model ensemble and post-processing strategies, the proposed solution obtain the 5th place on the YouTube-VOS 2022 testset, as listed in Table 1. From the result, we see that our solution surpasses most solutions in the seen category (as shown in $\mathcal{J}_{seen}$ and $\mathcal{F}_{seen}$), which is a characteristic of our solution. In addition, we also show some of our quantitative results in Figure 3. It can be seen that the proposed solution can accurately segment objects in some difficult scenarios which have severe changes in object appearance, confusion of multiple similar objects and small objects.

In order to demonstrate the effectiveness of different components, we conduct several ablation experiments. Quantitative results are shown in Table 2. We boost the performance of the original AOT network to 86.6% on YouTube-VOS 2019 validation set without any test-time augmentation such as multi-scale testing or flip testing.

Components	Overall
AOTL-R50(baseline)	85.3
+ Swinb backbone	85.5
+ LSTT Block v2	85.7
+ More real video data	86.2
+ Turn off strong augmentation	86.6

Table 2. Ablation study on YouTube-VOS 2019 validation set.

4. Conclusion

In this paper, we propose a solution for the video object segmentation task, and make nontrivial improvements and attempts in many stages such as data, model, ensemble, and post-processing strategies. In the end, we achieve the 5th place on the YouTube-VOS 2022 Video Object Segmentation Challenge with an overall score of 86.1%.

References

[1] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu. Transformer tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

[2] H. K. Cheng, Y.-W. Tai, and C.-K. Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] H. K. Cheng, Y.-W. Tai, and C.-K. Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 37(3):569–582, 2014.

[5] T. M. Daniel DeTone and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018.

[6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision (IJCV)*, 88(2):303–338, 2010.

[7] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2011.

[8] Y. F. L. Yang and N. Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[11] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[12] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[14] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[15] T. M. Paul-Edouard Sarlin, Daniel DeTone and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, page 4938–4947, 2020.

[16] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[17] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[18] J. Qi, Y. Gao, Y. Hu, X. Wang, X. Liu, X. Bai, S. Belongie, A. Yuille, P. Torr, and S. Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision (IJCV)*, 2022.

[19] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 38(4):717–729, 2015.

[20] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

[21] C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, and X. Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021.

[22] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.

[23] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[24] W. Yang, Z. Lv, J. Liu, and D. Huang. Feature aligned memory network for video object segmentation. 2021.

[25] Z. Yang, J. Miao, X. Wang, Y. Wei, and Y. Yang. Associating objects with scalable transformers for video object segmentation. *arXiv preprint arXiv:2203.11442*, 2022.

[26] Z. Yang, Y. Wei, and Y. Yang. Collaborative video object segmentation by foreground-background integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[27] Z. Yang, Y. Wei, and Y. Yang. Collaborative video object segmentation by foreground-background integration. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[28] Z. Yang, Y. Wei, and Y. Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[29] S. Zhang, J. H. Liew, Y. Wei, S. Wei, and Y. Zhao. Interactive object segmentation with inside-outside guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12234–12244, 2020.