Language-Bridged Spatial-Temporal Interaction for Referring Video Object Segmentation

Zihan Ding[1,4,5] Tianrui Hui[2,3] Junshi Huang[4] Xiaoming Wei[4] Jizhong Han[2,3] Si Liu[1,5*]
[1] Institute of Artificial Intelligence, Beihang University
[2] Institute of Information Engineering, Chinese Academy of Sciences
[3] School of Cyber Security, University of Chinese Academy of Sciences
[4] Meituan [5] Hangzhou Innovation Institute, Beihang University

Abstract

Referring video object segmentation aims to predict foreground labels for objects referred by natural language expressions in videos. Previous methods either depend on 3D ConvNets or incorporate additional 2D ConvNets as encoders to extract mixed spatial-temporal features. However, these methods suffer from spatial misalignment or false distractors due to delayed and implicit spatial-temporal interaction occurring in the decoding phase. To tackle these limitations, we propose a Language-Bridged Duplex Transfer (LBDT) module which utilizes language as an intermediary bridge to accomplish explicit and adaptive spatial-temporal interaction earlier in the encoding phase. Concretely, cross-modal attention is performed among the temporal encoder, referring words and the spatial encoder to aggregate and transfer language-relevant motion and appearance information. In addition, we also propose a Bilateral Channel Activation (BCA) module in the decoding phase for further denoising and highlighting the spatial-temporal consistent features via channel-wise activation. Extensive experiments show our method achieves new state-of-the-art performances on four popular benchmarks with 6.8% and 6.9% absolute AP gains on A2D Sentences and J-HMDB Sentences respectively, while consuming around 7× less computational overhead [1].

1. Introduction

Referring video object segmentation (RVOS), which aims to segment the target object referred by a natural language expression in video frames, is an emerging task at the intersection of computer vision and natural language processing. Different from semi-automatic video object segmentation (SVOS) [3, 7, 31, 38], where the target object is referred by the manually annotated mask in the first frame, RVOS is more challenging for identifying the targets due to the variance of free-form expressions. Providing a more natural way for human-computer interaction, RVOS opens up a wide range of applications including language-based video editing [5], language-guided video summarization [29], and video question answering [14, 39], *etc*.

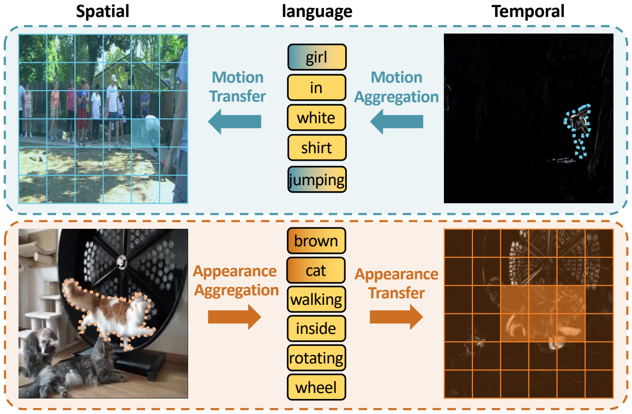

Figure 1. Illustration of our main idea. (Top) For *temporal→language→spatial* transfer, the referring words (*e.g.*, "jumping") can aggregate language-relevant motion information from the temporal features, which can help the spatial encoder recognize correct actions. (Bottom) For *spatial→language→temporal* transfer, the referring words (*e.g.*, "brown") can aggregate the language-relevant appearance information from the spatial features, which helps the temporal encoder remove the disturbance of background motion (*e.g.*, the rotating wheel).

The keys to solving RVOS are *spatial-temporal interaction* and *cross-modal alignment* [12, 41]. Existing methods mainly focus on the latter and design several mechanisms (*e.g.*, cross-modal attention [30, 37], capsule routing [25], and dynamic convolution [6, 36]) to mine the semantic correspondence between vision and language modalities. However, all these methods have limitations on spatial-

*Corresponding author
[1] https://github.com/dzh19990407/LBDT

temporal interaction due to the reliance on 3D ConvNets (*e.g*., I3D [1]). Concretely, since the poses and locations of moving objects vary in adjacent frames, aggregating spatially misaligned multi-frame features via 3D operators (*e.g*., 3D convolution and 3D pooling) may confuse the original appearance information in the target frame, leading to inaccurate segmentation results.

To alleviate this phenomenon, CSTM [12] introduces an additional 2D spatial encoder (*e.g*., ResNet [8]) to extract undisturbed appearance information of the target frame, which is fused with features of the temporal encoder in the later decoding phase. However, the spatial encoder of CSTM lacks motion information since it doesn't explicitly interact with the temporal encoder, making it hard to distinguish among objects with similar appearances while performing different actions. Thus, it tends to generate high responses on false objects and introduces noises inevitably.

In this paper, we argue that an explicit interaction between spatial and temporal features should be established earlier in the encoding phase, forming a more sufficient and effective information exchange process between encoders. Moreover, naive spatial-temporal interaction still tends to introduce noises due to redundant information contained in language-irrelevant distractors. Therefore, we believe *the language expression can be exploited as the medium to bridge spatial and temporal interaction*, where only language-relevant information can be transferred between encoders for valid context aggregation. To this end, we propose a novel Language-Bridged Duplex Transfer (LBDT) module for effective spatial-temporal interaction in the encoding phase. As illustrated in Figure 1, motion information from the temporal encoder is first aggregated to the referring words by cross-modal attention. Then, the spatial encoder can obtain language-relevant motion clues from the referring words by reversed cross-modal attention, which assists in identifying the referred object by recognizing correct actions (Figure 1 top). Similarly, appearance information from the spatial encoder is also transferred to the temporal encoder through the language bridge, which facilitates the temporal encoder to distinguish language-relevant foreground objects from complex backgrounds (Figure 1 bottom). In addition, we also remove the dependence on 3D ConvNets and approximate motion information with frame difference processed by a 2D ConvNet. By this means, the model complexity is significantly reduced as 2D ConvNet occupies nearly 30× less computational overhead compared to 3D ConvNet (e.g., 3.6 *vs*. 107.9 GFLOPs) [2].

