

Supplementary Material for “Hierarchical Interaction Network for Video Object Segmentation from Referring Expressions”

Zhao Yang^{1*}

Yansong Tang^{1*}

Luca Bertinetto³

Hengshuang Zhao^{1, 2}

Philip H.S. Torr¹

¹ Torr Vision Group

University of Oxford

² The University of Hong Kong

³ Five AI Ltd.

Abstract

This supplementary document provides implementation and experimental details complementing those reported in the main paper. The contents are organized as follows:

- Section 1 describes the implementation details.
- Section 2 discusses the datasets and the evaluation metrics we adopted.
- Section 3 compares the performance of our model with that of Khoreva *et al.* [1] on subgroups of DAVIS-16 videos which are organized based on their key characteristics.
- Section 4 visualizes the differences in prediction results from several ablated versions of our model and complements our quantitative ablation study in Section 4.3 of the main paper.
- Section 5 discusses and compares the benefits and the downside of utilizing optical flow in our architecture.
- Section 6 visualizes our prediction results on several video segments in which the queried objects temporarily stay stationary.
- Section 7 visualizes prediction results on all of the four benchmark datasets.

1 Implementation details

During pre-training, we adopt batch size 24 and 50,000 training iterations with a standard cross-entropy loss. The initial learning rate is 0.01 and follows a “poly” adjustment policy [2], where the initial learning rate is multiplied by $(1 - \frac{iter}{total_iters})^{0.9}$ at each iteration. Network parameters are optimized via stochastic gradient descent with weight decay 0.0001. To improve training speed and memory efficiency, we use mixed precision during training.

The BERT model has a hidden size of 768, which corresponds to c_s in Section 3.1 of the main paper and is initialized with weights provided by Bellver *et al.* [10]. The official BERT model was pre-trained using two unsupervised tasks, “masked language model” (MLM) and “next sentence prediction”, with data from BookCorpus [24] (a dataset containing 11,038 unpublished books) and English Wikipedia (excluding lists, tables, and headers). WordPiece [24] with a 30,000 token vocabulary was used as the tokenizer. Bellver *et al.* [10] further fine-tuned the model for one epoch with natural language expressions from the RefCOCO dataset [25] using the MLM objective. Every input sentence is prepended with a special token [CLS] and appended a special token [SEP] indicating the start and the end of the sentence respectively.

We initialize ResNet-101 [9] with weights pre-trained on ImageNet [6] and randomly initialize all other layers in our model (except BERT). The ASPP module follows the same configuration as in [9]. It presents a total stride of 8 and the output channel number is 256, which corresponds to c_4 in Section 3.1 of the main paper. In addition, we set c_1 , c_2 , and c_3 in Section 3.1 of the main paper to 48, motivated by the practice in [9], which sets a smaller channel number for low-level and intermediate feature maps to prevent their outweighing the importance of the high-level feature maps. We employ RAFT [17], a state-of-the-art deep model for optical flow estimation.

Training. It is common practice for referring segmentation methods to first pre-train their models on a large-scale dataset (*e.g.* RefCOCO [10], MSRA [9], Kinetics [9], YouTubeVOS [16, 23], *etc.*) for bootstrapping purpose and then fine-tune them on the evaluation datasets [8, 9, 11, 16, 19]. We follow the same practice and pre-train our model on YouTubeVOS and then fine-tune our model on DAVIS-16, DAVIS-17, and A2D for evaluation respectively. During fine-tuning on DAVIS-16, DAVIS-17, and A2D, we adopt batch size 8, an initial learning rate of 0.001 with the “poly” adjustment policy [9], and weight decay 0.0001, and optimize network parameters via stochastic gradient descent. Following prior work on semi-supervised VOS [18, 24], we only optimize the losses of the hardest 15% pixels. Similarly to previous work [12, 19, 20], we use class-balanced cross-entropy loss with weight 0.9 for the background class and weight 1.1 for the object class, which we found useful for compensating for the object class as objects are usually much smaller than the background. Finally, for DAVIS-16 and DAVIS-17, we clip the gradient norm to 1.0 during training, which we empirically found important for stabilizing training. The model is trained for 10 epochs on DAVIS-16 and DAVIS-17, and 20 epochs on A2D. During both pre-training and fine-tuning, we resize each frame to the height of 480 pixels and then randomly crop a region of size 480×480 pixels from the frame. Raw predictions are upsampled via bilinear interpolation to the size of the ground-truth masks.

