Temporal Context Enhanced Referring Video Object Segmentation

Xiao Hu [1], Basavaraj Hampiholi[2], Heiko Neumann[2], and Jochen Lang✉[1]

[1]University of Ottawa, Canada, {xhu008, jlang}@uottawa.ca
[2]Ulm University, Germany, {basavaraj.hampiholi, heiko.neumann}@uni-ulm.de

Abstract

The goal of Referring Video Object Segmentation is to extract an object from a video clip based on a given expression. While previous methods have utilized the transformer's multi-modal learning capabilities to aggregate information from different modalities, they have mainly focused on spatial information and paid less attention to temporal information. To enhance the learning of temporal information, we propose TCE-RVOS with a novel frame token fusion (FTF) structure and a novel instance query transformer (IQT). Our technical innovations maximize the potential information gain of videos over single images. Our contributions also include a new classification of two widely used validation datasets for investigation of challenging cases. Our experimental results demonstrate that TCE-RVOS effectively captures temporal information and outperforms the previous state-of-the-art methods by increasing the J&F score by 4.0 and 1.9 points using ResNet-50 and VSwin-Tiny as the backbone on Ref-Youtube-VOS, respectively, and +2.0 mAP on A2D-Sentences dataset by using VSwin-Tiny backbone. The code is available at https://github.com/haliphinx/TCE-RVOS

1. Introduction

Video understanding [11, 27] has great potential, as it draws upon visual spatio-temporal information along with audio and language information. Various video-based tasks have been put forward, including but not limited to video classification [3], temporal video action segmentation [21], and video object detection [12]. The transformer structure [10, 36] has shown strong ability in both visual and language understanding, and most importantly, to serve a unified structure for different data formats. All these efforts have led to the task of Referring Video Object Segmentation (RVOS). RVOS is a cross-modal task that takes a video clip with a text expression as input and segments the referred object in all the video frames. Compared with the Referring Image Object Segmentation task (RIOS), the referring expression in RVOS can describe not only an object in space but also motion in the spatio-temporal dimension. Furthermore, RVOS methods also require data association to track the referred object across multiple frames.

Initially, researchers utilized complicated model structures with multi-step training strategies [17, 25, 31]. Recently, benefiting from the transformer structure, various end-to-end learning structures [2, 35, 39] have been published that have achieved state-of-the-art performance on various benchmarks. However, previous methods focused on either text and visual information aggregation or image-based feature learning, but paid less attention to the inter-frame temporal understanding. This leads to a problem that the segmentation performs well on most frames individually but lacks the ability to fuse information across frames. As a consequence, these methods have limitations in handling motion blur and occlusions. To overcome these limitations and fully exploit the inter-frame temporal information in the video, we present a novel approach called Temporal Context Enhanced Referring Video Object Segmentation (TCE-RVOS). Figure 2 shows two sets of comparison results between the proposed method with state-of-the-art methods. The result shows TCE-RVOS outperforms previous state-of-the-art methods by handling some challenging scenarios (e.g., occlusion, and motion) better.

Our main contributions are as follows:

- We designed a frame token fusion (FTF) module as encoder to aggregate features between frames in the video clip using memory tokens. The memory tokens first distill information for each frame independently and then enrich the overall encoding with information from other frames.
- We propose an instance query transformer (IQT) module in the decoding stage to directly aggregate queries about the same object in different frames. This overcomes issues caused by insufficient visual information

