Segment Every Reference Object in Spatial and Temporal Spaces

Jiannan Wu[1], Yi Jiang[2], Bin Yan[3], Huchuan Lu[3], Zehuan Yuan[2], Ping Luo[1,4]
[1]The University of Hong Kong [2]ByteDance
[3]Dalian University of Technology [4]Shanghai AI Laboratory

Abstract

The reference-based object segmentation tasks, namely referring image segmentation (RIS), referring video object segmentation (RVOS), and video object segmentation (VOS), aim to segment a specific object by utilizing either language or annotated masks as references. Despite significant progress in each respective field, current methods are task-specifically designed and developed in different directions, which hinders the activation of multi-task capabilities for these tasks. In this work, we end the current fragmented situation and propose UniRef to unify the three reference-based object segmentation tasks with a single architecture. At the heart of our approach is the multiway-fusion for handling different task with respect to their specified references. And a unified Transformer architecture is then adopted for performing instance-level segmentation. With the unified designs, UniRef can be jointly trained on a broad range of benchmarks and can flexibly perform multiple tasks at runtime by specifying the corresponding references. We evaluate the jointly trained network on various benchmarks. Extensive experimental results indicate that our proposed UniRef achieves state-of-the-art performance on RIS and RVOS, and performs competitively on VOS with a single network.

1. Introduction

The reference-guided object segmentation aims at segmenting the specified object with the given references (*e.g.*, language or first-frame mask annotation). The three representative tasks include referring image segmentation (RIS) [94], referring video object segmentation (RVOS) [34] and semi-supervised video object segmentation (VOS) [61]. They are the fundamental tasks for vision understanding and have wide applications in image/video editing, video surveillance, etc. Over time, many advanced methods have ballooned in their respective fields and rapidly improves the state-of-the-art performance.

Despite witnessing the significant progress, these tasks are separately tackled with specialized designed models.

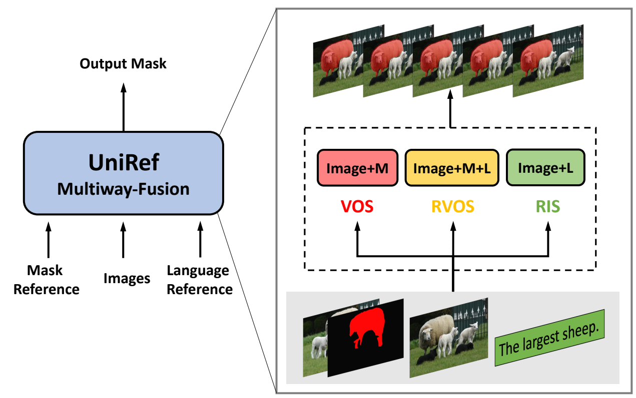


Figure 1: A single, jointly trained UniRef can perform three different reference-based tasks by specifying the corresponding references. 'L' and 'M' represent language and mask reference, respectively.

In that regard, the individual methods need extra training time and produce different sets of model weights on each task. This would cause expensive computational cost and yield redundant parameters. More importantly, the independent designs prevent the synergy and facilitation of different tasks. We argue that the current fragmented situation is unnecessary as the three tasks have essentially the same definition in a high-level aspect: they all use the references (language or first-frame annotated mask) as guidance to perform the per-pixel segmentation of the target object. This motivates us to build a unified model within the same parameters which can perform different tasks at runtime by specifying the corresponding references.

Towards the unification of reference-based object segmentation tasks, it poses great challenges in connecting the isolated landscapes as a whole: **(i) The mainstream methods in different fields vary greatly.** RIS and RVOS methods [91, 29, 86, 66] mostly focus on the deep cross-modal fusion of vision and language information. While the space-time memory network for pixel-level matching has long dominated the VOS domain [59, 16, 88, 15]. **(ii) Previous RVOS and VOS tasks are solved in two dif-**

ferent paradigms. The previous state-of-the-art RVOS methods [6, 78] take the whole video as input and generate the prediction results for all frames in one single step, which termed as offline methods. In contrast, VOS methods [59, 16] operate in an online fashion where they readout the historical features to propagate the target masks frame by frame. **(iii) The image-level RIS methods cannot be simply extended to the video domain for RVOS.** RIS only requires to segment the referred target in a single image. For the video tasks, however, the objects may encounter occlusion, fast motion or disapperance-reappearance in many complex scenes, which requires the networks to leverage the spatio-temporal information to track the objects throughout the whole video. Hence, simply adopting the image-level methods for each frame independently cannot ensure the temporal consistency for target object in videos.

