# 1st Place Solution for YouTubeVOS Challenge 2022: Video Object Segmentation

Rui Sun[1*], Naisong Luo[1*], Yuan Wang[1*], Yuwen Pan[1], Huayu Mai[1], Zhe Zhang[2], Tianzhu Zhang[1†]
[1]University of Science and Technology of China
[2]Lunar Exploration and Space Engineering Center of CNSA
{issunrui, lns6, wy2016, panyw, mai556}@mail.ustc.edu.cn, cnclepzz@126.com, tzzhang@ustc.edu.cn

## Abstract

*Memory-based methods in semi-supervised video object segmentation task achieve competitive performance by performing dense matching between query and memory frames. However, most of the existing methods struggle to distinguish similar objects caused by global-to-global matching. Besides, they ignore the context pixel information which can be utilized to capture the local differences between the predicted mask and the ground truth. To mitigate these limitations, we propose a dynamic matching network (DMNet) [21] to jointly model pixel-level matching and part-level matching for semi-supervised VOS. The proposed DMNet model enjoys several merits. First, we propose a dynamic pixel-aware correspondence module (Pixel-CM) and a dynamic part-aware alignment module (Part-AM), and these two modules are trained via an adversarial process, where Pixel-CM will generate the more accurate predicted mask approaching the ground truth to fool Part-AM. Second, the proposed Pixel-CM is responsible for further dynamically optimizing the correspondences within the local window to reduce false matches, and Part-AM aims at dynamically dividing different target objects into diverse parts in an adaptive manner and accurately discriminating detailed local differences between the predicted mask and the ground truth. After applying test-time augmentations and model ensemble, we rank 1st in Track 1 (Video Object Segmentation) of the 4th Large-scale Video Object Segmentation Challenge (CVPR2022), achieve the $\mathcal{G}$ score of 87.2% on test set.*

## 1. Introduction

Semi-supervised Video Object Segmentation (VOS) is a fundamental task to perform pixel-wise classification of a set of class-agnostic objects in video sequences, which has been widely applied to autonomous driving [31], video editing [14], augmented reality [13], *etc*. Since the object mask is only given in the first frame without any other prior information assumptions, how to fully exploit limited information to perform accurate segmentation in the subsequent frames is thus extremely challenging.

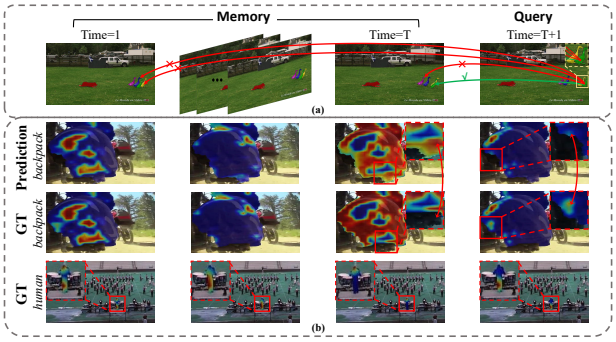


Figure 1. Illustration of our motivation. (a) shows the pixel-level mismatching caused by the global-to-global correspondences. (b) shows the effect of the dynamic part-aware alignment module. The first and the second rows show that the differences between the predicted mask and the ground truth are usually in local parts. The second and the third rows show that various target objects are divided into diverse parts in an adaptive manner.