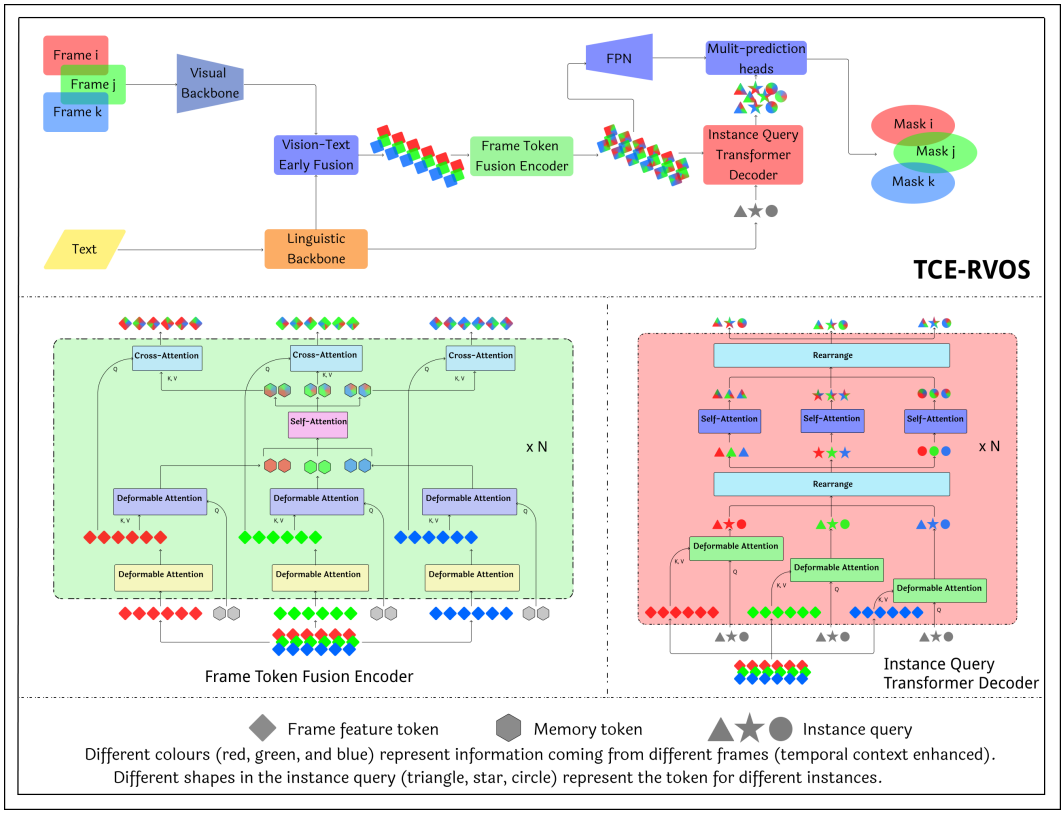


Figure 1. Overview of the TCE-RVOS framework. The sub-graph on the top shows the overall structure. The bottom left and bottom right sub-graphs present the frame token fusion encoder and instance query transformer decoder respectively. The attention blocks in the same level with the same background color share the same weights.

in the current frame due to, e.g., occlusion or motion blur.

- We further organize the Refer-Youtube-RVOS [35] validation set into subcategories including occlusion, motion, crowded, interaction between objects, ambiguous queries, and object presence to investigate how our method improves the SOTA.

Our proposed method improves the *J&F* by a large margin of 4.0 and 1.9 points over the previous SOTA ReferFormer [39] using spatial backbone ResNet-50 [13], and spatio-temporal backbone Video Swin Transformer tiny [27], respectively.

2. Related Work

Video Object Segmentation is commonly solved using two different model types: offline and online models. The offline model solves all frames at once [14, 18, 38], while the online model segments the object in the first frame then propagates to the rest [32, 40, 41]. The offline models usually demand large memory storage which leads to computational overhead, since the segmentation of objects among all conjugative frames at once requires large temporal receptive field. Offline methods are also weak in data association. VisTR [38] builds upon the image-based object segmentation structure DETR [4] by processing the object queries from all the frames together. IFC [18] extends the image-based object segmentation Mask2Former [6] into video-based segmentation and utilizes a token-based inter-frame communication to overcome the data association problem. On top of IFC, VITA [14] fully tokenizes the frame features to distill the visual information. The online model IDOL [40] shows that the inter-frame data association is the bottleneck in current video instance segmentation tasks. IDOL utilizes contrastive learning to enhance the data association across frames.

Referring Video Object Segmentation is a relatively new task that was first introduced by Gavrilyuk *et al.* [12] in 2018 to segment the actors and actions in video clips. Since RVOS involves video object segmentation, neural language processing, and cross-modal learning, early research [17, 24, 25, 31] typically combined models from different tasks, resulting in complex structures that are difficult to train end-to-end. A straightforward approach is to extend image-based methods [7, 16, 19, 29] to process each frame

of the video clip separately. However, this approach neglects temporal information. To incorporate temporal information, spatio-temporal backbones such as I3D [5], and VSwin Transformer [27] have been used. Nonetheless, the processing after the backbone still treats each frame independently. URVOS [35] splits the task into an image-based referring object segmentation and a mask propagation task. Recurrent neural networks and memory mechanisms are used in [17, 25] to provide an online strategy. Another direction is to integrate linguistic features into the video object segmentation task [38, 42]. MTTR [2], and ReferFormer [39] are two state-of-the-art methods that borrow ideas from a VOS structure called VisTR [38]. However, cross-frame information exchange only occurs in the backbone stage (if a spatio-temporal backbone is used), and the instance sequence matching at the final stage. The encoder and decoder stages still process each frame independently. Some researches stipulated that the imbalance between the language pipeline and the vision pipeline would affect the model performance. VLT [8] generates a number of language features for a single sentence to close the gap between the two types of feature. R^2-VOS [22] first enriched the original dataset with mismatched video-text pairs, then proposed a contrastive learning structure to filter out the mismatched pairs to help the model understand the language feature better.