To exploit rich multi-scale contexts of hierarchical visual features for finer mask predictions, we also propose a Bilateral Channel Activation (BCA) module to adjust different feature channels in the decoding phase. Concretely, we first upsample and add multi-level features together in temporal and spatial decoders respectively to obtain the decoded features, on which linguistic features are utilized to filter out language-irrelevant motion and appearance information by channel-wise activation. Meanwhile, the global contexts of the decoded features are further extracted to activate the spatial-temporal consistent channels for highlighting features of the referred object.

In a nutshell, our contributions are three-fold: 1) We propose a Language-Bridged Duplex Transfer (LBDT) module to conduct spatial-temporal interaction explicitly between two independent 2D ConvNets in the encoding phase for RVOS, where we use referring words as the medium to transfer language-relevant motion and appearance information. 2) In the decoding phase, we propose a Bilateral Channel Activation (BCA) module to obtain the language-denoised spatial-temporal consistent features for segmenting the referred object. 3) Extensive experiments show that our proposed method outperforms previous methods on four popular RVOS benchmarks, with significant AP gains of 6.8% on A2D Sentences and 6.9% on J-HMDB Sentences, while consuming around 7× less computational overhead.

2. Related Work

2.1. Referring Image Segmentation

The goal of referring image segmentation (RIS) is to segment the corresponding object referred by a natural language expression in a static image. The task is first proposed by Hu *et al*. [10], where visual features extracted by FCN [23] and language features extracted by LSTM [9] are directly concatenated and fused to form the cross-modal features, based on which the segmentation mask of the referred object is predicted. Most recent methods follow this one-stage paradigm and design sophisticated ways of cross-modal interaction involving fine-grained dependency modeling and structural analysis [11, 13, 15, 20, 41]. Moreover, as RIS and referring expression comprehension (*i.e*., predicting the bounding box for the referred object instead of mask) are highly related, MCN [24] proposes a multi-task collaborative network to achieve joint learning of the two tasks. In this paper, we also follow the one-stage paradigm but focus more on realizing effective and efficient spatial-temporal interaction under the mediation of language.

2.2. Referring Video Object Segmentation

Referring video object segmentation (RVOS) can be regarded as an extension of RIS, where both motion and appearance information is required to segment the correct object in a dynamic video. With the availability of diverse benchmarks [6,16,34], RVOS has achieved notable progress recently. Most existing methods mainly feed a video clip centered on the target frame to a 3D ConvNet (*e.g*., I3D [1]), and then obtain mixed spatial-temporal features of the target frame via 3D convolution and pooling. Similar to RIS

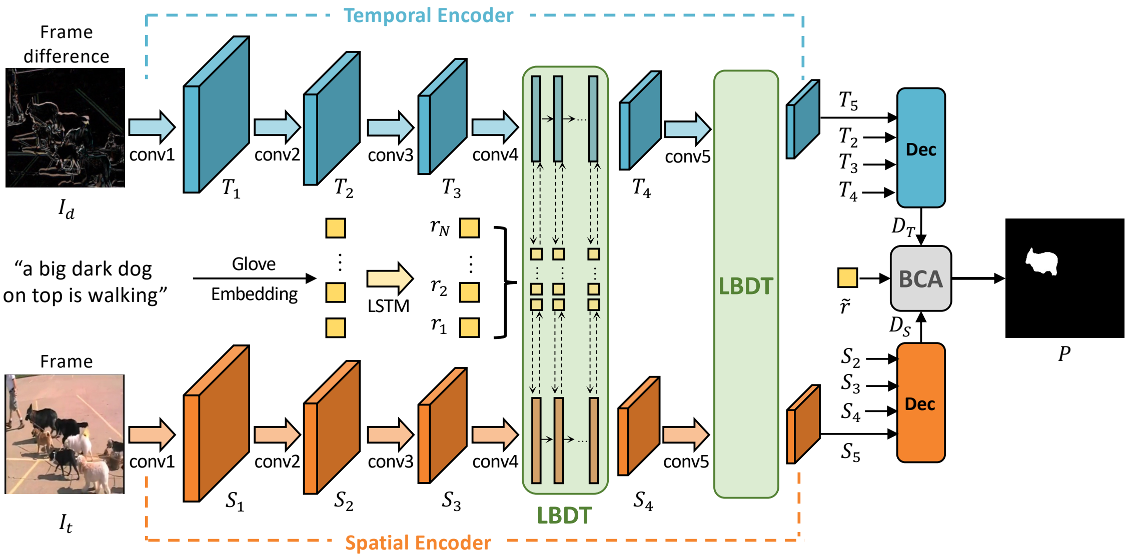

Figure 2. Overview of our proposed method. We feed the target frame I_t and its frame difference I_d into the spatial encoder (bottom) and temporal encoder (top) respectively. And in the LBDT module, we stack several LBDT layers to conduct spatial-temporal interaction with referring words as the medium. In the decoding phase, we denoise the language-irrelevant motion and appearance information and activate the spatial-temporal consistent channels for the decoded spatial features D_S and temporal features D_T respectively in the proposed BCA module. Finally, we apply convolutions and sigmoid function on the outputs of BCA module to get the prediction P.

methods, they focus on designing different mechanisms for better mining the semantic correspondence between video and language features [6, 25, 30, 36, 37, 41], while neglecting the spatial misalignment issue caused by 3D operators. Accordingly, CSTM [12] leverages an additional 2D spatial encoder to complement undisturbed spatial information of the target frame with temporal features in the decoding phase, but still introduce noises with the absence of motion clues in the encoding phase. In this paper, we propose to establish an explicit spatial-temporal interaction earlier in the encoding phase, where language is exploited as a bridge between spatial and temporal encoders to transfer only language-relevant motion and appearance information while suppressing other irrelevant distractors.