By studying the existing memory-based methods, we sum up two limitations that need to be mitigated for building a robust VOS model. (1) **Pixel-level mismatching.** Most memory-based methods only consider global-to-global matching at the pixel level, so they tend to struggle to distinguish the objects with similar appearances, increasing the risk of false matches, as shown in Figure 1 (a). To alleviate this problem, KMN [18] attempts to conduct additional memory-to-query matching, but neglects the inherent temporal information of the video sequences. Actually, the target objects appear only in local regions in each frame. Therefore, it is more reasonable to restrict the possible pixel correspondences to a local window to reduce matching redundancy. Recently, several methods [7, 19, 24] consid-

*Equal contribution
†Corresponding author

er taking advantage of the temporal information to diminish pixel-level mismatching. However, these methods will inevitably bring correspondence noise when similar distractors are very close to the target object. Therefore, the pixel matching restricted to the local window should be further optimized to guarantee that the true correspondences enjoy higher weights. (2) **Part-level matching.** Intuitively, humans can quickly identify a specific target object from the cluttered background by automatically decomposing the object into multiple local parts, and then discriminate them in a fine-grained manner. Inspired by this, we believe that the VOS model should not only be constrained at the pixel-level, but should also be aligned with the ground truth in a part-level manner, which is not considered by the previous methods. In specific, the differences between the predicted mask and the ground truth are usually in local parts, especially the drastic changes in the appearance of the target object across frames caused by object movements, camera movements and occlusions. Please see the first row and the second row in Figure 1 (b). Thus, it is necessary to make full use of the context information to merge the neighboring pixel features to conduct part-level matching. However, since the class-agnostic objects in various video sequences have large distribution differences in size and shape. For example, as shown in the second row and the third row in Figure 1 (b), the object *human* accounts for only a small part in the current video, while the object *backpack* occupies a dominant position in another video. Therefore, it is impractical to decompose objects into different parts in a fixed manner such as grid dividing. And how to dynamically divide different target objects into different parts in an adaptive manner is extremely challenging.

To mitigate the above limitations, we propose a dynamic matching network (DMNet) [21] based on adversarial learning framework including a dynamic pixel-aware correspondence module (Pixel-CM) and a dynamic part-aware alignment module (Part-AM) for robust VOS. **To alleviate pixel-level mismatching**, we employ a dynamic pixel-aware correspondence module that combines the kernel guidance constraint and the optimal transport algorithm [5, 22] to further dynamically optimize the correspondences within the local window. Specifically, we leverage kernel priori [19] to impose temporal smoothness constraints on the global-to-global correspondences calculated by the cross-attention mechanism between the query and the memory. Therefore, the marginal distribution of the kernel prior can be served as the initial marginal distribution of the optimal transport algorithm to optimize the correspondences within the local window. Then, after obtaining the correspondences with temporal smoothness and the initial marginal distribution, we can attain the optimal transport plan dynamically, which can be regarded as the refined local-to-local matching. In this case, the relatively minor pixels will be suppressed while the dominant ones are highlighted to reduce the correspondence noise. **To model part-level matching**, we propose a dynamic part-aware alignment module, which can dynamically divide different target objects into diverse parts in an adaptive manner, thus the detailed local differences between the predicted mask and the ground truth can be accurately discriminated. In specific, we introduce a set of part-aware prototypes and take advantage of the cross-attention mechanism to extract part-aware features from the object feature map. In this way, we compare the similarity among these part-aware features, and select the most different part pair to optimize the model. For training, we optimize the model under the framework of adversarial learning to make it more robust. In this way, Part-AM (*Discriminator*) can accurately discriminate detailed local differences, and Pixel-CM (*Generator*) will generate the more accurate predicted mask approaching the ground truth to fool Part-AM via an adversarial process.

After applying common test-time augmentations including multi-scale and flipping, and ensembling DMNet [21] with AOT [28] and STCN [10], we rank 1st place in the Track 1 (Video Object Segmentation) of the 4th Large-scale Video Object Segmentation Challenge (CVPR2022), achieve the $\mathcal{G}$ score of 87.2% on test set.

