Hierarchical Interaction Network for Video Object Segmentation from Referring Expressions

Zhao Yang¹* Yansong Tang¹* 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

In this paper, we investigate the problem of video object segmentation from referring expressions (VOSRE). Conventional methods typically perform multi-modal fusion based on linguistic features and the visual features extracted from the top layer of the visual encoder, which limits these models' ability to represent multi-modal inputs at different semantic and spatial granularity levels. To address this issue, we present an endto-end hierarchical interaction network (HINet) for the VOSRE problem. Our model leverages the feature pyramid produced by the visual encoder to generate multiple levels of multi-modal features. This allows more flexible representation of various linguistic concepts (e.g., object attributes and categories) in different levels of the multi-modal features. Moreover, we further extract signals of moving objects from optical flow input, and utilize them as complementary cues for highlighting the referent and suppressing the background with a motion gating mechanism. In contrast to previous methods, this strategy allows our model to make online predictions without requiring the whole video as input. Despite its simplicity, our proposed HINet improves over the previous state of the art on the DAVIS-16, DAVIS-17, and J-HMDB datasets for the VOSRE task, demonstrating its effectiveness and generality.

1 Introduction

As a fundamental task in computer vision, video object segmentation (VOS) has various realworld applications including augmented reality [5], robotics [25], video surveillance [50], and video editing [59]. According to the level of supervision provided at test time, the VOS problem has traditionally been tackled in the *semi-supervised* (initialized by pixel-level masks or scribbles) [50], 50 or *unsupervised* settings [6], 50. In comparison, the recently introduced task of video object segmentation from referring expressions (VOSRE) aims to segment out an object from a video with the guidance of a natural language description. This setting offers several advantages: (1) Language annotations are much more economical and scalable than pixel-level masks [27], [23], while they are also more natural and intuitive

^{© 2021.} The copyright of this document resides with its authors.

It may be distributed unchanged freely in print or electronic forms.

^{*} Both authors have contributed equally. yansong.tang@eng.ox.ac.uk. zhao.yang@eng.ox.ac.uk.

Query 1: "A horse doing high jumps." Query 2: "A woman riding a horse."



Figure 1: The goal of VOSRE is to predict a pixel-level mask, in every frame, that delineates the object described by a language query. Masks in this example are our predictions.

compared with scribbles for providing target identification, especially on mobile devices. (2) Differently from the unsupervised setting where no target specification is given, language queries greatly reduce the ambiguity in problem definition.

Due to the reasons above, VOSRE has attracted growing attention in recent years. With the purpose of aligning semantic concepts between language and vision, the most existing methods focus on developing new mechanisms for combining linguistic and visual features, $(e.g., dynamic convolutions [\begin{tabular}{ll} 23, \begin{tabular}{ll} 23,$

In other related fields including object detection and segmentation, the importance of utilizing not only top-level feature maps but an array of feature maps from different levels is well known and has been widely explored. For instance, the feature pyramid network [29] extracts region-of-interest features from outputs of five different stages in the backbone encoder to effectively capture objects of different scales, and PSPNet [53] and DeepLabV3+ [5] combine low-level and intermediate features with high-level features for recovering more accurate spatial and appearance details. As for our problem, a linguistic phrase typically describes multiple concepts about the referent object, such as its category, attributes, actions, *etc.* These different concepts may not all be best represented by the same level of visual feature map, *e.g.*, clues for a described color may be more evident in lower-level features, and those for a category are more likely to be embedded in high-level features.