2.3. Spatial-Temporal Interaction

Recently, the improvement of spatial-temporal interaction has been widely witnessed in the field of unsupervised video object segmentation [4, 19, 21, 33, 40, 42]. For example, Zhou *et al*. [42] propose a motion-attentive transition to reinforce spatial-temporal object representations with motion information. Ren *et al*. [33] propose a reciprocal transformation network to discover primary objects by correlating both motion and appearance clues. However, these methods only consider spatial-temporal interaction but ignore the vision-language alignment, while the latter is also crucial for RVOS. In this paper, we exploit language as a medium to bridge spatial-temporal interaction for extracting comprehensive multimodal representations.

3. Method

The overall architecture of our model is shown in Figure 2. For the input video clip, we feed its target frame I_t annotated with ground-truth mask and the computed frame difference I_d into two independent ResNet-50 [8] backbones respectively, which are denoted as spatial encoder and temporal encoder in the following. For the input referring expression, we extract language features from the pretrained GloVe embeddings [32] with LSTM [9], which are denoted as $R = \{r_n\}_{n=1}^{N}$ where N is the length of referring expression. To explicitly transfer the language-relevant motion and appearance information between the two encoders, we propose a LBDT module and insert it into different encoder stages. In the decoding phase, we integrate multi-scale contexts and propose a BCA module to denoise the language-irrelevant information and activate the spatial-temporal consistent features via channel-wise activation.

3.1. Visual and Linguistic Feature Extraction

Given a video clip, we feed the target frame $I_t \in \mathbb{R}^{3\times H_0\times W_0}$ and the frame difference $I_d = |I_t - I_{t-\delta}| \in \mathbb{R}^{3\times H_0\times W_0}$ to the spatial encoder and temporal encoder respectively, where δ is the interval between the target frame and the previous frame for calculating the frame difference. Instead of using I3D [1] as the temporal encoder, we build our spatial and temporal encoders upon the 2D ResNet-50 [8]. We denote features of the five stages as $\{S_s\}_{s=1}^{5}, S_s \in \mathbb{R}^{C_s\times H_s\times W_s}$ and $\{T_s\}_{s=1}^{5}, T_s \in$

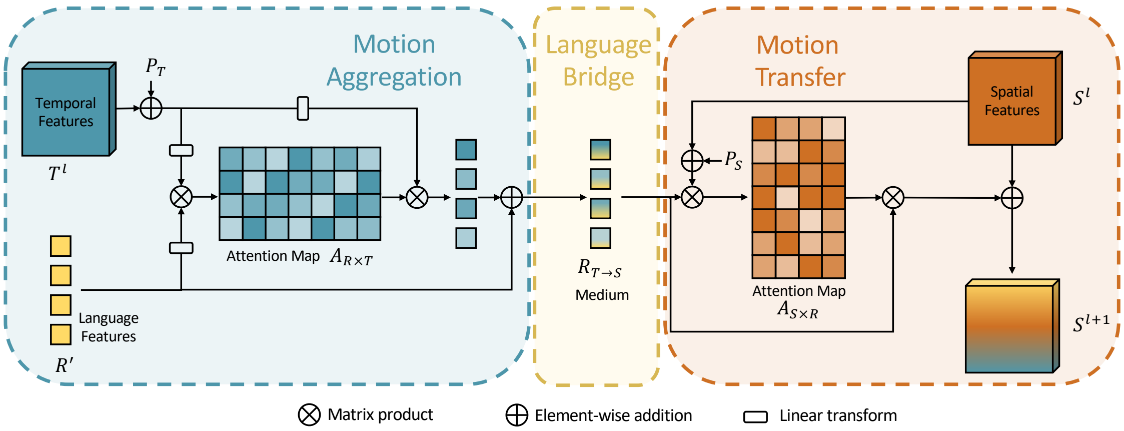

Figure 3. Illustration of the *temporal→language→spatial* information transfer process in our proposed LBDT module. The language-relevant motion information from the temporal features T^l is aggregated into the language medium $R_{T\rightarrow S}$, from which each pixel in the spatial features S^l can select the cross-modal motion information according to the semantic relevance. The *spatial→language→temporal* information transfer process is conducted similarly.

$\mathbb{R}^{C_s\times H_s\times W_s}$ for spatial and temporal encoders respectively, where $H_s, W_s = \frac{H_0}{2^s}, \frac{W_0}{2^s}$ and C_s are the height, width, and channel numbers of features in the s-th stage.

For the referring expression, we embed each word as a 300-dimensional vector [32] and use LSTM [9] as the text encoder to extract the words features $R = \{r_n\}_{n=1}^{N} \in \mathbb{R}^{N\times C_m}$, where N is the max length of the referring expressions and C_m is the channel number.