## 2. Our Method

Following the STCN [4], given $t$ memory frames (only images as input) and a query frame at time $t+1$, the memory feature map $\mathbf{Z}^t \in \mathbb{R}^{t\times h\times w\times c}$ and the query feature map $\mathbf{X}^{t+1} \in \mathbb{R}^{h\times w\times c}$ are respectively extracted from a feature extractor $\varphi$, where $h$, $w$ and $c$ denote the height, width and channel number of the feature map, respectively. Besides, $t$ memory frames (images and masks as input) are fed into the feature extractor $\psi$ to attain the mask embeddings $\mathbf{M}^t \in \mathbb{R}^{t\times h\times w\times d}$. Note that we take `res4` features with stride 16 from both feature extractors $\varphi$ and $\psi$, for simplicity, we omit the superscript $t$. To diminish pixel-level mismatching and conduct part-level matching simultaneously, as shown in Figure 2, the proposed DMNet consists of a dynamic pixel-aware correspondence module (Pixel-CM) and a dynamic part-aware alignment module (Part-AM), which are trained via an adversarial process.

### 2.1. Dynamic Pixel-Aware Correspondence Module

We leverage kernel guidance [19] to impose temporal smoothness constraints on the global-to-global correspondences calculated by the cross-attention mechanism to obtain $\hat{\mathbf{S}}$. Our goal is to compute the minimum cost transmission plan $\tilde{\mathbf{S}}$ with each entry $\tilde{s}_{i,j}$ representing the further dynamic optimization between the $i$-th query pixel and the $j$-th memory pixel. Specifically, $\tilde{\mathbf{S}}$ is computed by solving the optimal transport problem as

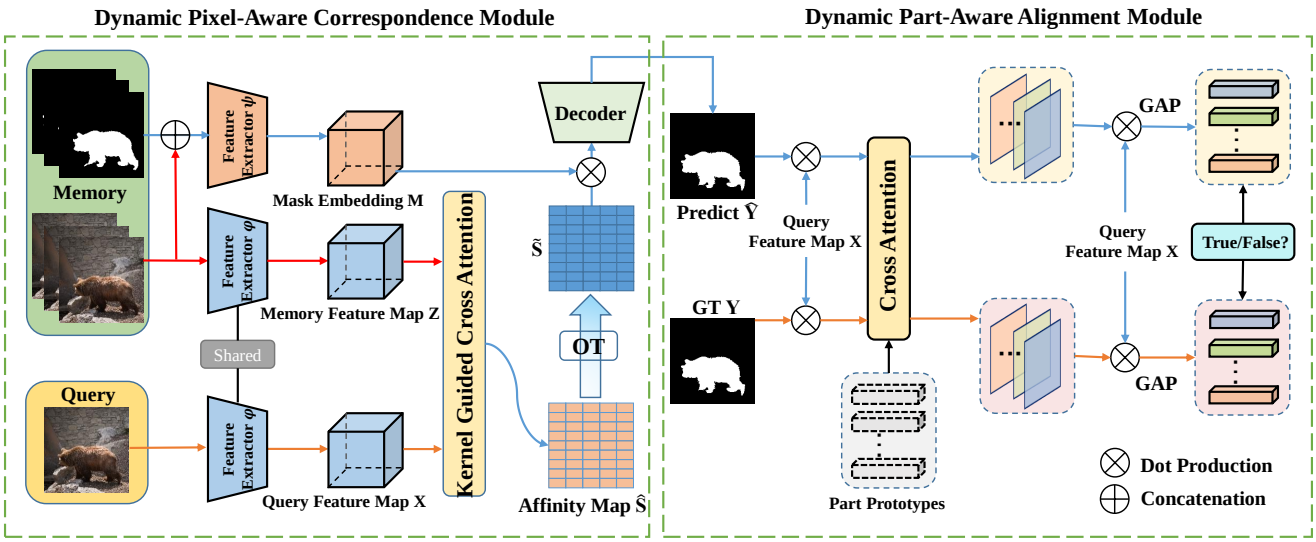


Figure 2. Illustration of the proposed DMNet. DMNet is mainly composed of a dynamic pixel-aware correspondence module to further dynamically optimize the correspondences within the local window, and a dynamic part-aware alignment module to discriminate detailed local differences between the predicted mask and the ground truth. In the figure, "GAP" represents a global average pooling layer [11].