3. Method

After analyzing the framework of previous works [2, 39], we found that the temporal context aggregation only happens during the feature extraction when using a spatio-temporal backbone like VSwin Transformer, and in the post processing stage. However, the encoder and decoder stages handle each frame independently. TCE-RVOS utilizes a newly designed frame token fusion encoder and an instance query transformer decoder to enhance communication across video frames. The model contains four main stages overall: (1) Backbone and early fusion (Sec 3.1), (2) Frame Token Fusion encoder (Sec 3.2), (3) Instance Query Transformer Decoder (Sec 3.3), and (4) Post-processing and Prediction (Sec 3.4). The TCE-RVOS model structure is presented in Figure 1. Frames from the video clip and the text expression are passed through the visual backbone and linguistic backbone separately. Once the visual and linguistic features are fused by the vision-text early fusion block, the aggregated features are sent into the frame token fusion encoder to further process the information as well as communicate between frames. The temporally enhanced features are used to guide the instance queries to distill the instance-related features in the instance query transformer decoder. Finally, each instance query is used to predict the instance mask, bounding box, and reference score.

Given a video clip $V = \{I_i\}_{i=1}^{T}, I_i \in \mathbb{R}^{D\times H\times W}$, with T frames and a text expression $E = \{t_i\}_{i=1}^{L}$ with L words. D represents the number of frame channels, H and W represent the height and width of the frame. The proposed model will predict T frames of binary segmentation masks of the referred object, $M = \{m_i\}_{i=1}^{T}, m_i \in \mathbb{R}^{H\times W}$.

3.1. Backbone & Early Fusion

Visual Backbone. Since the proposed model can easily adapt to different backbone structures, both image based backbone ResNet [13], and video based backbone Video Swin Transformer [27] are tested in our model. The output for a given video clip is $F_{vis} = \{v_i\}_{i=1}^{T}, v_i \in \mathbb{R}^{S\times C}, S = \sum_l H_l \times W_l$. C is the channel size of the feature. l represents the set of feature layers from the backbone.

Linguistic Backbone. We follow ReferFormer [39] and use RoBERTa [26] as a backbone to extract features of the text expression. RoBERTa returns a set of the word based features corresponding to each word separately, $F_{word} = \{f_i\}_{i=1}^{L}, f_i \in \mathbb{R}^{C}$, and a sentence based feature $F_{sentence} \in \mathbb{R}^{C}$.

Vision-Text Early Fusion. This block is a vanilla multi-head cross-attention model. The visual feature F_{vis} is added to a fixed 3D positional encoding as the $query$. The key and $value$ are based on F_{word} . This stage allows the resultant features to carry both vision and language context.

3.2. Frame Token Fusion Encoder

Previous works [2, 39] encode frames in a batch by treating each frame individually during encoding. The features between different frames are not aggregated and thus the temporal context is not well encoded. Inspired by IFC [18], a video instance segmentation framework, we have designed an efficient yet simple structure as shown in the bottom left of Figure 1 to enhance the temporal context aggregation between frames in the encoding stage. Compared with the inter-frame communication used in IFC, which directly concatenates memory tokens to the end of each feature token query and performs self-attention, the proposed structure not only saves memory but also incorporates multi-scale features that benefit the visual understanding.

The deformable self-attention model [46] is used to reduce the overall memory usage and support multi-scale feature processing. The features from each frame are first processed by an identical deformable attention block to independently aggregate the spatial information in each frame (shown in the yellow blocks in Figure 1). A set of randomly initialized trainable memory tokens, denoted by $F_{memory} = \{j_i\}_{i=1}^{T}, j_i \in \mathbb{R}^{N\times C}$ are utilized to distill and represent information for each frame. The hyperparameter N determines the number of memory tokens used for each frame. This step also uses the deformable cross-attention model, F_{memory} as $query$, and F_{vis} as key and $value$, so that the memory tokens obtain multi-scale information from

the frame features (shown as the dark blue blocks in Figure 1). The resulting memory tokens, which now contain information from all frames independently, are forwarded as input to a self-attention block (shown as the pink block in Figure 1) for communication between frames. Finally, we update frame feature tokens F_{vis} corresponding to the memory tokens using a naive cross-attention block. Each attention block is followed by a simple FFN. The output features are now carrying the information from both the corresponding frame, and other frames.

3.3. Instance Query Transformer Decoder

We design the Instance Query Transformer Decoder block to enhance temporal information learning by operating multi-head attention between the features of an object from all frames.