3.2. Language-Bridged Duplex Transfer

Our LBDT module aims to explicitly transfer the language-relevant motion and appearance information between temporal and spatial encoders with language as the bridge, where we stack L layers of LBDT module to conduct this duplex transfer approach (Figure 3). To clearly elaborate the transfer process in the LBDT module, we take the s-th stage of the encoders as an example and omit the superscript s for simplicity. We change the channel numbers of both spatial features $S \in \mathbb{R}^{C\times H\times W}$ and temporal features $T \in \mathbb{R}^{C\times H\times W}$ to C_m via linear transformation:

$$T^1 = \mathit{Linear}(T), \;\; S^1 = \mathit{Linear}(S), \tag{1}$$

where $S^1 \in \mathbb{R}^{C_m\times H\times W}$ and $T^1 \in \mathbb{R}^{C_m\times H\times W}$ are the visual inputs to the 1-st LBDT layer.

For the linguistic inputs, we first enhance the words features $R \in \mathbb{R}^{N\times C_m}$ by the self-attention mechanism [35] and denote the enhanced words features as $R' \in \mathbb{R}^{N\times C_m}$, which can be formatted as follows:

$$\begin{aligned} R^{Q/K/V} &= \mathit{Linear}(R + P_R), \\ R' &= \mathit{Softmax}(\frac{R^Q (R^K)^T}{\sqrt{C_m}})R^V + R, \end{aligned} \tag{2}$$

where $P_R \in \mathbb{R}^{N\times C_m}$ is the sinusoids positional encoding [35].

Our LBDT module follows the implementation practice of Transformer [35] and revises it to a cross-modal version. In each LBDT layer, we feed the enhanced language features R' and the outputs from the previous layer to it as inputs:

$$S^{l+1}, T^{l+1} = \mathit{LBDT}(S^l, T^l, R'), l = 1, ..., L-1. \tag{3}$$

As the duplex transfer process happens in a symmetrical way, we take the *temporal→language→spatial* transfer process in the l-th LBDT layer as an example (Figure 3). For *motion aggregation*, we first add the 2D sinusoids positional encoding $P_T \in \mathbb{R}^{C_m\times H\times W}$ to the temporal features T^l, and then reshape it to $T^{l'} \in \mathbb{R}^{HW\times C_m}$. We obtain the attention map $A_{R\times T} \in \mathbb{R}^{N\times HW}$ by calculating the similarity between each word and each pixel:

$$T^{l'} = \mathit{Reshape}(T^l + P_T), \tag{4}$$

$$\begin{aligned} R^Q &= \mathit{Linear}(R'), \;\; T^K = \mathit{Linear}(T^{l'}), \\ A_{R\times T} &= \mathit{Softmax}(\frac{R^Q (T^K)^T}{\sqrt{C_m}}), \end{aligned} \tag{5}$$

where $A_{R\times T} = \{A^i_{R\times T}\}_{i=1}^{N}$ and $A^i_{R\times T} \in \mathbb{R}^{HW}$ is the attention map for the i-th word. We use $A_{R\times T}$ to adaptively aggregate the language-relevant motion information from the reshaped temporal features $T^{l'}$, and then add it to the words features R' to obtain the *language medium* $R_{T\rightarrow S} \in \mathbb{R}^{N\times C_m}$ with multimodal representations:

$$\begin{aligned} T^V &= \mathit{Linear}(T^{l'}), \\ R_{T\rightarrow S} &= A_{R\times T} T^V + R'. \end{aligned} \tag{6}$$

For *motion transfer*, we let the spatial features to adaptively select cross-modal motion information from the

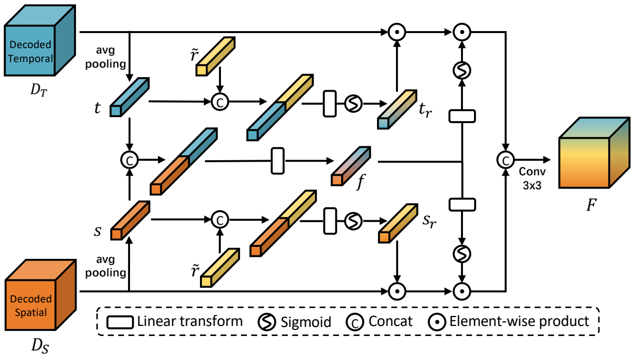

Figure 4. Illustration of BCA module. We filter out the language-irrelevant information with the language denoisers t_r and s_r, and use f_t and f_s to highlight spatial-temporal consistent channels.

medium $R_{T\to S}$. Similarly, we first add the positional encoding P_S to the spatial features S^l, and reshape it to $S^{l'} \in \mathbb{R}^{HW\times C_m}$. Then we calculate the cross-attention map $A^i_{S\times R} \in \mathbb{R}^N$ between the i-th pixel of the spatial features and the medium $R_{T\to S}$, which measures the semantic relevance between these two features:

$$S^{l'} = \mathit{Reshape}(S^l + P_S), \tag{7}$$

$$\begin{aligned} R^K_{T\to S} &= \mathit{Linear}(R_{T\to S}), \quad S^Q = \mathit{Linear}(S^{l'}), \\ A_{S\times R} &= \mathit{Softmax}\Big(\frac{S^Q (R^K_{T\to S})^T}{\sqrt{C_m}}\Big). \end{aligned} \tag{8}$$

Afterwards, we transfer the language-relevant motion information to the spatial features with the cross-attention map $A_{S\times R}$:

$$\begin{aligned} R^V_{T\to S} &= \mathit{Linear}(R_{T\to S}), \\ S^{l+1} &= \mathit{MLP}(A_{S\times R} R^V_{T\to S}) + S^l, \end{aligned} \tag{9}$$

where *MLP* denotes the multi-layer perception and S^{l+1} is the output spatial features of the l-th LBDT layer.