$$\max_{\tilde{\mathbf{S}}\in\tilde{\mathcal{S}}} \mathrm{Tr}\left(\tilde{\mathbf{S}}^{\top}\hat{\mathbf{S}}\right) + \epsilon H(\tilde{\mathbf{S}}), \tag{1}$$

$$\tilde{\mathbf{S}} = \left\{\tilde{\mathbf{S}} \in \mathbb{R}_{+}^{hw\times thw} : \tilde{\mathbf{S}}\mathbf{1}_{thw} = \boldsymbol{\mu}, \tilde{\mathbf{S}}^{\top}\mathbf{1}_{hw} = \boldsymbol{\nu}\right\}, \tag{2}$$

where $\mathbf{1}_{thw} \in \mathbb{R}^{thw}$ and $\mathbf{1}_{hw} \in \mathbb{R}^{hw}$ denote vectors of ones. In Eq. (1), the second term (i.e., $H(\tilde{\mathbf{S}}) = -\sum_{i=1}^{hw}\sum_{j=1}^{thw}\tilde{s}_{i,j}\log\tilde{s}_{i,j}$) measures the entropy regularization of $\tilde{\mathbf{S}}$, and $\epsilon$ is the weight for the entropy term. A large value of $\epsilon$ usually leads to a trivial solution where each query pixel has the same correspondence to each memory pixel. Thus, we use a small value of $\epsilon$ in our experiments to avoid the above trivial solution. Besides, in Eq. (2), both $\boldsymbol{\mu} \in \mathbb{R}_{+}^{hw}$ and $\boldsymbol{\nu} \in \mathbb{R}_{+}^{thw}$ represent the initial marginal distribution.

## 2.2. Dynamic Part-Aware Alignment Module

Since the differences between the predicted mask $\hat{\mathbf{Y}}$ and the corresponding ground truth $\mathbf{Y}$ are usually in local parts, it is necessary to make full use of the context information to merge the neighboring pixel features to conduct part-level matching. Then we design a dynamic part-aware alignment module, which can dynamically divide different target objects into diverse parts in an adaptive manner.

In specific, we introduce a set of part prototypes $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^{K}$ focusing on different target object parts dynamically, each of which represents a part filter to discover pixels of the query feature map belonging to the part $i$. Given the down-sampled predicted mask $\hat{\mathbf{Y}}$ and the corresponding ground truth $\mathbf{Y}$, the object feature map $\mathbf{F}$ and $\mathbf{F}^r$ can be obtained by multiplying with the query feature map $\mathbf{X}$, respectively. Then we adopt a cross-attention layer to generate part masks for the object feature map $\mathbf{F}$ to obtain part features $\Pi = \{\boldsymbol{\pi}_i\}_{i=1}^{K}$ by a weighted pooling over all values. Similarly, we feed the object feature map $\mathbf{F}^r$ to $D$, and acquire the corresponding part masks $\mathbf{A}^r = \{\boldsymbol{a}_i^r\}_{i=1}^{K}$ and part features $\mathbf{\Pi}^r = \{\boldsymbol{\pi}_i^r\}_{i=1}^{K}$.

Finally, we calculate the cosine similarity among part features $\mathbf{\Pi}^r = \{\boldsymbol{\pi}_i^r\}_{i=1}^{K}$ and $\mathbf{\Pi} = \{\boldsymbol{\pi}_i\}_{i=1}^{K}$, and select the most different part pair to feed into a fully-connected layer to output real/fake results for adversarial training.