Inference. Each frame is evaluated at its original size. We take the *argmax* along the channel dimension of the score maps as the predicted labels. When evaluating our method on DAVIS-17, we adopt a simple strategy to combine the prediction results of multiple objects in a single frame when conflicts arise. Specifically, the predicted mask of an object with a higher ID number overrides that of an object with a lower ID number.

2 Datasets and evaluation metrics

Datasets. We evaluate the performance of our model on four datasets, DAVIS-16 [10, 24], DAVIS-17 [10, 19], A2D [6, 22], and J-HMDB [6, 8]. The DAVIS-16 dataset annotates a single foreground entity for each video and contains 30 videos for training and 20 videos for

validation. The DAVIS-17 dataset annotates up to several objects per video and contains 60 videos for training and 30 videos for validation. Khoreva *et al.* [14] augmented the DAVIS-16 and DAVIS-17 datasets with natural language expressions from human annotators for all the original objects. For both DAVIS-16 and DAVIS-17, two expressions are provided based on annotators’ observation in the first frame. For DAVIS-17, two more expressions are provided based on annotators’ observation throughout the whole video. We train our algorithm on the training set and evaluate on the validation set. The Actor-Action Dataset (A2D) by Xu *et al.* [22] is a benchmark dataset for the task of actor-action video segmentation. There are 43 valid actor-action pairs. The original task of actor-action segmentation requires to assign an actor-action class label (or a background label) to each pixel in each frame of the input video. Gavriluk *et al.* [8] introduced the task of actor-action segmentation from a natural sentence, and extended the A2D dataset with natural language descriptions of what action an actor is performing in a video, leading to 6,656 sentences describing 6,656 target objects. There are 3,036 videos for training and 746 videos for testing. We train our model on the training set and evaluate on the test set. The Joint-annotated Human Motion Data Base (J-HMDB) [8] is a fully annotated dataset for human actions and human poses. There are 928 videos divided into 21 action classes. It provides pixel-level mask annotations in the form of articulated human puppets. Gavriluk *et al.* [8] extended this dataset with 928 natural language expressions each describing what the actor is doing in each video. Following previous work [14, 19, 20], we evaluate our method on all 928 videos from this dataset using weights fine-tuned with A2D.

Evaluation metrics. On DAVIS-16 and DAVIS-17, we adopt the official evaluation metrics of mean region similarity \mathcal{J} , which is the intersection-over-union of the prediction and ground truth, and mean contour accuracy \mathcal{F} , which is the F-measure defined on contour points from the prediction and the ground truth. On DAVIS-17, we report performance under two settings. The first is the standard setting which requires that no pixel be assigned more than one object ID. The second setting is for fair comparison with URVOS [16], which did not address potential overlaps in the predictions of multiple objects (therefore an easier setting) and we refer to it as *binary evaluation* in Table 2 of the main paper. On A2D and J-HMDB, we adopt the common metrics of overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision at five threshold values. The overall IoU is measured as the ratio between the total intersection area and the total union area of all test samples (each test sample is a language query and a video frame). This metric favors large objects. The mean IoU is the IoU between the prediction and ground truth averaged across all test samples. This metric treats large and small objects equally. The precision metric measures the percentage of test samples that passes a certain IoU threshold. We evaluate precision at the common IoU thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9.