We denote the outputs of the last LBDT layer S^L and T^L as the outputs of our LBDT module. Finally, we increase the channel numbers of S^L and T^L to C and add them with original spatial and temporal features respectively to form a residual connection for easier optimization.

3.3. Bilateral Channel Activation

To obtain strong semantic representation and maintain local details of the frame simultaneously, we upsample the low-resolution spatial and temporal features in the last three stages $\{S_s\}^5_{s=3}$ and $\{T_s\}^5_{s=3}$ to the same size with the 2-nd stage features S_2 and T_2. The resulted features are denoted as $\{S^{up}_s\}^5_{s=3} \in \mathbb{R}^{C_d\times H_2\times W_2}$ and $\{T^{up}_s\}^5_{s=3} \in \mathbb{R}^{C_d\times H_2\times W_2}$ where C_d is the channel number in the decoder. Then, we add them with S_2 and T_2 to get the decoded features $D_S \in \mathbb{R}^{C_d\times H_2\times W_2}$ and $D_T \in \mathbb{R}^{C_d\times H_2\times W_2}$ for spatial and temporal decoders respectively. Given D_S and D_T, we also propose a Bilateral Channel Activation (BCA) module to adaptively filter out language-irrelevant information while highlighting consistent spatial-temporal features, which is illustrated in Figure 4.

Concretely, since D_T and D_S may contain the language-irrelevant motion and appearance information, we propose to exploit the sentence feature $\tilde{r} = \sum_{n=1}^N r_n \in \mathbb{R}^{C_r}$ as the denoiser to filter out the language-irrelevant information. We first conduct average pooling on D_S and D_T to squeeze them into $s \in \mathbb{R}^{C_d}$ and $t \in \mathbb{R}^{C_d}$. Then, we obtain the language-specific spatial denoiser $s_r \in \mathbb{R}^{C_d}$ and temporal denoiser $t_r \in \mathbb{R}^{C_d}$ as follows:

$$s_r = \sigma(\mathit{Linear}([s;\tilde{r}]), \quad t_r = \sigma(\mathit{Linear}([t;\tilde{r}]), \tag{10}$$

where σ is sigmoid function and $[;]$ denotes concatenation.

Meanwhile, we also concatenate s and t on the channel dimension and apply linear transformation to obtain the spatial-temporal consistent feature $f \in \mathbb{R}^{C_d}$:

$$f = \phi(\mathit{Linear}([t;s])), \tag{11}$$

where $\phi(\cdot)$ is ReLU [28] function. We transform f to the channel activators $f_s \in \mathbb{R}^{C_d}$ and $f_t \in \mathbb{R}^{C_d}$ for D_S and D_T respectively using the sigmoid function σ:

$$f_t = \sigma(\mathit{Linear}(f)), \quad f_s = \sigma(\mathit{Linear}(f)). \tag{12}$$

Next, the language-specific denoisers (*i.e.*, t_r and s_r) and spatial-temporal consistent activators (*i.e.*, f_t and f_s) are combined to process decoded spatial features D_S and temporal feature D_T before fusing them:

$$D'_T = f_t \odot t_r \odot D_T, \quad D'_S = f_s \odot t_s \odot D_S, \tag{13}$$

where $\odot$ is the element-wise product with broadcasting operation.

Finally, we concatenate the refined spatial feature D'_S and temporal feature D'_T together and use the 3×3 convolution to obtain the fused feature $F \in \mathbb{R}^{C_d\times H_2\times W_2}$. We further apply convolutions and sigmoid function on F to obtain the logit map and up-sample it to the same spatial size with inputs as the prediction $P \in \mathbb{R}^{1\times H_0\times W_0}$.

4. Experiments

4.1. Datasets and Evaluation Criteria

We evaluate the performance of our method on four popular referring video object segmentation benchmarks: A2D Sentences [6], J-HMDB Sentences [6], Refer-YouTube-VOS [34], and Refer-DAVIS$_{17}$ [16]. For A2D Sentences and J-HMDB Sentences, we use IoU and Precision@X (P@X) as the evaluation criteria following [12, 30, 36].

Method	Pub.			Overlap			AP	IoU	
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
Gavrilyuk *et al.* [6]	CVPR18	47.5	34.7	21.1	8.0	0.2	19.8	53.6	42.1
Gavrilyuk *et al.* † [6]	CVPR18	50.0	37.6	23.1	9.4	0.4	21.5	55.1	42.6
ACGA [37]	ICCV19	55.7	45.9	31.9	16.0	2.0	27.4	60.1	49.0
VT-Capsule [25]	CVPR20	52.6	45.0	34.5	20.7	3.6	30.3	56.8	46.0
CMDy [36]	AAAI20	60.7	52.5	40.5	23.5	4.5	33.3	62.3	53.1
PRPE [30]	IJCAI20	63.4	57.9	48.3	32.2	8.3	38.8	66.1	52.9
CMSA [41]	TPAMI21	48.7	43.1	35.8	23.1	5.2	-	61.8	43.2
CSTM [12]	CVPR21	65.4	58.9	49.7	33.3	9.1	39.9	66.2	56.1
CMPC-V [22]	TPAMI21	65.5	59.2	50.6	34.2	9.8	40.4	65.3	57.3
Ours (LBDT-1)	-	**71.1** (+5.6)	**66.1** (+6.9)	**57.8** (+7.2)	**41.6** (+7.4)	**12.0** (+2.2)	**46.1** (+5.7)	**70.1** (+3.9)	**61.2** (+3.9)
Ours (LBDT-4)	-	**73.0** (+7.5)	**67.4** (+8.2)	**59.0** (+8.4)	**42.1** (+7.9)	**13.2** (+3.4)	**47.2** (+6.8)	**70.4** (+4.2)	**62.1** (+4.8)

Table 1. Comparison with state-of-the-art methods on A2D Sentences testing set. Our method significantly outperforms previous methods relying on the 3D ConvNets for spatial-temporal interaction. "†" denotes utilizing additional optical flow inputs. "LBDT-x" indicates stacking x LBDT layers in each LBDT module.

