Visual-Language Transformer for Referring Video Object Segmentation

E-Ro Nguyen[1,3], Nhat Hoang-Xuan[1,3], Minh-Triet Tran[1,2,3]
[1]University of Science, VNU-HCM, [2]John von Neumann Institute, VNU-HCM
[3]Vietnam National University, Ho Chi Minh city, Vietnam
{nero, hxnhat}@selab.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn

Abstract

Referring Video Object Segmentation task (R-VOS) aims to segment the target object in all video frames referred with a language expression. In this work, we present a Visual-Language Transformer (VLFormer), a query-based network to tackle the R-VOS task. Its key component is the visual-language transformer block (VLB), which associates visual and linguistic features with the object queries effectively and simultaneously. We use the object queries as a set of candidate objects, the Transformer decoder with VLB blocks that guide and interact with candidate objects to find the referred target object. The object tracking is achieved automatically by linking the corresponding queries across frames. Afterwards, a post-processing technique is used to refine and re-track the mask prediction among all the frames. Our model ranks 6^{th} place on Track 3: Referring Video Object Segmentation of the CVPR 2022 Youtube-VOS Challenge.

1. Introduction

Referring video object segmentation(RVOS) adopts a referring expression to segment the target object in the entire video. RVOS is a challenging task since it requires a comprehensive understanding of the semantics of multiple modalities without any ground-truth mask as the traditional semi-supervised video object segmentation.

Existing RVOS models can be primarily categorized into three types. First, the bottom-up methods conduct the interaction between visual and linguistic features for highlighting the visual features that are matched with the corresponding linguistic clues in the visual encoder phase and then adopt a FCN [9] as a decoder to predict the target object masks. Second, the top-down methods design a two-stage with a top-down perspective for object tracklet generation and tracklet-language grounding, respectively. They first construct a set of object candidate objects from the video, and then the referred target is selected from the tracklet to choose the best matching candidate. Finally, MTTR [1] and ReferFormer [13] recently demonstrated a promising solution that relies on the query-based mechanism. ReferFormer views the language as queries with a meta-architecture as [4,5], which provides a simple and unified framework for referring video object segmentation.

Benefiting from the efficient and powerful query-based architecture of Mask2Former for universal image segmentation, we propose an end-to-end network for referring video object segmentation, named VLFormer. We carefully design a Visual-Language Transformer block, which is the key component of VLFormer to guide the object queries by both the visual and linguistic features simultaneously. After generating the masks for the referred object, we select uniformly reference frames as a memory bank for STCN [6] to predict the high-quality mask for the entire video to obtain the final prediction.

2. Related works

Semi-supervised Video Object Segmentation. [**?**] This task aims to segment target objects throughout a video, given only a fully-annotated mask in the first frame. Most recent works [6,11,14] perform a memory feature matching. In particular, STM [11] leverages a memory bank from past frames and utilizes the query-key-value attention mechanism to predict the mask of the current frame. AOT [14] constructs hierarchical matching based on a Long Short-Term Transformer, STCN [6] uses directly image-to-image correspondence for more robust matching, leading to a simple, fast and efficient framework.

Query-based methods. Recently, query-based architectures have achieved marvellous success in most computer vision tasks such as object detection, object tracking and image segmentation. DETR [2] introduces the new query-based paradigm for object detection. VisTR [12] extends the framework for video instance segmentation (VIS), Mask2Former [4, 5] builds upon a simple meta-framework with a new Transformer decoder using masked attention and obtains top results in a variety of image and video seg-