Motivated by these observations, we propose a hierarchical interaction network (HINet) to learn more effective multi-modal representations for VOSRE. The HINet incorporates a separate set of linguistic features with the visual feature maps from the different convolutional blocks of the visual encoder. This simple strategy generates multi-level, multi-modal representations at different semantic and spatial granularity levels, which greatly improves the accuracy of object localization. Furthermore, unlike previous methods which typically employ a 3D convolutional network [12], 53, 52] or space-time attention network [13] to extract temporal information from multiple frames, we extract motion signals of moving objects from optical flow input and directly exploit them as complementary source of guidance for object localization. Specifically, we utilize optical flow magnitude input in two ways: (1) incorporating it with the RGB input to *implicitly* embed motion information in the resulting visual feature pyramid; (2) encoding it into high-dimensional features and using them as gating values to explicitly strengthen multi-modal features of moving objects and suppress those of static regions. We conduct experiments on the DAVIS-16 [11], DAVIS-17 [12], A2D [56], and J-HMDB [20] datasets. Despite its simplicity and online operability, our method sets new state of the art on three datasets and performs competitively on the other.



Figure 2: Comparison of different multi-modal feature fusion methods for VOSRE. (a) Single-level multi-modal feature fusion $[\square, \square3, \square3, \square3]$. (b) Single-level multi-modal feature fusion followed by deconvolutions $[\square3]$. (c) Multi-resolution multi-modal feature fusion in which **only the top-level visual features** are upsampled to multiple resolutions and then combined with linguistic features $[\square3, \square3]$. (d) Our hierarchical feature interaction approach, which combines linguistic features with visual features from **multiple levels of the encoder**.

2 Related Work

Referring image segmentation. In the image domain, a variety of methods for referring object segmentation perform multi-modal feature fusion by different strategies, *e.g.*, concatenation [13], recurrent interaction [23, 50], cross-modal attention [6, 13, 14, 51], multi-modal graph reasoning [13], and linguistic structure-guided context modeling [13]. Different from previous methods which leverage multi-level features for multi-modal fusion[6, 13, 14, 13], [14], [15], [14], [15], [16], [17], [18], [19], [19], [19], [10], [10], [10], [10], [10], [11], [12], [12], [13], [14], [15], [16],

VOS. The VOS problem is traditionally tackled in the semi-supervised or unsupervised settings. For *semi-supervised VOS*, the "supervision" refers to human guidance at test time and comes in the form of pixel-level masks or scribbles in the first frame [[12]]. While early methods [[0, [22], [22], [32], [41], [51]] apply an online fine-tuning strategy to adapt to object appearance at test time, they are particularly slow and impractical in many use cases. Instead, recent methods directly propagate the first-frame annotations to the rest frames with sequential modeling approaches, *e.g.*, leveraging space-time memory network [[52]], collaborative foreground-background integration [[50]] and many others [[0, [22], [32], [32], [33], collaborative foreground-background integration [[50]] and many others [[0, [22], [32], [32], [33], assume a single target object per video and the methods rely on optical flow [[22], [32], [32], [33], [33], assume a single target object per video and the methods rely on optical flow [[22], [32], [32], [33], [33], [33], [33], [33], [33], [34], [33], [34], [35]] for modeling appearance changes of the target object. Recent benchmarks [<math>[33] address the more practical and challenging multiple-object scenario. Researchers generally first predict an initial set of masks by an off-the-shelf instance segmentation model (*e.g.*, Mask R-CNN [[13]), and then associate individual mask predictions across different frames into consistent object tracks [[13], [53], [53]].

VOSRE. Fig. 2 displays a comparison of recent representative VOSRE approaches and our method in the multi-modal fusion stage. Note that we group the language-guided actoraction segmentation (L-AAS) methods [**12**, **19**, **28**, **51**, **53**, **54**] into the category of VOSRE, since the only minor difference is that expressions in L-AAS put more emphasis



Figure 3: Schematic illustration of our approach. We combine the multi-modal features with a hierarchical feature interaction strategy, which allows flexible integration of various concepts between the language and vision data, and optical flow information further helps for highlighting moving objects. In this figure, \odot denotes element-wise multiplication. Φ_i and Θ_i are 1×1 convolutions, where i = 1, 2, 3, 4. Ψ_j is two 3×3 convolutions connected by batch normalization and ReLU nonlinearity, where j = 1, 2, 3.

on describing what an actor is doing in each video. Besides the multi-modal fusion strategy, we further design a gating mechanism to strengthen the moving object compared with [23], which also utilizes optical flow as input. Moreover, our HINet does not require any iterative refinement procedure at inference time compared with [23].