Method	Pub.			Overlap			AP	IoU	
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
Gavrilyuk *et al.* [6]	CVPR18	69.9	46.0	17.3	1.4	0.0	23.3	54.1	54.2
Gavrilyuk *et al.* ‡ [6]	CVPR18	71.2	51.8	26.4	3.0	0.0	26.7	55.5	57.0
ACGA [37]	ICCV19	75.6	56.4	28.7	3.4	0.0	28.9	57.6	58.4
VT-Capsule [25]	CVPR20	67.7	51.3	28.3	5.1	0.0	26.1	53.5	55.0
CMDy [36]	AAAI20	74.2	58.7	31.6	4.7	0.0	30.1	55.4	57.6
PRPE [30]	IJCAI20	69.0	57.2	31.9	6.0	0.1	29.4	-	-
CMSA [41]	TPAMI21	76.4	62.5	38.9	9.0	0.1	-	62.8	58.1
CSTM [12]	CVPR21	78.3	63.9	37.8	7.6	0.0	33.5	59.8	60.4
CMPC-V [22]	TPAMI21	81.3	65.7	37.1	7.0	0.0	34.2	61.6	61.7
Ours (LBDT-1)	-	**86.4** (+5.1)	**75.1** (+9.4)	**50.7** (+11.8)	**11.6** (+2.6)	0.1	**40.3** (+6.1)	**64.6** (+1.8)	**65.2** (+3.5)
Ours (LBDT-4)	-	**86.4** (+5.1)	**74.4** (+8.7)	**53.3** (+14.4)	**13.2** (+4.2)	0.0	**41.1** (+6.9)	**64.5** (+1.7)	**65.8** (+4.1)

Table 2. Comparison with state-of-the-art methods on J-HMDB Sentences dataset using the best model trained on A2D Sentences. Our method shows notable generalization ability. ‡ denotes training more layers of I3D backbone on A2D Sentences.

Method	Pub.	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
URVOS † [34]	ECCV20	41.34	-	-
URVOS [34]	ECCV20	45.27	49.19	47.23
CMPC-V [22]	TPAMI21	45.64	49.32	47.48
Ours (LBDT-4)	-	**48.18** (+2.54)	**50.57** (+1.25)	**49.38** (+1.90)

Table 3. Comparison with state-of-the-art methods on the Refer-Youtube-VOS validation set. † indicates removing multiple iterations of the second stage inference step.

Method	Pub.	Pretrained	$\mathcal{J}\&\mathcal{F}$	
			1st video	full video
Khoreva *et al.* [16]	ACCV18	RefCOCO [27]	39.30	37.10
URVOS [34]	ECCV20	RefCOCO [27]	44.10	-
URVOS [34]	ECCV20	Refer-Youtube-VOS [34]	51.63	-
Ours (LBDT-4)	-	Refer-Youtube-VOS [34]	**54.08** (+2.45)	**54.52** (+17.42)

Table 4. Comparison with state-of-the-art methods on the Refer-DAVIS$_{17}$ validation set.

Specifically, the overall IoU is the ratio of the total intersection area divided by the total union area over all testing samples, and the mean IoU is the averaged IoU of all testing samples. P@X measures the percentage of test samples whose IoU is higher than a predefined threshold X, where X $\in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. We also compute the average precision (AP) over the interval of $[0.50 : 0.05 : 0.95]$. For Refer-YouTube-VOS and Refer-DAVIS$_{17}$, we use region similarity ($\mathcal{J}$) and contour accuracy ($\mathcal{F}$) following [34].

4.2. Implementation Details

We use ResNet-50 [8] pretrained on ImageNet [18] dataset as our spatial and temporal encoders. For linguistic inputs, we adopt LSTM [9] to extract language features from GloVe word embeddings [32] pretrained on Common Crawl with 840B tokens. The maximum length of the input sentence is 25. We set the frame interval $\delta = 6$ for calculating the frame difference unless otherwise stated. The input frames are resized to 320×320. Adam [17] is utilized as the optimizer. We train the whole network in an end-to-end way with batch size 8 and learning rate $1e^{-4}$ for 15 epochs on NVIDIA Tesla V100 GPUs, which is supervised by cross-entropy loss and dice loss [26]. The learning rate is divided by 2 for every 2 epochs started from the 10-th epoch. As for Refer-DAVIS$_{17}$, we finetune the best model trained on Refer-Youtube-VOS for 1 epoch with a learning rate $1e^{-5}$.

LBDT		BCA		Overlap					AP	IoU	
S→L→T	T→L→S	LD	STC	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
				60.5	54.9	47.8	35.0	11.0	38.8	61.6	54.5
✓				64.5	58.8	49.8	35.9	11.2	40.6	67.3	56.1
	✓			68.0	63.2	54.8	38.4	11.9	43.7	68.6	58.3
✓	✓			70.0	64.1	55.7	39.3	11.5	44.5	69.3	59.8
✓	✓		✓	70.3	65.3	56.7	40.4	12.2	45.3	69.5	60.4
✓	✓	✓		70.2	64.7	56.5	39.9	**12.3**	45.0	69.9	59.9
✓	✓	✓	✓	**71.1**	**66.1**	**57.8**	**41.6**	12.0	**46.1**	**70.1**	**61.2**

Table 5. Verification of the effectiveness of our proposed LBDT module and BCA module. "S→L→T" and "T→L→S" denote *spatial→language→temporal* transfer and *temporal→language→spatial* transfer respectively. "LD" and "STC" denote *language denoiser* and *spatial-temporal consistent activator*.