The proposed decoder follows a DETR-like structure that initializes a number of instance queries, $F_{instance} = \{h_i\}_{i=1}^{T}, h_i \in \mathbb{R}^{Q\times C}$, to generate Q instance candidates for each frame, guided by the frame feature tokens output from the encoder. However, different from previous models, the newly designed decoder establishes a two step feature aggregation structure to enhance the temporal context learning. The instance query aggregates the spatial features from each frame independently at first, then temporal features from all frames belonging to the same instance are fused. The instance queries are initialized by the sentence feature $F_{sentence}$ and a fixed 2D positional embedding. Then deformable cross-attention is used to extract information from the frame feature tokens output from the encoder (shown as the green blocks in Figure 1). The instance queries from different frames are rearranged and combined into groups based on the corresponding instance (shown as the light blue blocks in Figure 1), and processed by a self-attention block to communicate between them (shown as the dark blue blocks in Figure 1), which provides the temporal context aggregation. The functionality of the rearrange block can be described as $Rearrange(\mathbb{R}^{T\times Q\times C}) = \mathbb{R}^{Q\times T\times C}$. This strategy enables instances from different frames with different viewpoints to benefit the instance segmentation in the current frame, especially when the referred instance is occluded or is motion blurred in some frames. Finally, the Instance Query Transformer Decoder block outputs the processed instance queries for the multi-task prediction.

3.4. Post-processing and Prediction

The post-processing stage in the proposed method is similar to that used in MTTR and ReferFormer. However, we found that the cross-modal feature pyramid network (CM-FPN) used in ReferFormer takes a large amount of memory but accuracy improvements are limited. We use instead a classic Feature Pyramid Network (FPN) reducing the model size but with the same overall performance. Three outputs are predicted for each instance query sequence: The bounding box, denoted as $P_{bbox} \in \mathbb{R}^{4\times T\times Q}$; the reference score, denoted as $P_{ref} \in \mathbb{R}^{T\times Q}$, which indicates the confidence of each instance query being referred to by the sentence, and the segmentation mask, denoted as $P_{mask} \in \mathbb{R}^{H\times W\times T\times Q}$ for each instance query in each frame. The instance sequence with the highest overall P_{ref} score is selected as the final output prediction for the video.

3.5. Loss Functions

The overall loss consists of three terms as follows

$$\mathcal{L}_{overall} = \lambda_{bbox}\mathcal{L}_{bbox} + \lambda_{ref}\mathcal{L}_{ref} + \lambda_{mask}\mathcal{L}_{mask}. \quad (1)$$

The bounding box prediction loss $\mathcal{L}_{bbox}$ is composed of the L1 loss and the generalized IoU (GIoU) loss [34]. $\mathcal{L}_{ref}$ is a focal loss to supervise whether the prediction is the referred instance. $\mathcal{L}_{mask}$ consists of the DICE loss and per-pixel focal loss [30]. $\lambda_k, k \in \{bbox, ref, mask\}$, represents the corresponding coefficients to balance different losses (see Sec. 4.1 for their settings).

4. Experimental Evaluation

4.1. Experimental Setup

Datasets. The proposed method was evaluated on Ref-Youtube-VOS [35], A2D-Sentences [12] and Ref-DAVIS17 [20] datasets to compare the performance with state-of-the-art methods. The Ref-Youtube-VOS dataset is a large-scale benchmark that contains 3978 high-resolution videos from YouTube, with 15K language expressions and 131K high-quality manual annotations. The dataset is split into 3471 training videos, 202 validation videos and 305 test videos. Since the test video set is not publicly available, all models were evaluated on the validation set for fair comparison. Ref-DAVIS17 is built from DAVIS17 [33] by adding text expressions to the VOS dataset. Although the dataset only contains 90 videos, it is still widely used in the R-VOS task. A2D-Sentences dataset extends the original A2D video object segmentation dataset with text expressions. A2D-Sentences contains 3,782 video samples with 3-5 frames annotations per sample.

Evaluation Matrics. The standard evaluation metrics for Ref-Youtube-VOS and Ref-DAVIS17 are Jaccard index (J) for region similarity, contour accuracy $F1$ score (F), and their average ($J\&F$). The (J) score focuses on the overall segmentation quality, and the (F) score focuses more on the segmentation details (boundary accuracy).

To evaluate the proposed method and compare with previous works on A2D-Sentences dataset, we adopt precision@K ($K \in [0.5, 0.6, 0.7, 0.8, 0.9]$), overall & mean IoU, and mean average precision (mAP) over 0.50:0.05:0.95.

Implementation Details. Various backbones were evaluated in our model. The outputs from the last three layers of

the backbone are used to generate multi-scale features with spatial down-sampling rates of $\{8, 16, 32\}$ respectively. The number of memory tokens used in the Frame Token Fusion Encoder is $N = 8$. All the experiments were conducted using 4 Nvidia V100 GPUs with 32GB memory.

In order to compare with state-of-the-art methods, we use a similar hyperparameter settings than [39]. In particular, both the encoder and decoder are 4 layers. The dimension for both visual and linguistic features is $C = 256$. Batch size is set to 1. The number of instance queries for each frame in the Instance Query Transformer Decoder is $Q = 5$. We use AdamW [28] with an initial learning rate $1e-4$ as the optimizer. The coefficients for the loss function are $\lambda_{bbox} = 2$, $\lambda_{ref} = 2$, and $\lambda_{mask} = 5$. By default, the input video frame number for training is $T = 5$, and the model is pretrained on the image based referring object segmentation dataset Ref-COCO [45].