3 Method

Fig. 3 illustrates the pipeline of our proposed *hierarchical interaction network* (HINet). Specifically, it employs three dedicated branches to extract high-dimensional features from the language query input, the video frame input, and the optical flow magnitude input, respectively. To effectively align language and vision semantics at different scales, we first fuse the extracted visual and linguistic features at multiple levels to obtain the "language-vision" feature maps. Then we extract multiple scales of motion feature maps from our motion gating branch and fuse them with "language-vision" feature maps to produce our final sets of multi-level, multi-modal feature maps, which capture motion information about objects that are moving. Finally, the multi-level, multi-modal feature maps are fused in a top-down manner for mask prediction. In the following, we describe our pipeline in more detail.

3.1 Multi-modal Feature Encoding

Encoding visual features. The input to our vision branch is a 4-channel image, which is a concatenation of the 3-channel RGB frame with the 1-channel optical flow magnitude image. We compute the optical flow magnitude after subtracting the mean flow vector from the optical flow field. We use the ResNet-101 [13] network and an Atrous Spatial Pyramid Pooling (ASPP) module [12] to extract multiple levels of visual feature maps and apply a 1×1

convolution for channel reduction to each level of features. The multi-level visual features, $V_i \in R^{c_i \times h_i \times w_i}$, $i \in \{1, 2, 3, 4\}$, correspond to outputs from the first three residual blocks in ResNet and the output from the ASPP module respectively, where c_i , h_i , and w_i refer to the number of channels, the height, and the width of the corresponding feature maps.

Encoding language features. We employ BERT [\square], a deep language representation model pre-trained from unlabeled text to extract a linguistic feature matrix, $G \in \mathbb{R}^{l \times c_s}$, from the input sentence. Here, *l* denotes the number of words and c_s denotes the number of channels. To generate sentence-level feature vectors for each level of the visual feature maps, we apply average pooling to *G* along the word dimension, and then apply four independent linear layers followed by batch normalization operations. We denote the output sentence-level feature vectors as $L_i \in \mathbb{R}^{c_i}$, $i \in \{1, 2, 3, 4\}$, where c_i denotes the number of channels and are the same ones defined for the visual feature maps.

Encoding optical flow features. We employ four independent 7×7 convolutional layers with stride 1 to extract four sets of motion feature maps from the optical flow magnitude image input (same as the one described earlier). Each of the four sets of motion feature maps are downsampled to the same spatial resolution as the corresponding visual feature maps. We denote the output motion feature maps as $O_i \in R^{c_i \times h_i \times w_i}$, $i \in \{1, 2, 3, 4\}$, where c_i , h_i , and w_i denote the number of channels, the height, and the width of the feature maps. These notations are consistent with those defined for the linguistic and visual features.

3.2 Hierarchical Feature Interaction

To effectively utilize various linguistic concepts (*e.g.*, object attributes and categories) for the VOSRE task, we fuse the multi-modal representations at different semantic and spatial granularity levels with a hierarchical feature interaction strategy. To achieve this, we first combine the extracted vision and language features, V_i and L_i , to generate "language-vision" multi-modal features. Then we fuse them with the motion features, O_i , to generate the final multi-modal feature maps. Concretely, our step to combine the vision and language features is described mathematically as follows

$$F'_i = \Phi_i(V_i \odot L_i), \quad i = 1, 2, 3, 4.$$
 (1)