4.3. Comparison with State-of-the-Art Methods

We compare our method with previous state-of-the-art methods on four popular benchmarks mentioned before. As shown in Table 1, our method outperforms previous works by large margins on the A2D Sentences test set [6]. Compared with CSTM [12], our LBDT-1 model achieves 5.7%, 3.9%, and 3.9% absolute improvements on AP, Overall IoU, and Mean IoU respectively, indicting that using language as the medium to conduct explicit spatial-temporal interaction in the encoding phase is superior to existing methods using 3D ConvNets and implicit interaction in the decoding phase. By stacking LBDT layers, spatial and temporal features can be iteratively optimized, and the best performance is obtained using the LBDT-4 model with 4 layers.

We further verify the generalization ability of our method on J-HMDB Sentences dataset [6]. Following prior works [12, 30, 36], we use the best model trained on A2D Sentences to directly evaluate all the samples in J-HMDB Sentences without finetuning. As shown in Table 2, our method accomplishes significant performance gains over previous state-of-the-arts, indicating that our method can obtain more robust multi-modal representations and generalize the learned knowledge to unseen datasets.

We also conduct experiments on the newly proposed Refer-YouTube-VOS benchmark [34] with richer object categories and denser annotated frames. As shown in Table 3, our method outperforms URVOS [34] and CMPC-V [22] by 2.15% and 1.90% on the $\mathcal{J}\&\mathcal{F}$ metrics respectively, demonstrating that our approach can perform well even in complex scenarios. Moreover, following URVOS [34], we use the best model trained on the Refer-YouTube-VOS and finetune it on the Refer-DAVIS$_{17}$ dataset [16], where we also achieve the best performance as shown in Table 4.

4.4. Ablation Studies

We conduct ablation studies on the A2D Sentences dataset to evaluate the different designs of our model. All experiments are based on our LBDT-1 model.

Component Analysis. We summarize the ablation results of our proposed components in Table 5. The 1-st row is the baseline method, where we first fuse the language features with the visual features and then conduct the duplex spatial-temporal interaction with the cross-attention mechanism [35] directly between spatial and temporal features without the bridging of language, whose computational complexity is $\mathcal{O}((HW)^2C)$. As shown in the 2-nd and 3rd rows, both *motion transfer* and *appearance transfer* can bring notable improvements, validating the effectiveness of the language-bridged duplex strategy. Moreover, the complexity of our LBDT module is $\mathcal{O}(HWNC)$, which is more lightweight as $N \ll HW$. As for the BCA module, we conduct ablation experiments on two key components: *language denoiser* and *spatial-temporal consistency activator*. It shows the decoded spatial and temporal features can benefit each other by activating the spatial-temporal consistent channels (the 5-th row), and reducing the language-irrelevant motion and appearance information can improve the vanilla feature fusion in the decoder (the 6-th row).

Metrics	Interval						
	1	2	3	4	5	6	7
AP	45.0	45.2	45.1	45.6	45.0	**46.1**	45.2
Mean	68.9	68.7	69.4	69.3	69.0	**70.1**	69.0
Overall	60.0	60.5	60.1	60.8	60.0	**61.2**	60.8

Table 6. Interval for calculating the frame difference. "Mean" and "Overall" are Mean IoU and Overall IoU respectively.

Interval for Calculating the Frame Difference. We demonstrate the influence of the interval value δ for calculating the frame difference in Table 6. We found that the best performance is achieved when the interval is 6, which achieves a balance of modeling short and long actions.

Inserting Stages of LBDT Module. We evaluate different inserting positions of LBDT module and summarize the results in Table 7. Inserting LBDT into the 4-th and 5-th stages of our spatial and temporal encoders can bring significant improvements, but the performance decreases as we

Stages				AP	IoU	
2	3	4	5	0.5:0.95	Overall	Mean
			✓	42.3	67.3	57.8
		✓	✓	**46.1**	**70.1**	**61.2**
	✓	✓	✓	45.3	68.2	60.8
✓	✓	✓	✓	38.7	63.7	55.9

Table 7. Inserting stages of LBDT module.

insert it into the earlier stages (*i.e.*, 2-nd and 3-rd stages). It indicates that the language-bridged spatial-temporal interaction is more suitable for transferring the high-level semantic information.

Method	Input Size	GFLOPs	FPS	AP
ACGA [37]	16×512×512	630.83	9.5	27.4
CSTM † [12]	8×320×320	213.06	11.4	39.9
Ours (LBDT-1)	2×320×320	**32.51**	**19.2**	46.1
Ours (LBDT-4)	2×320×320	38.03	12.5	**47.2**

Table 8. Computational overhead. The RGB input size is *frames* × *height* × *width* (channels are omited). † denotes the inference code is obtained by contacting the authors.

4.5. Computational Overhead

We compare the computational overhead of our method and previous ones in Table 8. Without the dependency on 3D ConvNet, our model outperforms existing methods by significant margins while consuming around 7× less GFLOPs and a much smaller input size. Moreover, we evaluate the FPS of these methods on a single NVIDIA 1080Ti GPU. It shows that our method is more efficient, which increases the possibility of practical applications for RVOS.