In this work, we conquer the challenges above and propose a unified model, UniRef, for the reference-based object segmentation tasks. The key idea behind our approach is to formulate all three tasks as instance-level segmentation problems, and the information of references can be injected into the network through an attention-based fusion process regardless of their modalities. As illustrated in Figure 1, for different tasks, UniRef receives the current frame and utilizes the corresponding references to perform the fusion process, termed as multiway-fusion. Specifically, the language description and annotated mask in the first frame are leveraged as references for RIS and VOS tasks, respectively. And we emphasize that, for RVOS, both the language and mask references are used. This design not only tackles RVOS in an online fashion, but also can utilize the historical information for mask propagation to ensure the temporal consistency for target object, and thus establishing a new paradigm for RVOS. Practically, we introduce a UniFusion module to fuse the visual features and the specified references. Afterwards, the visual features of current frame are fed into a unified Transformer architecture, where queries are employed for instance-level segmentation of the target object. Thanks to the unified architecture, our model can be jointly trained on the broad range of benchmarks to learn the general knowledge, and can flexibly perform multi-tasks at runtime by specifying the corresponding references.

To summarize, our contributions are as follows:

- We propose UniRef, a unified model to perform three reference-based object segmentation tasks (RIS, RVOS, VOS) with the same model weights.
- We introduce a UniFusion module to inject the reference information into the network regardless of their modalities. And we establish a new online paradigm for RVOS by leveraging both language and mask as references.
- Extensive experiments demonstrate that our models achieve state-of-the-art performance for RIS and RVOS, and perform competitively for VOS.

2. Related Work

2.1. Unified Model

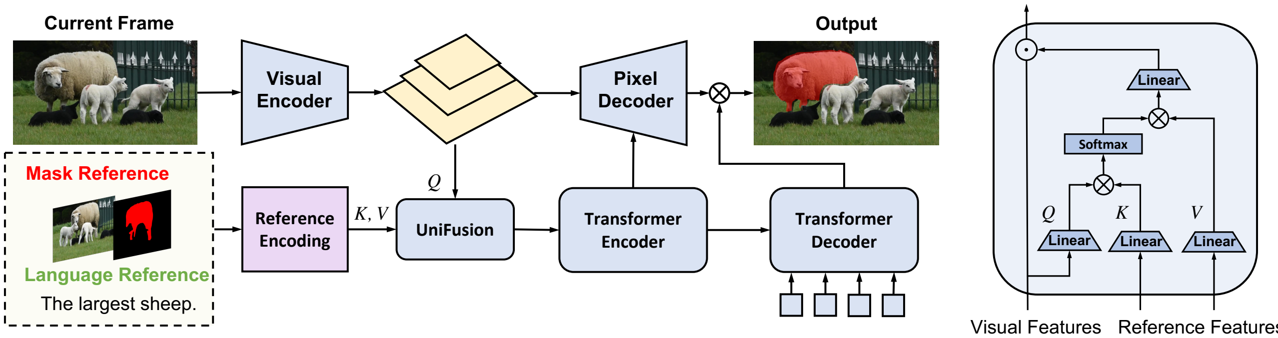
Towards achieving general artificial intelligence, the vision community has clearly witnessed the trend of building unified models recently. One line of works [1, 95, 92, 76, 9, 10, 101, 36, 74, 53] is to design the *general interface* for vision or vision-language (VL) tasks. For example, Unified-IO [53] unifies broad range of image-level tasks (*e.g.*, image classification [17], image caption [11] and VQA [2]) in a sequence-to-sequence generation paradigm. Another line of works [13, 33, 39, 97, 84, 102, 31, 5, 3, 85] is to build the *unified architecture* for the closely related tasks. GLIP [39] formulates both the object detection and phrase grounding tasks as the word-region alignment problem. OneFormer [31] rules the universal image segmentation tasks with a single Transformer network. Unicorn [84] proposes the designs of target priors to tackle four tracking tasks. However, these methods with unified architecture either focus on the image domain or only consider the visual-only tasks. We aim to bridge the gap by building the unified model for the reference-based object segmentation tasks.