Here \odot denotes element-wise multiplication, and Φ_i is a 1 × 1 convolution applied for feature fusion with the same number of input and output channels. The language feature vector L_i is element-wise multiplied with the visual feature vector at each spatial location on the visual feature maps V_i . Note that as described in the previous section, V_i and L_i have the same number of channels, which makes them compatible for element-wise multiplication. The 1 × 1 convolution Φ_i is employed to project the results to a different high-dimensional space. On the other hand, we combine the obtained "language-vision" features with the motion features by utilizing a motion gating branch, which aims to highlight the moving objects and suppress the static regions. Specifically, we design the fusion process as below

$$F_i = \Theta_i(F'_i \odot (1 + \tanh(O_i))), \quad i = 1, 2, 3, 4,$$
(2)

where tanh is the hyperbolic tangent non-linearity, and Θ_i is a 1×1 convolution applied for feature fusion. In this step, the tanh non-linearity first scales each element in O_i to the range of (-1,1), and then the addition of the constant 1 further scales each element to the range of (0,2). Therefore, $1 + tanh(O_i)$ is a scaling function which may scale up or scale down each element in F'_i . Empirically, as detailed in the ablation studies of Table 4, we found that this gating branch improves the performance of the model. The 1×1 convolution Θ_i has the same number of input and output channels, and performs a final projection to embed the re-scaled features into a different high-dimensional space.

3.3 Multi-modal Feature Decoding

To fully exploit the information embedded in the feature maps at different granularity levels, we employ a top-down decoding scheme for combining the multi-level, multi-modal feature maps, F_i , $i \in \{1, 2, 3, 4\}$, into a single set of feature maps for mask prediction. The fusion process can be described by the following recursive function

$$\begin{cases} T_4 = F_4, \\ T_i = \Psi_i([U(T_{i+1}); F_i]), & i = 3, 2, 1. \end{cases}$$
(3)

Here '[;]' denotes feature concatenation along the channel dimension, U represents upsampling via bilinear interpolation, and Ψ_i is two 3 × 3 convolutions connected by batch normalization and ReLU nonlinearity. In words, progressing in the order of F_4 , F_3 , F_2 , and F_1 , we repeat an "upsample-concatenate-project" procedure three times in a cascade manner. At each step, we fuse feature maps from the top-down path with feature maps at the current level. The final feature maps, T_1 , are projected into two class score maps by a classification layer, which is a 1 × 1 convolution with stride 1 and 2 output channels. The entire network is trained end-to-end with a cross-entropy loss. During inference, *argmax* along the channel dimension of the score maps are used as the prediction.

4 **Experiments**

To evaluate the effectiveness of the proposed HINet, we conduct experimental evaluations on four challenging datasets, including DAVIS-16 [23, 21], DAVIS-17 [23, 22], A2D [12, 56], and J-HMDB [12, 20]. In the following, we first describe the implementation details and evaluation metrics in Sec. 4.1, then we present extensive comparisons with the state-of-the-art methods in Sec. 4.2, and finally discuss several ablation studies in Sec. 4.3.

4.1 Implementation Details and Evaluation Metrics

We implement the HINet in PyTorch. We use the SGD optimizer with momentum and weight decay set to 0.9 and 0.0001, respectively. During training, we follow the procedure of previous approaches $[\square, \square, \square, \square, \square]$ of pre-training first and then fine-tuning. We pretrain our model on YouTubeVOS $[\square]$ and then fine-tune it on the experimental datasets for evaluation respectively. During pre-training, we adopt a batch size of 24 and train for 50K iterations. The initial learning rate is set to 0.01 with polynomial learning rate decay. During fine-tuning, we adopt a batch size of 8 and train for 10 epochs on DAVIS-16 and DAVIS-17, and 20 epochs on A2D, with initial learning rate 0.001. We use BERT [\square] and RAFT [\square] for extracting language and optical flow features, respectively. For evaluation metrics, on DAVIS-16 and DAVIS-17, we adopt the official evaluation metrics of mean region similarity \mathcal{J} [\blacksquare] and mean contour accuracy \mathcal{F} [\blacksquare]. On A2D and J-HMDB, we adopt the metrics of overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision at five threshold values (P@ α where α is a threshold value) [\square , \square , \square , \square]. We include detailed definitions for these metrics in the supplementary material.