The optimization process includes two steps: (1) in the generation step, fix Part-AM ($D$) and maximize the generator loss to update Pixel-CM ($G$). (2) In the discrimination step, fix Pixel-CM ($G$) and minimize the discriminator loss to update Part-AM ($D$). Alternatively optimizing the min and max steps allows Part-AM divide objects into diverse parts and discriminate between the predicted mask and the ground truth based on detailed local differences from part pairs, then Pixel-CM will adjust itself to generate more accurate segmentation for fooling Part-AM.

# 3. Implementation Details

For a fair comparison with previous methods [4, 15], we first pretrain the model on static image datasets [2, 9, 20, 23, 30] with synthetic deformation, then perform main training on DAVIS 2017 [16] and YouTube-VOS 2019 [26]. We use a batch size of 16 during pretraining and a batch size of 8 during main training, and the bootstrapped cross entropy is adopted by following [4]. The Adam optimizer [8] is employed with default momentum $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the initial learning rate is set as 1e-5. During the inference, we employ the soft aggregation operation following [15] when multiple target objects exist in a video.

We evaluate our DMNet [21] on YouTube-VOS [25], which is the latest large-scale benchmark for multi-object video segmentation. Specifically, For the 2018 version, its validation set contains 474 videos, including 65 training (seen) categories and 26 unseen categories.

We measure the area similarity ($\mathcal{J}_\mathcal{S}$, $\mathcal{J}_\mathcal{U}$) and contour accuracy ($\mathcal{F}_\mathcal{S}$, $\mathcal{F}_\mathcal{U}$) for the seen object categories and the unseen ones separately, and finally the averaged overall score $\mathcal{G}$ can be attained.

# 4. The 4th Large-scale Video Object Segmentation Challenge

In this section, we introduce our solution on the 4th Large-scale Video Object Segmentation Challenge. We mainly adopt three frameworks, including DMNet [21], AOT [28] and STCN [10]. In specific, AOT [28] introduces an identification embedding mechanism to embed the masks of multiple different targets into the same feature space for propagation. Besides, a Long Short-Term Transformer (LSTT) is designed for constructing hierarchical object matching and propagation. And STCN [10] establishes correspondences between current frames and memory ones for every object by calculating affinity with the negative squared Euclidean distance rather than the conventional cosine similarity for exploiting the rich memory information. With the strength of model ensembling, we finally achieve the 1st rank on the test split of the 4th Large-scale Video Object Segmentation Challenge.

## 4.1. Compare DMNet with SOTA Methods

As shown in Table 1, we can observe that our approach achieves superior performance (84.0% in $\mathcal{G}$) on YouTube-VOS compared to the previous state-of-the-art methods. Besides, to look deeper into the proposed method, we perform a series of ablation studies on both DAVIS 2017 validation set and YouTube-VOS 2018 set to analyze each component of our DMNet, including the dynamic pixel-aware correspondence module (Pixel-CM) and the dynamic part-aware alignment module (Part-AM). As shown in Table 2, each module is integral and coincides with its own design purpose.

**Qualitative Comparison**. Figure 3 shows qualitative comparison with some state-of-the-art methods including STM [15], STCN [10] and HMMN [19]. We can observe that STM fail to predict target objects when multiple similar objects have appeared (DAVIS example). Benefit from utilizing the optimal transport theory to further dynamically optimize the correspondences, our method yields more precise segmentation. And for fast moving target object (YouTube example), DMNet can obtain more accurate segmentation thanks to the dynamic part-aware alignment module which has the ability of discriminating the detailed local differences between the predicted mask and the ground truth.