2.2. Task-specific Object Segmentation

Referring Image Segmentation. The objective of RIS [27] is to generate a pixel-level mask for the target object described by a language description in an image. Prior research has primarily focused on the multi-modal feature interaction techniques, either employing attention mechanism in CNNs [93, 8, 91, 29, 28, 32, 86] or using multi-modal Transformers [18, 35, 77]. As RIS is closely related to referring expression comprehension (REC), which aims to predict the bounding box of the referred object, some works [54, 41, 99] also explore the unified frameworks that can accomplish these two tasks simultaneously.

Referring Video Object Segmentation. RVOS can be considered as an extension of RIS in the video domain. Some previous methods process the video frames independently [91, 98, 20, 38] or simply adopt 3D CNNs [73, 57, 72] to extract the spatio-temporal features for a video clip. Recently, state-of-the-art methods [6, 78] are based on query-based Transformers and process the videos in an offline fashion. They receive the whole video as input and employ the queries to segment and track the target object simultaneously. However, such methods are not suitable for long videos or ongoing videos. In contrast to these works, our UniRef belongs to the online method and can utilize the historical information for mask propagation, which can ensure the temporal consistecy of target object and improve segmentation accuracy.

Video Object Segmentation. Given a video with the target mask annotations in the first frame, the VOS algorithms need to propagate the masks to the entire video. The pre-



(a) Main Pipeline of Our Framework (b) UniFusion Module

Figure 2: Illustration of (a) the overall framework of UniRef. For sake of clarity, we omit the visualization of prediction heads which are on top of Transformer decoder. The core network (in blue) is shared for all tasks. (b) The details of UniFusion module. The reference features come from the language or mask references.

vious approaches could be broadly categorized into two groups: (i) Template-based methods. These works [70, 75, 71, 65, 12] regard the annotated frame as template and investigate how to fuse the template information into the current frame. (ii) Memory-based methods. The pioneering work STM [59] leverages a memory network to embed past-frame predictions and learns the space-time pixel-level correspondence on the memory to propagate the mask information. This type of works has achieved significant improvement and dominated the VOS community. The subsequent works mainly focus on improving memorized embeddings [87, 89, 40, 43, 83, 56], designing novel memory networks [81, 15] or proposing reliable memory readout strategies [67, 16, 88, 90]. These previous works view the VOS task as the pixel-level binary classification problem, lacking the understanding of object. In contrast to theirs, we tackle the VOS as the instance segmentation problem.

3. Method

3.1. Overview

We present UniRef, a simple and unified architecture that can segment arbitrary objects with the given references in images/videos. Conceptually, it allows us to train a single network jointly on all related benchmarks and simultaneously solves the aforementioned tasks (RIS, RVOS, VOS).

The overall architecture of UniRef is illustrated in Fig. 2. Our framework consists of a visual encoder, two reference encoders (for text and mask, respectively), a proposed UniFusion module and a transformer-based object detector. Given an image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ and the corresponding references, we first use the visual encoder $\mathbf{Enc}_V$ to extract the multi-scale features $\boldsymbol{\mathcal{F}} = \{\boldsymbol{F}_\ell\}_{\ell=1}^{4}$ of current image, where ℓ denotes the level index of the hierarchical visual features, with the spatial strides from 4 to 32. Then the reference encoders are applied to encode the reference information, followed by the UniFusion module to inject the information into the visual features. Finally, the network can produce a binary mask for the target object $\boldsymbol{m} \in \mathbb{R}^{H \times W}$ via a unified Transformer-based architecture.