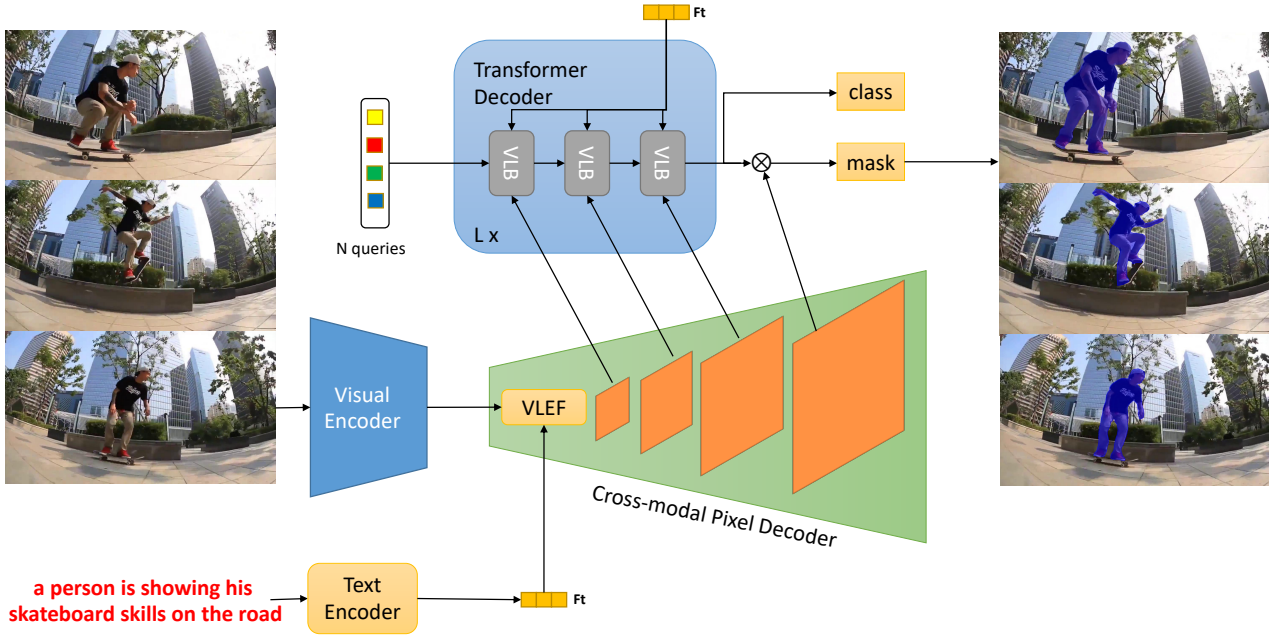

Figure 1. **VLFormer overview.** It contains three major components: Backbone(visual encoder, text encoder), Cross-modal Pixel Decoder, and Transformer Decoder.

mentation tasks. ReferFormer [13] even use language as queries and achieves the state-of-the-art performance on R-VOS task. Inspired by these works, our work also relies on the query-based mechanism of Transformer and enhance the linguistic feature for each Transformer block.

3. Method

Given a video clip $V = \{V_1, V_2, ..., V_T\}$ and a referring expression with L words. We propose an efficient and simple end-to-end framework named VLFormer for referring video object segmentation, as shown in Figure 1. It consists of 3 main components: Backbone, Cross-modal Pixel Decoder and Transformer Decoder. A backbone includes Visual Encoder and Text Encoder, which extract low-resolution feature maps from video frames and linguistic features from the corresponding referring expression, respectively. A Cross-modal Pixel Decoder fuses low-resolution features from the backbone and linguistic features by a Visual-Language Early Fusion (VLEF) and upsamples these features to generate the high-resolution per-pixel embeddings. Finally, a Transformer decoder operates on linguistics and visual features to process the object queries. The final binary mask predictions are generated by associating the per-pixel embeddings and object queries. We select the query with the highest average confidence score as the target object during inference. Regarding refinement and re-tracking the target object, we use an off-the-shelf semi-supervised VOS model named STCN [6] that uses some reference frames as a memory bank to predict the entire video.

3.1. VLFormer

Cross-modal Pixel Decoder. After extracting the entire video by the visual encoder, we obtain the multi-scale low-resolution feature maps $\{F_v^i\}_{i=1}^T$ for each frame i. Firstly, these features are fused with the linguistic features F_t, which are extracted by the text encoder to perform an early interaction between visual and linguistic features by a Visual-Language Early Fusion block (VLEF) for highlighting regions that are matched with the referring expressions.

$$\hat{F}_v^i = F_v^i \odot MSA(F_v^i W^Q, F_t W^K, F_t W^V) \quad (1)$$

where MSA(q, k, v) is the multi-head attention layer and $W^Q, W^K, W_V \in \mathbb{R}^{C \times d_{head}}$ are learnable parameters. This multi-head attention layer is used as a compatibility weight between the visual features and linguistic features. And then, we multiply this weight with the visual feature to focus on region high related to referring expressions.

The rest of Pixel Decoder is adopted from Mask2Former [4] with a multi-scale deformable attention decoder to produce a multi-scale features map for video frames.

Vision-Language Transformer Block Inspired by the

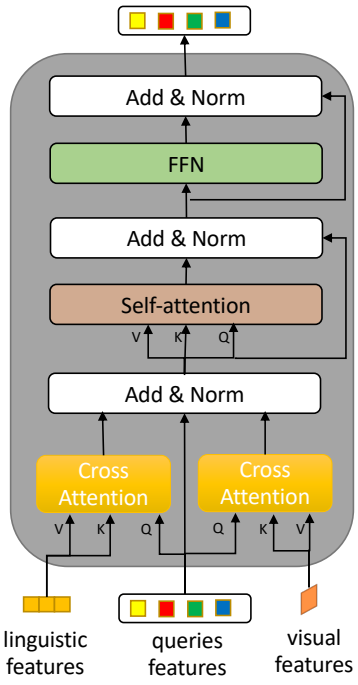

Figure 2. **Visual-Languagge Transformer BLock(VLB) –** The linguistic and visual features are used to update the queries features simultaneously.