4.2. Comparison Results

Table 1 presents a comparison between the proposed TCE-RVOS and state-of-the-art methods on Ref-Youtube-VOS dataset. The upper half of the table shows the comparison of methods using a spatial backbone. TCE-RVOS outperforms all other methods with a ResNet-50 backbone, with an improvement of at least 4.0 points in the *J&F* index. Moreover, our model performs better than previous models with the larger backbone ResNet-101 by at least 2.3 points. By using ResNet-101, our method achieves state-of-the-art. The bottom half of the table shows the comparison for all models using a spatio-temporal backbone. TCE-RVOS outperforms other models using the same VSwin-Tiny backbone with at least 1.9 points improvement, and even outperforms a model using the large-scale backbone VSwin-Small with a 1.2 points increase. Figure 2 shows two prediction examples. In each sub-figure, the top, middle, and bottom sequences represent the result from MTTR, ReferFormer, and TCE-RVOS, respectively. Two challenging scenarios are selected as the examples. In Figure 2a, the person is partially occluded by the window in some of the video frames, but never fully occluded. In Figure 2b, the parachute is fully occluded by a person in the front in several frames at the start. The result shows that our proposed method increases the capability of handling occluded instance segmentation to predict more accurate masks, and has the ability to verify if the referred instance is visible in the frame by using the information from other frames.

Table 2 shows a comparison result on Ref-Davis17 dataset for various methods by using the ResNet-50 backbone. Since Ref-Davis17 is a small dataset with only 90 videos in total, we directly evaluate the model trained on Ref-Youtube-VOS dataset on Ref-Davis17 without finetuning. The result shows that our model has a good generalization which achieves similar accuracy on both dataset,

Method	Backbone	*J&F*	*J*	*F*
Spatial Backbone				
CMSA [43]	ResNet-50	34.9	33.3	36.5
CMSA+RNN [43]	ResNet-50	36.4	34.8	38.1
URVOS [35]	ResNet-50	47.2	45.3	49.2
PMINet [9]	ResNet-101	48.2	46.7	49.6
PMINet + CFBI [9]	ResNet-101	53.0	51.5	54.5
ReferFormer [39]	ResNet-50	55.6	54.8	56.5
ReferFormer [39]	ResNet-101	57.3	56.1	58.4
TCE-RVOS (ours)	ResNet-50	<u>59.6</u>	<u>58.3</u>	<u>60.8</u>
TCE-RVOS (ours)	ResNet-101	**60.8**	**59.4**	**62.2**
Spatio-Temporal Backbone				
MTTR [2]	VSwin-Tiny	55.3	54.0	56.6
ReferFormer [39]	VSwin-Tiny	59.4	58.0	60.9
ReferFormer [39]	VSwin-Small	<u>60.1</u>	<u>58.6</u>	<u>61.6</u>
TCE-RVOS (ours)	VSwin-Tiny	**61.3**	**59.8**	**62.7**

Table 1. Comparison with state-of-the-art methods on Ref-Youtube-VOS [35]. The top portion shows models with spatial backbone, and the bottom portion shows models with spatio-temporal backbone. The best results are in bold, and the second best results are underlined.

Method	*J&F*	*J*	*F*
CMSA [43]	34.7	32.2	37.2
CMSA+RNN [43]	40.2	36.9	43.5
URVOS [35]	51.5	47.3	56.0
ReferFormer [39]	<u>58.5</u>	<u>55.8</u>	<u>61.3</u>
TCE-RVOS (ours)	**59.4**	**56.5**	**62.4**

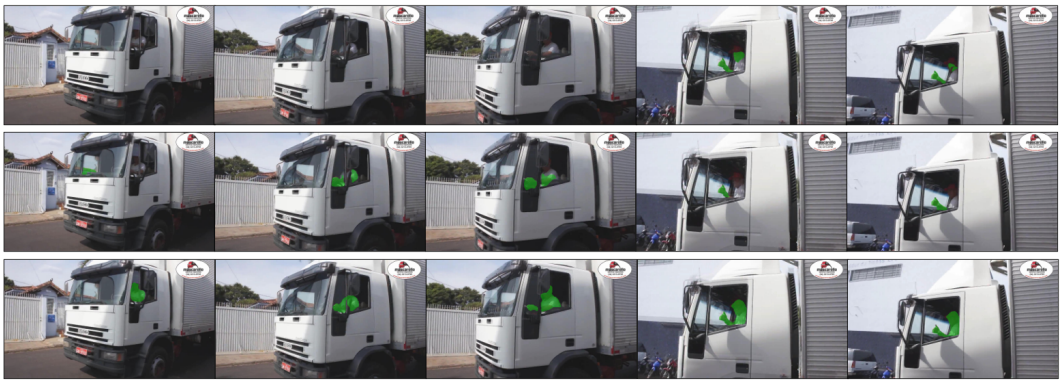
Table 2. Comparison with state-of-the-art methods by using ResNet-50 as a backbone on Ref-Davis17 [20]. The results for ReferFormer and TCE-RVOS are obtained with a model trained on Ref-Youtube-VOS datset without finetuneing.

and outperforms all the competitors.