Table 1. Comparisons between different methods on multi-object YouTube-VOS 2018 validation set.

| Method | $\mathcal{G}$ | $\mathcal{J}_\mathcal{S}$ | $\mathcal{F}_\mathcal{S}$ | $\mathcal{J}_\mathcal{U}$ | $\mathcal{F}_\mathcal{U}$ |
|---|---|---|---|---|---|
| STM[ICCV19] [15] | 79.4 | 79.7 | 84.2 | 72.8 | 80.9 |
| CFBI[ECCV20] [27] | 81.4 | 81.1 | 85.8 | 75.3 | 83.4 |
| KMN[ECCV20] [18] | 81.4 | 81.4 | 85.6 | 75.3 | 83.4 |
| CFBI+[TPAMI21] [27] | 82.8 | 81.8 | 86.6 | 77.1 | 85.6 |
| STCN[NIPS21] [10] | 83.0 | 81.9 | 86.5 | 77.9 | 85.7 |
| **DMNet *(ours)*** | **84.2** | **83.8** | **88.7** | **78.2** | **86.2** |

Table 2. Evaluation of the effectiveness of different components on DAVIS 2017 validation set [16] and YouTube-VOS 2018 [26] set by reporting $\mathcal{J}\&\mathcal{F}$ and $\mathcal{G}$ scores, respectively.

| Pixel-CM | Part-AM | DAVIS-17 | YouTube-18 |
|---|---|---|---|
| ✗ | ✗ | 85.4 | 83.0 |
| ✓ | ✗ | 86.0 | 83.4 |
| ✗ | ✓ | 86.5 | 83.7 |
| ✓ | ✓ | **87.1** | **84.2** |

Table 3. Comparison with other methods on the YouTube-VOS 2022 test set. Our team achieves a 1st place.

| Team | $\mathcal{G}$ | $\mathcal{J}_\mathcal{S}$ | $\mathcal{F}_\mathcal{S}$ | $\mathcal{J}_\mathcal{U}$ | $\mathcal{F}_\mathcal{U}$ |
|---|---|---|---|---|---|
| ***Ours*** | **87.2** | **85.5** | **91.4** | 81.7 | **90.3** |
| ux | 86.7 | 84.4 | 90.3 | **81.9** | 90.3 |
| zjmagicworld | 86.2 | 84.1 | 89.5 | 81.6 | 89.6 |
| whc | 86.2 | 84.0 | 89.4 | 81.8 | 89.6 |
| gogo | 86.1 | 84.7 | 90.1 | 80.8 | 89.0 |
| sz | 85.7 | 83.1 | 88.6 | 81.5 | 89.6 |

## 4.2. Model Ensemble

We use current state-of-the-art methods for model ensemble, including DMNet [21], STCN [4] and AOT [28], which are offline-learning memory-based. In specific, following the STCN [4], for DMNet and STCN, we get several model variations by replacing the default backbone ResNet50 [6] with Swin-B [12], WideResNet-50 [29] and WideResNet-50+ ASPP [1], respectively. For AOT, we use AOT-L derivatives [28] with different backbones, including Mobilenet-V2 [17], ResNet-50 [6] and Swin-B [12] to obtain various predictions.

For better performance we additionally train these models with the YouTube-VIS 2022 and the synthetic dataset BL30K [3]. We use test-time multi-scale and flipping augmentations for each of above model variations and average the output probabilities.