3 Attribute-based performance on DAVIS-16

In Table 1, we provide an analysis of our method when evaluated on DAVIS-16 videos assigned to different attribute groups. Our method achieves the best performance in most attribute groups and is robust against a wide variety of challenging scenarios.

| Attribute | AC | LR | SV | SC | CS | DB | BC | FM | MB | DEF | OCC |
|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Khoreva <i>et al.</i> [10] | 80.1 | 79.0 | 74.4 | 77.6 | 85.7 | 66.4 | 85.0 | 77.7 | 78.1 | 84.3 | 80.1 |
| RefVOS(re-implement) [10] | 68.3 | 67.2 | 67.2 | 64.5 | 79.7 | 48.5 | 89.3 | 63.7 | 66.5 | 71.2 | 68.2 |
| Ours | 85.1 | 88.3 | 83.3 | 77.1 | 92.3 | 69.2 | 90.3 | 82.6 | 80.7 | 83.7 | 84.7 |

Table 1: Attribute-based results (\mathcal{J}) on the DAVIS-16 validation set. AC: appearance change. LR: low resolution. SV: scale variation. SC: shape complexity. CS: camera shake. DB: dynamic background. BC: background clutter. FM: fast motion. MB: motion blur. DEF: deformation. OCC: occlusions.

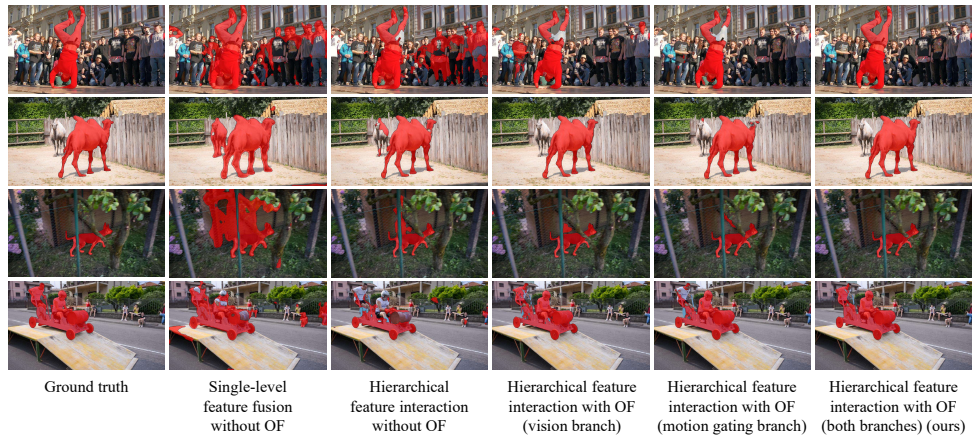
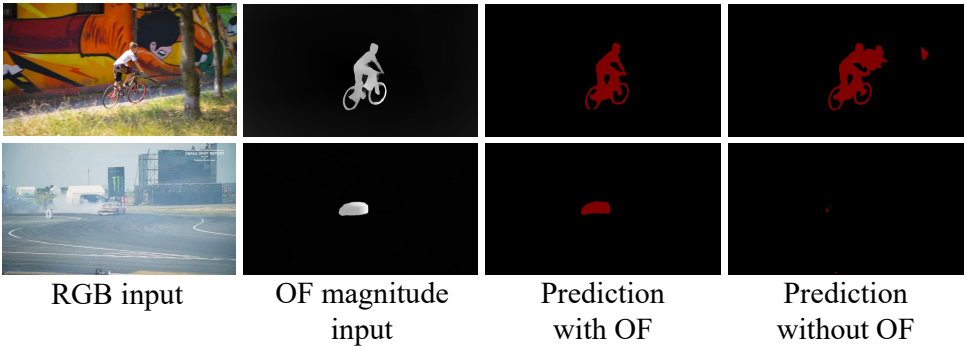


Figure 1: We visualize the effects of our key contributions on examples from DAVIS-16. From left to right, each column visualizes masks from the ground truth, model that employs single-level feature fusion without utilizing any optical flow input, models that employ hierarchical feature interaction only between language and vision features without any optical flow input, with optical flow input only to the vision branch, with optical flow input only to the motion gating branch, and with optical flow input to both branches, respectively.