Method	$\mathcal{J}(\%)$	$\mathcal{F}(\%)$	$\mathcal{J}\&\mathcal{F}(\%)$	year
Khoreva <i>et al</i> . [23]	82.8	-	-	2018
RefVOS [71.8	71.0	71.4	2020
HINet (Ours)	84.4	85.3	84.8	-

Table 1: Results on DAVIS-2016 validation set. \mathcal{J} and \mathcal{F} are defined in Section 4.1. $\mathcal{J}\&\mathcal{F}$ is the mean of \mathcal{J} and \mathcal{F} .

Method	1st frame	full video	binary eval.	year
Khoreva et al. [23]	39.3	37.1	×	2018
RefVOS [44.5	45.1	×	2020
HINet (Ours)	50.2	47.9	×	-
URVOS [51.6	_	\checkmark	2020
HINet (Ours)	52.0	50.4	\checkmark	-

Table 2: $\mathcal{J} \& \overline{\mathcal{F}}$ on DAVIS-2017 validation set. "1st frame" and "full video" refer to results using 1st-frame expressions and full-video expressions, respectively. For a fair comparison with URVOS[[1]], we also report performance under the "binary eval." setting, in which we evaluate each of the multiple referent objects in a video individually without resolving overlaps in the predicted masks. Please see our supplementary material for more details.

4.2 Comparison with Others

Results on DAVIS-16 and DAVIS-17. In Table 1, we evaluate our method against state-ofthe-art VOSRE methods on DAVIS-16. Our method outperforms that of Khoreva *et al.* [23] by 1.6 absolute point on the \mathcal{J} metric (only \mathcal{J} is available for their method). We outperform RefVOS [I] by more than 10 absolute points on both the \mathcal{J} and \mathcal{F} metrics. As Bellver *et al.* [I] did not report performance on DAVIS-16, we fine-tune their public model (which was pre-trained on RefCOCO [22]) on DAVIS-16 and report the performance.

In Table 2, we evaluate our proposed model on the DAVIS-17 validation set. We compare with the method of Khoreva *et al.* [23] and RefVOS [11] under the standard multipleobject segmentation setting. Our method outperforms both methods when leveraging either 1st-frame expressions or full-video expressions. RefVOS [II] is a static segmentation model which segments each frame independently and does not exploit any temporal information. In addition, it only generates multi-modal feature maps from the top-level visual feature maps and does not have a hierarchical feature interaction mechanism. Our performance gain compared with RefVOS [I] validates our approach to integrate motion information and perform multi-level, multi-modal feature fusion. Compared with the method of Khoreva *et al.* [23], our method achieves more than 10% (absolute points) improvement in J&F, while being end-to-end trainable and enjoys fewer components and significantly simpler training and inference procedures. We compare with URVOS [13] under the binary evaluation setting (see Sec. 2.1 in the supplementary material). Only performance on the 1st-frame expressions is available for URVOS, and we outperform them by a small margin. We note that our method operates in an online fashion without requiring access to future frames, while URVOS requires the whole video as input for inference.

Results on A2D and J-HMDB. In Table 3, we evaluate our method against state-of-theart language-guided actor-action segmentation methods on A2D and J-HMDB. On the A2D dataset, our method performs the best on 3 out of 7 evaluation metrics. Compared to our model, the contemporary models of Hui *et al.* [1] and Liu *et al.* [3] achieve better precision at the lower IoU thresholds of 0.5, 0.6, and 0.7 and a better mean IoU, but achieves lower precision at the higher IoU thresholds of 0.8 and 0.9. While they demonstrate an overall