Table 3 shows the comparison result of TCE-RVOS with previous methods. Only using the spatial backbone (ResNet-50), the proposed model already outperforms most of the previous methods. When a spatio-temporal backbone (VSwin-Tiny) is used, TCE-RVOS outperforms all the previous methods using the backbone no larger than VSwin-Tiny with at least 2.0 mAP improvement, and also outperforms the model using a larger backbone VSwin-Small with 0.9 mAP increment. When the large spatio-temporal backbone VSwin-Base is used, our proposed method outperforms all others, and achieves the state-of-the-art.

4.3. Ablation Study

To fully investigate where and how our model improves the previous state-of-the-art method, some ablation studies



(a) The prediction result for the expression "a person wearing a white shirt is driving a white truck moving down the road", shown in green masks.

(b) The prediction result for the expression "a white and red parachute blowing in the wind", shown in blue masks.

Figure 2. Comparison between qualitative result of MTTR (top sequence), ReferFormer (middle sequence), and TCE-RVOS (bottom sequence) from Ref-Youtube-RVOS dataset [35]. (a) Partial occlusion, and (b) Complete occlusion.

are made including a model component analysis, different scenarios observations, and a study of the impact of temporal window size. All the experiments in this section are using ResNet-50 as backbone and are trained on Ref-Youtube-VOS datset if nothing else is specified.

Model Components. In order to show that the designed Frame Token Fusion Encoder (Sec. 3.2) and Instance Query Transformer Decoder (Sec. 3.3) benefit the model performance, ReferFormer is selected as the baseline model since it is the previous state-of-the-art model and shares similar overall structure with TCE-RVOS. The comparison is done by changing the encoder and decoder. Table 4 shows the ablation study result for the model components. Both Frame token Fusion Encoder and Instance Query Transformer Decoder benefit the model performance. Notably the Frame Token Encoder improves the contour accuracy (F) score more than the region similarity (J) score. The state-of-the-art method ReferFormer performs well on the overall segmentation quality in common scenarios. However, some challenging scenarios (i.e., occlusion, and motion blur) will change the shape of the object and decrease the boundary prediction accuracy (F). By adding our newly designed structures to enhance the temporal context understanding in the network, the segmentation information from other frames will benefit the prediction in challenging frames.

Testcase Scenarios. In RVOS datasets, the text expression describes a single instance but the instance may not be in view in all the frames of a video. It may also be in the field-of-view but occluded by other objects in the scene. The instance segmentation is of varying difficulty because the instance may be stationary or moving relative to the camera, or because there may be many different instances of the same class of objects visible in the video. The different text expressions are also of varying quality and may describe the instance in absolute terms or only relative to other scene objects. Sometimes even multiple objects fit a

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
Hu *et al.* [15]	VGG-16	34.8	23.6	13.3	3.3	0.1	47.4	35.0	13.2
Gavrilyuk *et al.* [12]	I3D	47.5	34.7	21.1	8.0	0.2	53.6	42.1	19.8
CMSA+CFSA [44]	ResNet-101	48.7	43.1	35.8	23.1	5.2	61.8	43.2	-
ACAN [37]	I3D	55.7	45.9	31.9	16.0	2.0	60.1	49.0	27.4
RefVOS [1]	ResNet-101	57.8	-	-	-	9.3	67.2	49.7	-
CSTM [17]	I3D	65.4	58.9	49.7	33.3	9.1	66.2	56.1	39.9
CMPC-V [25]	I3D	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
ClawCraneNet [23]	ResNet-50/101	70.4	67.7	61.7	48.9	17.1	63.1	59.9	-
MTTR(w=8) [2]	VSwin-Tiny	72.1	68.4	60.7	45.6	16.4	70.2	61.8	44.7
MTTR(w=10) [2]	VSwin-Tiny	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
TCE-RVOS (ours)	ResNet-50	80.3	77.1	70.1	53.3	18.2	75.6	67.5	51.2
ReferFormer [39]	VSwin-Tiny	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
ReferFormer [39]	VSwin-Small	82.6	79.4	73.1	57.4	21.1	77.7	69.8	53.9
TCE-RVOS (ours)	VSwin-Tiny	83.0	79.9	73.6	56.7	20.5	77.5	69.9	54.8
ReferFormer [39]	VSwin-Base	<u>83.1</u>	<u>80.4</u>	<u>74.1</u>	<u>57.9</u>	<u>21.2</u>	**78.6**	<u>70.3</u>	<u>55.0</u>
TCE-RVOS (ours)	VSwin-Base	**83.3**	**80.6**	**74.6**	**58.6**	**22.2**	<u>78.4</u>	**70.5**	**56.0**

Table 3. Comparison with state-of-the-art Methods on the A2D Dataset [12]. The best results are in bold, and the second best results are underlined.