4 Ablative qualitative analysis

In Fig. 1, we compare the different prediction results generated by models that employ only single-level feature fusion and our hierarchical feature interaction scheme (both models do not utilize any optical flow input); and then we demonstrate the effects of utilizing optical flow input to the vision branch, to the motion gating branch, and to both branches, respectively, under our hierarchical feature interaction scheme. It can be seen that hierarchical feature interaction even without optical flow input generate fewer false positives and more accurate masks compared to its single-level feature fusion counterpart. Moreover, when employing optical flow input either to the vision branch or to the motion gating branch helps the model focus on the correct target object, which shows the advantage of utilizing motion information. Finally, incorporating optical flow in both branches generate the best results, which further shows that our two ways of utilizing optical flow complement each other and are both necessary for achieving the best results.

Positive effects of OF



Negative effects of OF

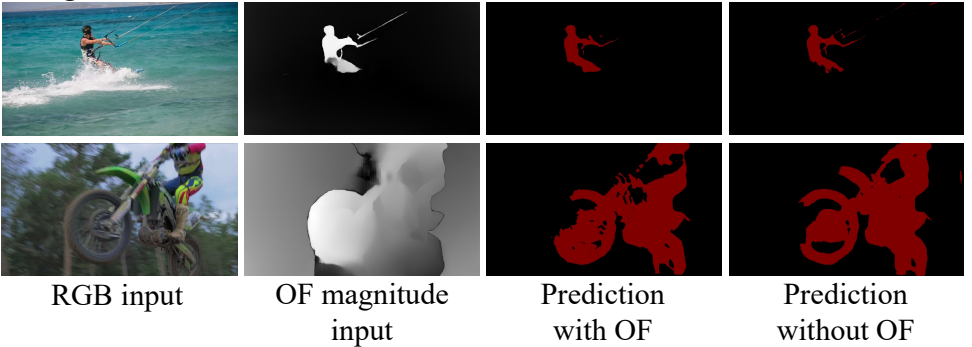


Figure 2: Incorporating optical flow for prediction have both positive and negative effects. However, as discussed in the text of this section as well as demonstrated in Table 4 of the main paper, the benefits of utilizing optical flow far outweighs the drawbacks.

5 The effects of optical flow

An important question is that if we employ optical flow for highlighting moving objects, then how much does poor optical flow signals or static target object that has no motion signals affect our predictions. We discuss the benefits and the downside of using optical flow in our architecture with a few examples in Fig. 2. First of all, as can be seen in the first two rows under “Positive effects of OF”, the main benefit for using optical flow is to complement language for more accurate object localization. In many cases, as language referral is an extremely challenging problem which we already know, the expression itself often generates false positive or false negative predictions. These are the cases where optical flow magnitude (computed after subtracting the mean vector) comes into play and helps correcting those false positives or false negatives.

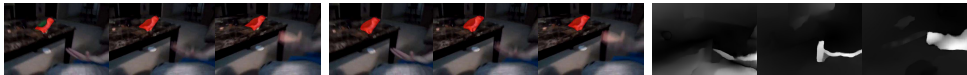
In some cases (the two rows under “Negative effects of OF”), incorporating OF produces slight negative effects. In the first case, motion of waves causes the model to also segment out water near the legs of the referent person. In the second case, due to large camera rotation angles and the target being near to the camera, subtracting the mean vector cannot produce decent separation result between the moving target and the static background. The messy and uninformative optical flow magnitude image does not provide much helpful information

and conversely includes many distracting signals. Even in this case, the prediction result from incorporating optical flow is only slightly worse than that from the RGB-only model, missing only small detailed parts of the motorbike and the biker. This shows that when optical flow signals are of poor quality, our model can still rely on language to generate decent predictions.