	Method	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	oIoU	mIoU	Year
	Gavrilyuk <i>et al</i> . [🗳]	50.0	37.6	23.1	9.4	0.4	55.1	42.6	2018
	Wang <i>et al</i> . [🛂]	55.7	45.9	31.9	16.0	2.0	60.1	49.0	2019
	McIntosh <i>et al</i> . [53]	52.6	45.0	34.5	20.7	3.6	56.8	46.0	2020
D	Wang et al. [🛂]	60.7	52.5	40.5	23.5	4.5	62.3	53.1	2020
A	RefVOS [57.8	53.4	45.6	31.1	9.3	67.2	49.7	2020
	Hui et al. * [🛄]	65.4	58.9	49.7	33.3	9.1	66.2	56.1	2021
	Liu <i>et al</i> . * [🛄] (R2D)	59.0	52.7	43.4	28.4	6.8	64.9	51.5	2021
	Liu <i>et al</i> . * [🛄] (I3D)	65.5	59.2	50.6	34.2	9.8	65.3	57.3	2021
	HINet (Ours)	61.1	55.9	48.6	34.2	12.0	67.9	52.9	-
	Gavrilyuk <i>et al</i> . [69.9	46.0	17.3	1.4	0.0	54.1	54.2	2018
	Wang et al. [🛂]	75.6	56.4	28.7	3.4	0.0	57.6	58.4	2019
В	McIntosh <i>et al</i> . [53]	67.7	51.3	28.3	5.1	0.0	53.5	55.0	2020
J-HMD	Wang et al. [🛂]	74.2	58.7	31.6	4.7	0.0	55.4	57.6	2020
	RefVOS [73.1	62.0	39.2	8.8	0.0	60.6	56.8	2020
	Hui et al. * [🛄]	78.3	63.9	37.8	7.6	0.0	59.8	60.4	2021
	Liu <i>et al</i> . * [💶] (I3D)	81.3	65.7	37.1	7.0	0.0	61.6	61.7	2021
	HINet (Ours)	81.9	73.6	54.2	16.8	0.4	65.2	62.7	-

Table 3: Results on the A2D test set and the entire J-HMDB dataset. * denotes contemporary work published or to be published in 2021.

Method	$\mathcal{J}(\%)$	$\mathcal{F}(\%)$	$\mathcal{J}\&\mathcal{F}(\%)$	$\Delta \mathcal{J} \& \mathcal{F}$
Full model	84.4	85.3	84.8	0.0
Full model w/o. hierarchical feature interaction	78.8	78.1	78.5	-6.3
Full model w/o. optical flow (motion gating branch)	81.9	82.2	82.0	-2.8
Full model w/o. optical flow (vision branch)	82.0	81.3	81.6	-3.2
Full model (original optical flow)	78.7	80.4	79.5	-5.3
Full model (concatenate)	80.2	79.0	79.6	-5.2
Full model (sum)	82.0	82.0	82.0	-2.8

Table 4: Ablation study on the DAVIS-16 validation set. $\Delta \mathcal{J} \& \mathcal{F}$ denotes absolute declines in the average of \mathcal{J} and \mathcal{F} .

better object localization ability on the A2D dataset, our model tends to generate finer, more detailed segmentation masks. Our method outperforms the rest of the methods with notable advantages across most of the metrics (except mean IoU on which our performance is lower by 0.2% than that of [5]). Both Hui *et al.* [1] and Liu *et al.* [5] employ an I3D [2] backbone which is pre-trained on Kinetics-400 [2] for action classification. When using the same backbone as ours, the method of Liu *et al.* [5] obtains lower performance than ours ("(R2D)" in Table 3). Following prior work [5], 53, 54, 54, 54, we test the generalization ability of our model on J-HMDB with weights fine-tuned on A2D. As shown in Table 3, our method outperforms all other methods on all metrics, mostly by a significant margin. In particular, we improve over the second best method in precision at the 0.8 and 0.7 thresholds by as large as 8.0 and 15.0 absolute points, respectively. We also achieve non-zero precision at the very hard IoU threshold of 0.9. These results highlight the remarkable generalization ability of our model.

4.3 Ablation Study

We conduct several ablations to evaluate the effectiveness of several key design choices we have made, and each corresponds to a row below *Full model* in Table 4.