Transformer decoder in Mask2Former [4], we propose a Visual-Language Transformer Block(VLB) to conduct the interaction between queries and visual features and between queries and linguistic features simultaneously. As illustrated in Fig. 2, given the visual features of the video frames, linguistic features of the referring expressions, and the object queries features First, Q interact independently with linguistic and visual features in a cross-attention way, where object queries and linguistic/visual features are query and key, respectively. And then, they will be summed up and fed into a multi-head self-attention module to update the object queries features and gather the contextual information of both linguistic and visual features. Finally, a FFN module is applied to these features to get the final object query features.

3.2. Inference

As mentioned in [3, 13], due to sharing of queries across frames, VLFormer can segment and track the referred object across frames. Given the video and language expression, VLFormer will generate N candidate instance for the whole video. We obtain the confidence score set $P_i = \{p_{i}\}_{j=1}^{N}$ by averaging the predicted probabilities over all the frames for each instance query. We select the instance with the highest score, and its index is donated as σ. The segmentation mask for each frame is obtained from the mask candidates set by selecting the corresponding query index σ.

First, we uniformly sample K = 10 keyframes for each video sequence. Then, these keyframes are memorized in a memory bank of the STCN [6] to re-predict target masks of the entire video.

	Team	$\mathcal{J}\&\mathcal{F}(\uparrow)$	$\mathcal{J}(\uparrow)$	$\mathcal{F}(\uparrow)$
1	Bo___	0.641	0.622	0.661
2	jiliushi	0.617	0.598	0.636
3	PENG	0.608	0.589	0.627
4	ds-hohhot	0.596	0.579	0.612
5	JQK	0.594	0.577	0.611
6	nero(**Ours**)	0.580	0.561	0.599

Table 1. Results in Ref-YouTube-VOS 2022 *test* set.

Backbone	STCN	$\mathcal{J}\&\mathcal{F}(\uparrow)$	$\mathcal{J}(\uparrow)$	$\mathcal{F}(\uparrow)$
ResNet101		0.563	0.546	0.581
Swin-B		0.622	0.602	0.642
Swin-B	✓	**0.632**	**0.610**	**0.654**

Table 2. **Ablation study** on *development* set.

4. Experiments

4.1. Dataset and Evaluation Metrics

The challenge dataset has 3,978 videos with about 15000 language expressions in total. There are 3,471 fully-annotated videos that are released for training. The rest videos are split into 202/304 for development/testing sets.

4.2. Training Details

We used Detectron2 to train our network. The loss function and training strategy in our work follows Mask2Former [4]. We train our network with two different backbone ResNet101 [7] and SwinTransformer-Base [8]. The text encoder is frozen during the training stage. In the Transformer Decoder, we choose $L = 3$ with 9 decoder layers in total. Following [13], we choose number of queries $N = 5$. The initial backbone's weights have been previously pretrained on ImageNet. Our model implementations mostly follows Mask2Former [4] with SwinTransformer-B as a backbone. STCN is pretrained mainly on COCO and finetuned over training split of semi-supervised VOS task.

We train the network for 10 epochs using the AdamW [10] optimizer with the initial learning rate 10^{-4}. A factor of 0.1 decreases the learning rate at the 6th epoch. We train the network with a small batch size of 2 on 2 Tesla V100 with 16GB GPU.

Figure 3. **Qualitative results** on *test* set. Each referring expression and the corresponding referred object are highlighted in the same color.

4.3. Results

Table 1 show the result of top teams in *testing* set. Our method ranks 6^{th} on Refer-Youtube-VOS 2022. Without any ensemble technique and lacking of resources, we can achieve a result of 0.58 in overall $\mathcal{J}\&\mathcal{F}$ on *testing* set. Figure 3 shows the qualitative results of our proposed model on *testing* set. Each referring expression and the corresponding referred object are highlighted in the same color.

4.4. Ablation study

As shown in Table 2, the backbone SwinTransformer [8] can improve the result of 3.9% in terms of overall $\mathcal{J}\&\mathcal{F}$ on *test* set from 56.3% to 62.2%.The introducing STCN [6] as post-processing phase to refine and propagate the coarse masks from the VLFormer can bring a 1.0% improvements.

5. Conclusions

In this paper, we propose a Vision-Language Transformer Network (VLFormer) for referring video object segmentation. Our approach achieves an overall score of 0.58, and ranks 6th on YouTube-VOS-2022 referring video object segmentation challenge.

References

[1] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. 1

[3] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. type: article. 3

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. type: article. 1, 2, 3

[5] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. type: article. 1

[6] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. type: article. 1, 2, 3, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 4

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2014. 1

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. 3

[11] Seoung Oh, Joon-Young Lee, Ning Xu, and Seon Kim. Video object segmentation using space-time memory networks. pages 9225–9234, 10 2019. 1

[12] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[13] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. type: article. 1, 2, 3

[14] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1