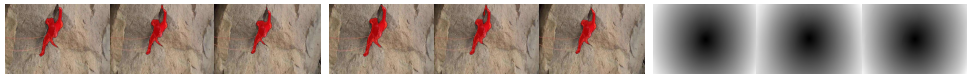
Finally, we want to clarify that the overall benefits of optical flow far outweighs its negative effects. There are two main reasons. First, consistent with the pattern shown in these four examples, optical flow tend to make *big* improvements by correcting false positives or negatives, and only introduce *small* errors in the cases where they are harmful. Second, in all four benchmarks, the cases where optical flow plays a positive role far outnumber those where optical flow negatively affect the prediction results. And this is due to the fact that the queried target objects in videos are more often moving objects and it is also relatively easy to separate their motions from the background by subtracting the mean vector from the flow field.

6 Static objects

Query 1: “Parrot is jumping from the table.”



Query 1: “Person climbing on the rock.”



Query 1: “Yellow cat at the bottom walking.”



Query 1: “A man in black is standing on the left and helping a girl.”



Predictions from the static model

Predictions from the full model

The optical flow magnitude images

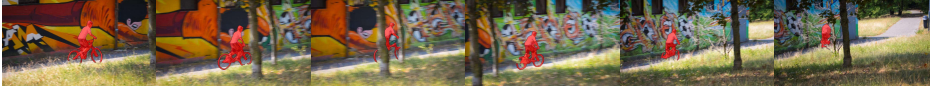
Figure 3: Predictions on a video segment in which the queried object is stationary. We include predictions from the static model which only employs language as guidance for comparison with predictions from our full model. When motion signals about the queried object are missing, our model can still leverage language to locate the object and generates similar or better masks compared with its static counterpart.

7 Qualitative evaluation on all benchmarks

Query 1: "A black swan."



Query 1: "A boy riding a bmx bike and a bike."



Query 1: "A man in red doing breakdance."



Query 1: "A brown camel."



Query 1: "A black car."



Query 1: "A silver car."



Query 1: "A cow with a bell around its neck."



Query 1: "A girl dancing."



Query 1: "A golden retriever walking in the grass."



Query 1: "A red and white car."



Figure 4: Segmentation results on the DAVIS-16 validation set.

Query 1: “A drift car on a straight road.”



Query 1: “A goat walking on rocks.”



Query 1: “A girl riding a horse and a horse.”



Query 1: “A man kite-surfing.”



Query 1: “A dog running in the garden.”



Query 1: “A man riding a motorbike in colorful biking gear and a motorbike.”



Query 1: “A man launching the paraglider.”



Query 1: “A man jumping across fences.”



Query 1: “A man riding a black scooter in suit and a black scooter.”



Query 1: “A blue wooden car and two men pushing it.”



Figure 5: Segmentation results on the DAVIS-16 validation set.



Figure 6: Segmentation results on the DAVIS-17 validation set, excluding repetitive videos also contained in the DAVIS-16 validation set.

Query 1: "A fat man on the right in a black jacket."

Query 2: "A cardboard box held by a man."

Query 3: "A man on the left with a beard wearing jeans."



Query 1: "A biker man in a white tshirt."

Query 2: "A black stunt bike."



Query 1: "A man riding a motorbike."

Query 2: "A green motorbike."



Query 1: "A black harness with an airbag."

Query 2: "A man launching a paraglider."

Query 3: "Wing risers with cascade."



Query 1: "A brown and white colored piglet."

Query 2: "A brown piglet in the middle."

Query 3: "An adult pig on the right."



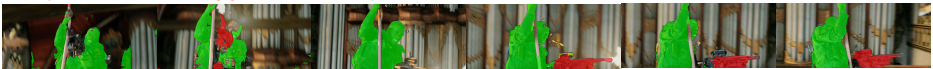
Query 1: "A man in a suit riding a scooter."

Query 2: "A black scooter ridden by a man."



Query 1: "A black shooting gun."

Query 2: "A black man."



Query 1: "A blue wooden car."

Query 2: "A man in a white helmet driving a wooden car."