Hierarchical feature interaction. We remove the hierarchical feature interaction mech-



Figure 4: Qualitative results on DAVIS-2017. Comparing with single-level feature fusion, our hierarchical feature interaction method could generate results with fewer false positives, as indicated in red dashed areas. Moreover, our model could better distinguish different concepts between a "pig" (I) and a "piglet" (II), and learn the concepts of "brown" (III), "white" (IV), "adult" (V), and "black" (VI).

anism and simply feed the multi-modal feature maps generated from the output of the toplevel ASPP module to the final classification layer. This leads to the biggest performance drop (6.3 absolute points) among all ablations, which demonstrates the importance of integrating multiple levels of features which capture context information at different scales and encode semantics at different levels of complexity. Fig. 4 shows an example where the hierarchical feature interaction design enables learning of different linguistic attributes and fine-grained semantic concepts. The input query "a pig" generates a mask largely covering all three pigs in the video frame as desired, whereas the input query "a piglet" generates a mask mostly covering only the two small pigs. When more attributes are successively added to describe an object, our model learns to leverage these attributes for more refined referral results. Such trends are easy to see when our predictions are viewed in the orders of "a pig" -> "an adult pig" -> "a black adult pig", and "a piglet" -> "a brown colored piglet" -> "a brown and white colored piglet".

Optical flow. As shown in Table 4 ("(motion gating branch)"), removing the motion gating branch leads to a moderate performance drop of 2.8 absolute points. We further evaluate whether it is necessary to include optical flow magnitude in our input to the vision branch ("(vision branch)"). This ablation leads to a decline of 3.2 absolute points in performance. The above experiments show that motion information can be helpful when integrated both at an early stage (to the RGB input) and at a late stage (to the "language-vision" representations). In another ablation experiment, we replace optical flow magnitude with the original optical flow as input to our full model ("(original optical flow)"). The resulting large performance drop of 5.3 absolute points illustrates the importance of choosing optical flow magnitude as the representation. A comparison between this variant and the optical flow-could only bring marginal benefits over not using optical flow information (79.5% vs. 79.1%).

Feature fusion methods. We evaluate two variants which differ from our default model in how they combine the language, vision, and motion features. In the "(concatenate)" vari-

Model	S	Н	H + OF(V)	H + OF(M)	H + OF(V&M)	OF
Inference time (ms)	82	88	88 + 30	92 + 30	92 + 30	30
$\mathcal{J}\&\mathcal{F}(\%)$	70.2	79.1	82.0	81.6	84.8	-

Table 5: Speed-accuracy trade-off analysis. S and H respectively denote single-level and hierarchical feature interaction; H + OF(V) denotes H with optical flow input to the vision branch; H + OF(M) denotes H with a motion gating branch; H + OF(V&M) denotes our full model; OF represents the external optical flow estimation model. For timings in "x+y" format, "x" denotes network inference time and "y" denotes extra optical flow computation time. Time is measured one sample per forward and is averaged on the DAVIS-16 validation set.

ant, we replace multiplication in Equations (1) and (2) with concatenation followed by a 1×1 convolution for channel reduction. This variant leads to a large performance drop of 5.2 points. In the "(sum)" variant, we replace multiplication in Equations (1) and (2) with element-wise summation, and the performance decreases by 2.8 absolute points. These results validate our choice of using multiplication for fusing features from different modalities. Network latency analysis. In Table 5, we analyze the latency trade-off when employing our two computationally heavy designs: the hierarchical feature interaction scheme and the incorporation of optical flow inputs. Compared with the single-level feature fusion baseline ("S"), the hierarchical model without optical flow inputs ("H") boosts the performance by 8.9 absolute points with only 6 milliseconds of extra inference time, a 12.7% improvement in accuracy with just 7.3% increase in latency. The efficiency is due to the dilated convolutions adopted in our visual encoder, which makes the lower-level features V_3 and V_2 have the same resolution as the top-level features V_4 . In the following, we analyze the performance gains and extra latency of our optical flow variants relative to the optical flow-free baseline "H". Our full model ("H + OF (V&M)") achieves 7.2% improvement in accuracy with 39.8% increase in inference time. Optical flow estimation accounts for most of the inefficiency. When excluding optical flow estimation time, each of the three variants "H + OF(V)", "H + OF (M)", and "H + OF (V&M)" brings 3.7%, 3.2%, and 7.2% performance improvement with just 0.0%, 4.5%, and 4.5% extra latency, respectively.