	J&F	*J*	*F*
ReferFormer	55.6	54.8	56.5
+FTF Encoder	56.3 (+0.7)	54.9 (+0.1)	57.6 (+1.1)
+IQT Decoder	58.1 (+2.5)	57.1 (+2.3)	59.0 (+2.5)
TCE-RVOS	59.6 (+4.0)	58.3 (+3.5)	60.8 (+4.3)

Table 4. Ablation study. Both, our frame token fusion (FTF) encoder and instance query transformer (IQT) decoder benefit the model. ReferFormer [39] is the baseline of our method.

given text expression. Therefore, we classify the validation set based on the testcase scenarios according to the following categories:

1. Occlusion: We categorize videos into no occlusion, partial occlusion, and full occlusion. If the referred instance is fully occluded in any of the frames of the video clip, we classify the clip as fully occluded. A video is classified as showing partial occlusion if the referred instance is partially occluded in at least one frame but never fully occluded in any frame.

2. Presence: We categorize videos into partial presence and full presence of the referred instance. Presence is lost if the referred instance completely exits or hasn't entered the field-of-view in one of the video frames, which is different from occlusion.

3. Object Motion: We categorize object motion based on the differences of the center and size (height, width) of the bounding box in the video frames. We use categories of no, slow and fast motion of the referred instance.

4. Crowded: We categorize a video to show crowding if there are multiple instances of objects with the same class than the referred instance (e.g., referring to one person in a group of people). The category is split into crowded and not crowded.

5. Interaction: We categorize videos whether the text expression describes the instance by an attribute related to other objects in the frame (e.g., the person near a tree). This category is divided into no interaction and interaction.

6. Ambiguity: We define a sample as ambiguous if there are multiple instances in the video clip which satisfy the text expression. The samples are divided into ambiguous and unambiguous.

More detailed explanations and some examples can be found in the supplemental material. Above categories can be clustered into two groups for different evaluations. (1) Temporal relationship: occlusion, presence, and motion can benefit from enhancing the temporal information because the poor visibility of the instance is limited to part of the frames. The segmented object from other frames can guide the prediction in frames with poor visibility. (2) Vision-text aggregation: crowded, interaction, and ambiguity are closely related to the textual description of the referred to instance but can benefit little from temporal relationships.

	Occlusion			Presence		Crowded		Interaction		Ambiguity		Motion		
	No	Partial	Full	Full	Partial	No	Yes	No	Yes	No	Yes	No	Slow	Fast
Samples	325	453	56	688	156	402	432	541	293	718	111	203	275	356
ReferFormer	63.3	52.9	33.2	60.0	34.9	63.7	48.2	61.9	44.1	59.3	31.9	51.9	62.8	52.2
+FTF Encoder	63.7	53.5	35.8	59.7	40.2	65.0	48.1	63.6	42.8	59.3	36.8	51.6	64.1	52.9
+IQT Decoder	64.6	55.9	37.9	61.5	41.8	66.0	50.7	64.6	45.9	61.6	35.3	**56.8**	63.3	54.7
TCE-RVOS	**65.9**	**57.4**	**40.1**	**62.6**	**45.3**	**66.1**	**53.4**	**65.5**	**48.6**	**62.6**	**40.5**	56.7	**67.0**	**55.5**

Table 5. Ablation study for testcase scenarios. Results are obtained with the ResNet-50 backbone on Ref-Youtube-RVOS [35]. The best results are in bold, and the second best results are underlined. Our frame token fusion (FTF) encoder and instance query transformer (IQT) decoder are most effective in categories where temporal information is relevant. ReferFormer [39] is the baseline of our method.

No. of frames	*J&F*	*J*	*F*
T=3	57.7	56.5	58.9
T=4	58.7	57.3	60.0
T=5	**59.6**	**58.3**	**60.8**

Table 6. Ablation study for the impact of the temporal windows size. We use Resnet-50 as backbone on Ref-Youtube-RVOS [35]. The accuracy increases as the number of input frames are increased.