Query 3: "A man wearing a white shirt on a wooden car without a helmet."



Figure 7: Segmentation results on the DAVIS-17 validation set, excluding repetitive videos also contained in the DAVIS-16 validation set.

Query 1: "Cat eating food in the bowl on the ground."
 Query 2: "Small fluffy puppy biting the cat."



Query 1: "Car jumping into the water."
 Query 2: "A car is parked on the left."



Query 1: "Woman in green dress is walking on the street."
 Query 2: "Man with a purple backpack walking on the right."



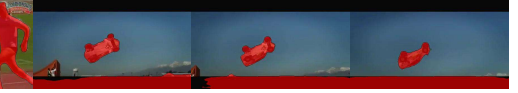
Query 1: "A baby in white shirt start walking."
 Query 2: "The person on the right is squatting."



Query 1: "Woman is staple running on the athletic track."



Query 1: "Car jumping from the ramp."



Query 1: "Black and white dog rolling on the meadow."
 Query 2: "The person is watching a dog."
 Query 3: "Small white dog walking on the right."



Query 1: "Baby crawling in the corridor."
 Query 2: "The dog on the right is crawling."
 Query 3: "Dog on the left crawling."



Query 1: "Guy is dribbling a ball around orange cones."
 Query 2: "The basketball is being dribbled."



Query 1: "Man is dribbling a ball on the basketball court."
 Query 2: "The basketball is being dribbled by the bald man."



Query 1: "Man in green shirt standing."
 Query 2: "Man in yellow shirt jumping over a man."



Query 1: "Man throwing a ball."
 Query 2: "A dark ball is flying on the air."



Query 1: "Small boy trying to walk with the help of his mother."
 Query 2: "Woman is supporting a baby."



Query 1: "The dog is standing behind a cat."
 Query 2: "Cat rolling on the asphalt near the dog."



Figure 8: Segmentation results on challenging videos from the A2D test set.

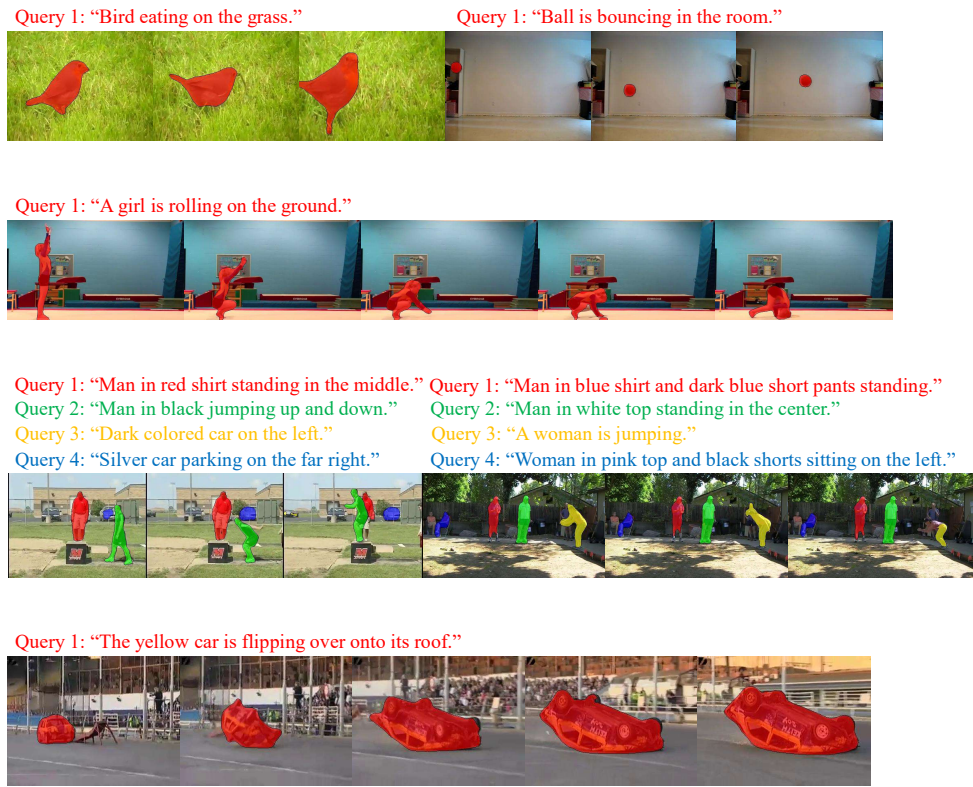


Figure 9: Segmentation results on challenging videos from the A2D test set.