4.6. Qualitative Analysis

Figure 5 presents the predictions of our method and CSTM [12] in the complex scenes. As the spatial encoder in CSTM lacks motion information, it tends to generate masks on false objects (2-nd column). By explicitly conducting spatial-temporal interaction in the encoding phase with language as the bridge, our methods can obtain the accurate masks of the referred objects (3-rd column). We further visualize the attended regions of the referring words in our LBDT module in Figure 6. Taking the 1-st row as an example, the motion-related word "jumping" attends to the region of the jumping girl, and the appearance-related word "white" and "blue" has the highest responses on the two people in the corresponding colors.

5. Conclusion and Discussion

In this paper, we reconsider the way of spatial-temporal interaction for RVOS and propose a Language-Bridged Duplex Transfer (LBDT) module to explicitly conduct spatial-temporal interaction in the encoding phase with language as the medium for transferring the language-relevant information. A Bilateral Channel Activation (BCA) module is also introduced in the decoding phase to denoise and activate the spatial-temporal consistent features via channel activation. Experiments show that our methods outperform previous methods by large margins on four popular benchmarks with much less computational overhead.

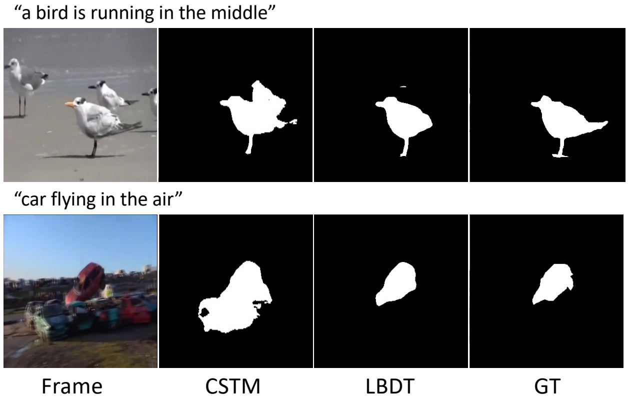

Figure 5. Visualization of the predicted masks of ours and CSTM [12] in the complex scenes.

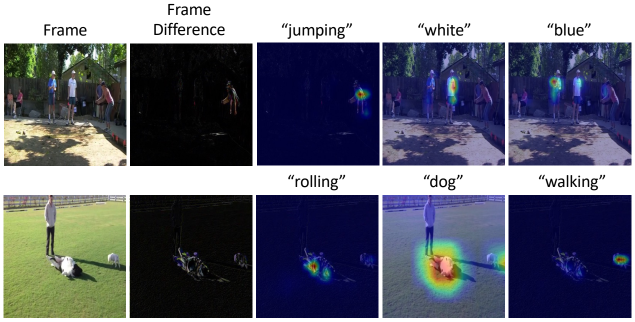

Figure 6. Visualization of attended regions of the referring words.

Limitation. The limitation of our paper is that static language descriptions may not always match the dynamic objects whose location and pose are various in continuous frames. In the future, we hope to address the mentioned mismatch problem by exploring the temporal coherence between the masks in different video frames, which is complementary to our focus in this paper.

Acknowledgments This research is partly supported by National Natural Science Foundation of China (Grant 62122010, 61876177), Fundamental Research Funds for the Central Universities, and Key R&D Program of Zhejiang Province (2022C01082).

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 2, 3

[2] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Multi-fiber networks for video recognition. In *ECCV*, 2018. 2

[3] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *CVPR*, 2021. 1

[4] Katerina Fragkiadaki, Pablo Arbelaez, Panna Felsen, and Jitendra Malik. Learning to segment moving objects in videos. In *CVPR*, 2015. 3

[5] Tsu-Jui Fu, Xin Eric Wang, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. Language-based video editing via multi-modal multi-level transformer. *arXiv preprint arXiv:2104.01122*, 2021. 1

[6] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018. 1, 2, 3, 5, 6, 7

[7] Wenbin Ge, Xiankai Lu, and Jianbing Shen. Video object segmentation using global and instance embedding learning. In *CVPR*, 2021. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 3, 6

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 2, 3, 4, 6

[10] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 2

[11] 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. 2

[12] 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. 1, 2, 3, 5, 6, 7, 8

[13] 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. 2

[14] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 1

[15] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, 2021. 2

[16] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018. 2, 5, 6, 7

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 6

[19] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, 2019. 3

[20] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 2

[21] 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. 3

[22] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE TPAMI*, 2021. 6, 7

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[24] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020. 2

[25] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *CVPR*, 2020. 1, 3, 6

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 6

[27] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 6

[28] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5

[29] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *arXiv preprint arXiv:2107.00650*, 2021. 1

[30] Ke Ning, Lingxi Xie, Fei Wu, and Qi Tian. Polar relative positional encoding for video-language segmentation. In *IJCAI*, 2020. 1, 3, 5, 6, 7

[31] Hyojin Park, Jayeon Yoo, Seohyeong Jeong, Ganesh Venkatesh, and Nojun Kwak. Learning dynamic network using a reuse gate function in semi-supervised video object segmentation. In *CVPR*, 2021. 1

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3, 4, 6

[33] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *CVPR*, 2021. 3

[34] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, 2020. 2, 5, 6, 7

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4, 7

[36] 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. 1, 3, 5, 6, 7

[37] 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. 1, 3, 6, 8

[38] Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *CVPR*, 2021. 1

[39] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1

[40] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *ICCV*, 2019. 3

[41] Linwei Ye, Mrigank Rochan, Zhi Liu, Xiaoqin Zhang, and Yang Wang. Referring segmentation in images and videos with cross-modal self-attention network. *IEEE TPAMI*, 2021. 1, 2, 3, 6

[42] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020. 3