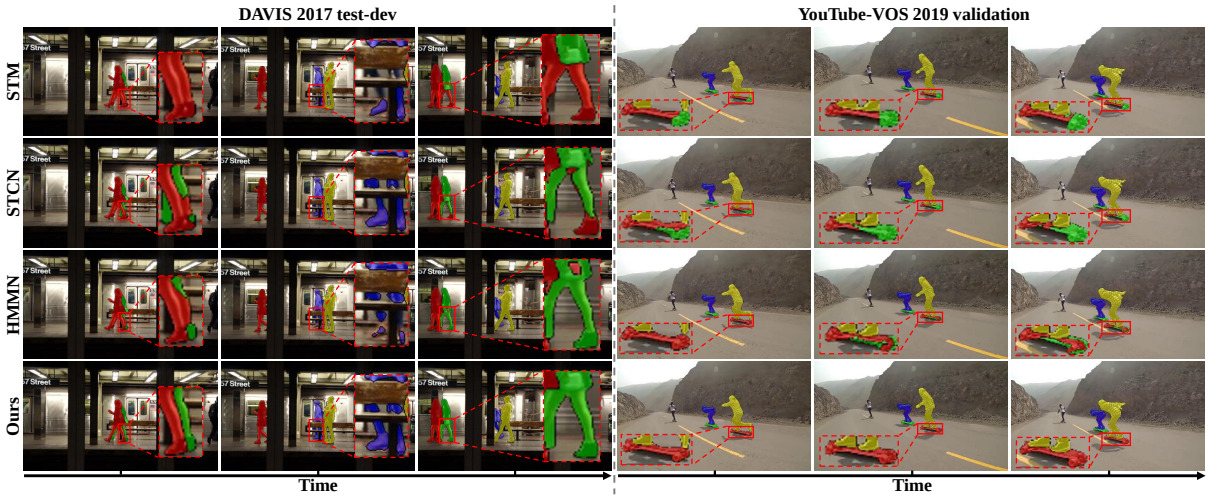


Figure 3. Qualitative comparison on DAVIS 2017 test-dev set and Youtube-VOS 2019 validation set. We compare DMNet with STM [15], STCN [10] and HMMN [19]. And we mark significant improvements using red boxes.

### 4.3. Challenge Results

We rank 1st place in the Track 1 (Video Object Segmentation) of the 4th Large-scale Video Object Segmentation Challenge (CVPR2022). Overall, we utilize 11 models from 3 frameworks, including 4 DMNet [21] models, 3 AOT-L [28] models, and 4 STCN [4] models. These models share the same framework but diverse backbones. As shown in Table 3, our team achieves the best performance on the overall and seen scores.

## 5. Conclusion

In this paper, we propose a dynamic matching network (DMNet) based on adversarial learning framework including a dynamic pixel-aware correspondence module (Pixel-CM) and a dynamic part-aware alignment module (Part-AM) for robust VOS. Specifically, Pixel-CM is designed to further optimize the correspondences within the local window and Part-AM is adopted to discriminate detailed differences between the predicted mask and the ground truth. Our solution achieves the 1st place with the $\mathcal{G}$ score of 87.2% on test set on the 4th Large-scale Video Object Segmentation Challenge.

## References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4

[2] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, pages 8890–8899, 2020. 3

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, pages 5559–5568, 2021. 4

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *NIPS*, 2021. 2, 3, 4, 5

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NIPS*, 26:2292–2300, 2013. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[7] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *CVPR*, pages 4144–4154, 2021. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[9] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, pages 2869–2878, 2020. 3

[10] Shuxian Liang, Xu Shen, Jianqiang Huang, and Xian-Sheng Hua. Video object segmentation with dynamic memory networks and adaptive object alignment. In *ICCV*, pages 8065–8074, 2021. 2, 4, 5

[11] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 3

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[13] King Ngi Ngan and Hongliang Li. *Video segmentation and its applications*. Springer Science & Business Media, 2011. 1

[14] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, pages 7376–7385, 2018. 1

[15] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019. 3, 4, 5

[16] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 3, 4

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 4

[18] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, pages 629–645. Springer, 2020. 1, 4

[19] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *ICCV*, pages 12889–12898, 2021. 1, 2, 4, 5

[20] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *PAMI*, 38(4):717–729, 2015. 3

[21] Rui Sun, Yuan Wang, Tianzhu Zhang, Zhe Zhang, Yongdong Zhang, and Feng Wu. Dynamic matching network for video object segmentation, 2022. 1, 2, 4, 5

[22] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 2

[23] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 3

[24] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *CVPR*, pages 1286–1295, 2021. 1

[25] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018. 4

[26] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 3, 4

[27] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, pages 332–348. Springer, 2020. 4

[28] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *NIPS*, 2021. 2, 4, 5

[29] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 4

[30] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7234–7243, 2019. 3

[31] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *CVPR*, pages 669–677, 2016. 1