Query 1: "Man clapping hands."



Query 1: "Man in black pants pulling up."



Query 1: "Man throwing something into the river."



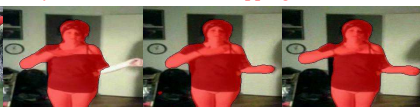
Query 1: "A top naked man is shooting a pistol."



Query 1: "Man shooting his bow."



Query 1: "Woman in black clapping her hands."



Query 1: "A girl is combing her hair."



Query 1: "Man picking up the basketball."



Query 1: "Woman throwing darts."



Query 1: "Man green shirt playing golf."



Query 1: "Man in black shirt shooting guns."



Query 1: "Woman in red top and black shorts pulling up."



Query 1: "Man shooting guns."



Query 1: "Boy swing the baseball."



Query 1: "A man in red jersey is shooting in a football game."



Query 1: "A man is throwing a ball toward the basket."



Figure 10: Segmentation results on challenging videos from the J-HMDB dataset.

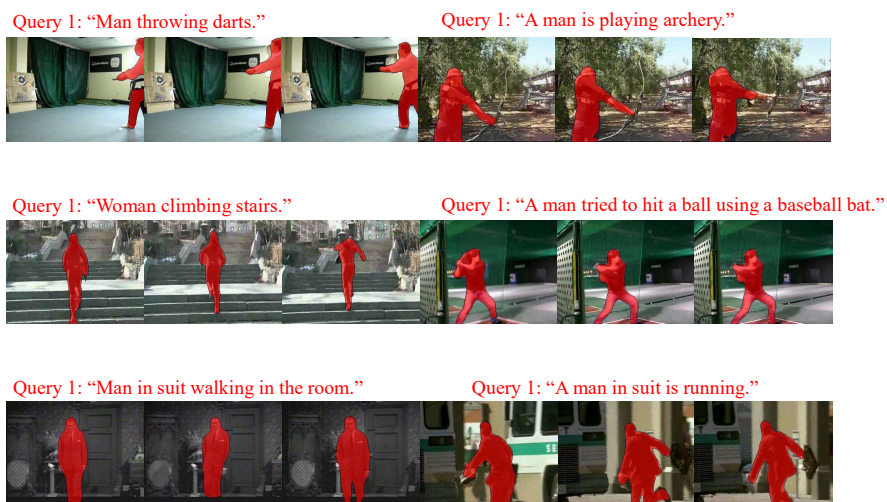


Figure 11: Segmentation results on challenging videos from the J-HMDB dataset.

References

- [1] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. Refvos: A closer look at referring expressions for video object segmentation. *arXiv preprint arXiv:2010.00263*, 2020.
- [2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Re-thinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.
- [4] Ming-Ming Cheng, Niloy J. Mitra, Xiaolei Huang, Philip H. S. Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 37(3):569–582, 2015.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.
- [6] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [8] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *ICCV*, 2013.
- [9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [10] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014.
- [11] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018.
- [12] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *CVPR*, 2017.
- [13] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *CVPR*, 2020.
- [14] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016.
- [15] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. In *CVPR Workshops*, 2017.

- [16] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020.
- [17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.
- [18] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019.
- [19] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *ICCV*, 2019.
- [20] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *AAAI*, 2020.
- [21] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- [22] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, 2015.
- [23] 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*, 2018.
- [24] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020.
- [25] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.
- [26] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015.