Table 5 shows the results for the testcase scenarios study. Since the ground truth for the Ref-Youtube-VOS validation set is not accessible, the motion status is classified empirically. We also classified the A2D validation set by two different thresholds to study the influence of thresholds on results for different motion categories (see the supplemental material for details). For temporal related categories, the improvement depends on sub-classes. For example in Table 5, comparing with the baseline model under the occlusion category, the improvements are 2.6, 4.5, and 6.9 points in the *J&F* index for no occlusion, partial occlusion, and full occlusion, respectively. TCE-RVOS improves the accuracy more for the occluded than the non-occluded instances. In the vision-text aggregation categories, the improvements are more equal with improvements for no interaction of 3.6 points and interaction of 4.5 points in the *J&F* index, respectively. This observation shows that our proposed novel encoder and decoder structures improve the model performance by handling the temporal related challenges well. When looking at the vision-text aggregation, similar improvements across sub-classes can be observed. This is expected as our model does not make any improvements specific to the vision-text aggregation structure.

Length of Temporal Window. We conduct an experiment to understand the impact of the temporal window size on the final performance. In Table 6, we observe that the performance of the model improves as the number of input frames are increased. These results imply that our proposed method is able to capture temporal context and thereby contributing to the enhancement of model performance.

Above experiments show the improvement of TCE-RVOS over state-of-the-art models by not only increasing the overall accuracy, but also handling challenging scenarios better. Both, the novel proposed encoder and decoder benefit model performance by enhancing the temporal context through all video frames. TCE-RVOS shares a similar model size with the ReferFormer model (e.g., 178M vs. 177M parameters when using the VSwin-Base backbone).

5. Conclusion

We proposed TCE-RVOS, an end-to-end referring video object segmentation approach. We first analyzed that the temporal context in previous work is weak. The data association between the same object in different frames limits the model performance in challenging scenarios. To overcome this weakness, our method enhances the temporal context learning in the model by a novel frame token fusion encoder with an instance query transformer decoder, and achieves state-of-the-art results with a clear margin (4.0 points in the *J&F* index on Ref-Youtube-VOS using ResNet50). We classified the Ref-Youtube-VOS validation dataset and A2D Dataset to investigate the performance in challenging scenarios. The ablation study shows that our improved structure achieves the goal of better handling the temporal context in challenging scenarios. However, there are still some limitations that can be directions for future works. The model as its baseline ReferFormer is not suitable for real time processing as it takes around 5 sec. for inference on a 10 frames video. There is also room in improving the vision-text aggregation including the handling of ambiguous text description.

6. Acknowledgements

Xiao Hu and Jochen Lang acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). The work was conducted while Jochen Lang was on academic leave and stay at the Institute of Neural Information Processing, UUlm.

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. 7

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 1, 3, 5, 7

[3] Darin Brezeale and Diane J Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3):416–430, 2008. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2

[7] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 2

[8] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[9] Zihan Ding, Tianrui Hui, Shaofei Huang, Si Liu, Xuan Luo, Junshi Huang, and Xiaoming Wei. Progressive multimodal interaction network for referring video object segmentation. *The 3rd Large-scale Video Object Segmentation Challenge*, page 7, 2021. 5

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[11] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 1

[12] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018. 1, 2, 4, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[14] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *arXiv preprint arXiv:2206.04403*, 2022. 2

[15] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 7

[16] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10488–10497, 2020. 2

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2021. 1, 2, 3, 7

[18] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 2, 3

[19] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[20] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 4, 5

[21] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6742–6751, 2018. 1

[22] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Yan Lu, and Bhiksha Raj. Rˆ 2vos: Robust referring video object segmentation via relational multimodal cycle consistency. *arXiv preprint arXiv:2207.01203*, 2022. 3

[23] Chen Liang, Yu Wu, Yawei Luo, and Yi Yang. Clawcranenet: Leveraging object-level relation for text-based video segmentation. *arXiv preprint arXiv:2103.10702*, 2021. 7

[24] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 2

[25] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension

for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2021. 1, 2, 3, 7

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[27] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 1, 2, 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] 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 *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 2

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4

[31] Ke Ning, Lingxi Xie, Fei Wu, and Qi Tian. Polar relative positional encoding for video-language segmentation. In *IJCAI*, volume 9, page 10, 2020. 1, 2

[32] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 2

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

[34] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[35] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 1, 2, 3, 4, 5, 6, 8

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3939–3948, 2019. 7

[38] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 2, 3

[39] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 2, 3, 5, 7, 8

[40] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 588–605. Springer, 2022. 2

[41] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V*, pages 332–348. Springer, 2020. 2

[42] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. Video object segmentation and tracking: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4):1–47, 2020. 3

[43] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10502–10511, 2019. 5

[44] Linwei Ye, Mrigank Rochan, Zhi Liu, Xiaoqin Zhang, and Yang Wang. Referring segmentation in images and videos with cross-modal self-attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3719–3732, 2021. 7

[45] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5

[46] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3