5 Conclusion

In this paper, we have proposed a Hierarchical Interaction Network (HINet) for the VOSRE problem, which fuses multiple modalities (*i.e.*, language, vision, and motion) in an end-toend and multi-level framework. Our method effectively associates different concepts in the language expression with the corresponding visual features which contain different levels of semantic and spatial details. Extensive experiments on four standard benchmarks demonstrate the advantage of our method with respect to the state of the art. In the future, we hope this work can serve as a strong baseline for VOSRE and inspire applications to other related tasks, such as video-based visual question answering.

Acknowledgement

This work is supported by the EPSRC grant: Turing AI Fellowship: EP/W002981/1, EP-SRC/MURI grant EP/N019474/1. We would also like to thank the Royal Academy of Engineering, FiveAI, and Tencent.

References

- 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] Sergi Caelles, Kevis-Kokitsi. Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, 2017.
- [3] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multiobject segmentation. In *CVPR Workshops*, 2019.
- [4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.
- [5] Dimitris Chatzopoulos, Carlos Bermejo, Zhanpeng Huang, and Pan Hui. Mobile augmented reality survey: From where we are to where we go. In *IEEE Access*, 2017.
- [6] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *ICCV*, 2019.
- [7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
- [8] 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.
- [9] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *CVPR*, 2018.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. In NAACL, 2019.
- [11] Shubhika Garg and Vidit Goel. Mask selection and propagation for unsupervised video object segmentation. In *WACV*, 2021.
- [12] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
- [14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, 2017.
- [15] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016.
- [16] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, 2020.

- [17] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, 2020.
- [18] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In ECCV, 2020.
- [19] Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *CVPR*, 2021.
- [20] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards understanding action recognition. In *ICCV*, 2013.
- [21] 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.
- [22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014.
- [23] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018.
- [24] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. In *IJCV*, 2019.
- [25] Vikash Kumar, Emanuel Todorov, and Sergey Levine. Optimal control with learned local models: Application to dexterous manipulation. In *ICRA*, 2016.
- [26] Ruiyu Li, Kai-Can Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018.
- [27] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C.-C. Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In ECCV, 2018.
- [28] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In CVPR, 2017.
- [29] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, 2017.
- [30] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017.
- [31] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. In *TPAMI*, 2021.
- [32] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018.

- [33] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *WACV*, 2020.
- [34] Kevis-Kokitsi. Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. In *TPAMI*, 2018.
- [35] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *CVPR*, 2020.
- [36] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. In *TPAMI*, 2013.
- [37] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In CVPR, 2018.
- [38] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019.
- [39] Federico Perazzi. Video Object Segmentation. PhD thesis, ETH Zurich, 2017.
- [40] 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.
- [41] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In CVPR, 2017.
- [42] 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.
- [43] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In ECCV, 2020.
- [44] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018.
- [45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for largescale image recognition. In *ICLR*, 2015.
- [46] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018.
- [47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In ECCV, 2020.
- [48] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017.
- [49] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017.

- [50] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *CSUR*, 46(2):1–37, 2013.
- [51] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *BMVC*, 2017.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. In *TIP*, 2015.
- [56] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In CVPR, 2015.
- [57] 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.
- [58] Zhao Yang, Qiang Wang, Song Bai, Weiming Hu, and Philip HS Torr. Video segmentation by detection for the 2019 unsupervised davis challenge. In *CVPR Workshops*, 2019.
- [59] Zhao Yang, Qiang Wang, Luca Bertinetto, Song Bai, Weiming Hu, and Philip H.S. Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019.
- [60] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020.
- [61] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019.
- [62] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018.
- [63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.