In the following subsections, we are going to details of UniRef by introducing the reference encoding (Sec. 3.2), the multi-scale UniFusioin module (Sec. 3.3), a unified encoder-decoder architecture (Sec. 3.4) and the training/inference process of UniRef (Sec. 3.5).

3.2. Reference Encoding

In this part, we introduce how to encode the reference information for the three reference-based tasks. Before that, we would like to clarify that the only task-specific design of UniRef is to use different reference encoders (text encoder $\mathbf{Enc}_T$ and mask encoder $\mathbf{Enc}_M$) for processing different modalities.

Video Object Segmentation. For VOS task, the target object mask annotation in the first frame is provided as the reference. And the network needs to propagate the mask throughout the video. Inspired by the spirit of STCN [16] that computing the similarity of two frames for once, we use the same visual encoder $\mathbf{Enc}_V$ to extract the hierarchical visual features $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{f}} = \left\{\boldsymbol{F}_{V,\ell}^{\mathrm{f}}\right\}$ of reference frame $\boldsymbol{I}_{\mathrm{ref}}$. Then, a lightweight mask encoder (*e.g.*, ResNet18 [25]) receives the reference frame $\boldsymbol{I}_{\mathrm{ref}}$, object mask annotation $\boldsymbol{m}_o$ and the encoded frame features $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{f}}$ to generate the multi-scale mask features $\boldsymbol{\mathcal{F}}_{\boldsymbol{V}}^{\mathbf{m}} = \left\{\boldsymbol{F}_{V,\ell}^{\mathrm{m}}\right\}$ for the target object in reference frame. Here, $\ell = 2, 3, 4$ for $\boldsymbol{F}_{V,\ell}^{\mathrm{f}}$ and $\boldsymbol{F}_{V,\ell}^{\mathrm{m}}$.

$$\boldsymbol{\mathcal{F}}_V^f = \mathbf{Enc}_V(\boldsymbol{I}_{\mathrm{ref}}) \tag{1}$$

$$\boldsymbol{\mathcal{F}}_V^m = \mathbf{Enc}_M(\boldsymbol{I}_{\mathrm{ref}}, \boldsymbol{m}_o, \boldsymbol{\mathcal{F}}_V^f) \tag{2}$$

Referring Image Segmentation. The reference for RIS task is the language description $\boldsymbol{T}$. To encode such linguistic information, we apply an off-the-shelf text encoder RoBERTa [49] to extract the language features $\boldsymbol{F}_T \in$

$\mathbb{R}^{L\times C}$, where L is the sentence length and C is the channel dimension.

$$\boldsymbol{F}_T = \mathbf{Enc}_T(\boldsymbol{T}) \tag{3}$$

Referring Video Object Segmentation. RVOS requires the model to not only understand the language description, but also track the referred object in the whole video. To this end, we encode both linguistic and visual information for this task. Similarly, we use Eq. 3 to extract language feature and apply Eq. 1 and Eq. 2 for mask features encoding. It should be noted that the mask annotation is available during training. And we use the predicted mask in the previous frame as the visual reference during inference.

3.3. Multi-scale UniFusion Module

After the reference encoding, a natural question is raised: *How to inject the reference information into the network*? In this subsection, we introduce our proposed multi-scale UniFusion module for the reference information injection.

We fuse the visual features $\mathcal{F}$ and the reference features in a hierarchical manner. For simplicity, we take the ℓ-th ($\ell = 2, 3, 4$) visual level for illustration. The UniFusion module receives three inputs: the ℓ-th level visual feature $\boldsymbol{F}_\ell$ of current image/frame, the key embedding and value embedding of the reference features. These inputs are first linearly projected and further reformulated as three vectors, namely $\boldsymbol{Q}_\ell$, $\boldsymbol{K}_\ell$ and $\boldsymbol{V}_\ell$. We first perform the attention operation between these vectors. Then, the output features are injected into the original visual features via element-wise multiplication. The process of UniFusion is represented as:

$$\boldsymbol{O}_\ell = \text{Attention}(\boldsymbol{Q}_\ell, \boldsymbol{K}_\ell, \boldsymbol{V}_\ell) = \text{Softmax}(\frac{\boldsymbol{Q}_\ell \boldsymbol{K}_\ell^\top}{\sqrt{C}})\boldsymbol{V}_\ell \tag{4}$$

$$\boldsymbol{F}'_\ell = \text{MLP}\big(\text{LN}(\boldsymbol{O}_\ell)\big) \odot \boldsymbol{F}_\ell \tag{5}$$

where $\boldsymbol{O}_\ell$ is the intermediate results after the attention operation. $\boldsymbol{F}'_\ell$ is the final output of UniFusion. $\odot$ means element-wise multiplication. Notably, the UniFusion module shares the same parameters in all visual scales. Thanks to the unifying fusion format, the reference information in different tasks can be injected into the visual features using the same UniFusion module.

$$\boldsymbol{F}'_\ell = \begin{cases} \mathbf{UF}(\boldsymbol{F}_\ell, \boldsymbol{F}^{\text{f}}_{V,\ell}, \boldsymbol{F}^{\text{m}}_{V,\ell}), & \text{VOS} \\ \mathbf{UF}(\boldsymbol{F}_\ell, \boldsymbol{F}_T, \boldsymbol{F}_T), & \text{RIS} \\ \mathbf{UF}(\boldsymbol{F}_\ell, \boldsymbol{F}^{\text{f}}_{V,\ell}, \boldsymbol{F}^{\text{m}}_{V,\ell}) \odot \mathbf{UF}(\boldsymbol{F}_\ell, \boldsymbol{F}_T, \boldsymbol{F}_T), & \text{RVOS} \end{cases} \tag{6}$$

where **UF** is the abbreviation of UniFusion module. We emphasize here, for RVOS task, both the language features and reference frame visual features are fused with the visual feature of current frame. In this fashion, the network can not only find the referred object by language, but also propagate the target mask across frames for tracking. This also unifies the paradigms of VOS and RVOS to the online pattern.

3.4. Unified Architecture

The fused multi-scale visual features $\mathcal{F}' = \{\boldsymbol{F}'_\ell\}_{\ell=2}^4$ have discriminative representations in highlighting the specific target by reference. We next adopt a unified Transform-based architecture to predict the target mask.

Transformer. We use the two-stage version Deformable-DETR [100] as our object detector. It receives the fused hierarchical visual features $\mathcal{F}'$ as input and perform multi-scale deformable self-attention in the encoder. In the decoder, N object queries are iteratively refined over stacked decoder layers and converted into the query representations $\boldsymbol{Q}_{\text{obj}} \in \mathbb{R}^{N\times C}$ finally. Three predictions heads (class head, box head and mask head) are further built on top of the decoder to predict the object scores $\boldsymbol{S} \in \mathbb{R}^{N\times 1}$, boxes $\boldsymbol{B} \in \mathbb{R}^{N\times 4}$ and mask dynamic convolution [69, 14, 13] kernel parameters $\mathcal{G} = \{\boldsymbol{g}_i\}_{i=1}^N$, respectively.

Mask Decoder. We take the output features (from strides 8 to 32) of Transformer encoder and hierarchically fuse them in a FPN-like [44, 79] manner. The feature map with 4× strides of backbone, namely $\boldsymbol{F}_1$, is also added in this process. This is helpful for preserving the reference-agnostic and fine-grained information of images. Consequently, we obtain the high-resolution mask features $\boldsymbol{F}_{\text{seg}} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$. Finally, the masks of target object are generated by performing dynamic convolution between $\boldsymbol{F}_{\text{seg}}$ and $\mathcal{G}$:

$$\boldsymbol{m}_i = \text{Upsample}\big(\text{DynamicConv}(\boldsymbol{F}_{\text{seg}}, \boldsymbol{g}_i)\big),\ i = 1, ..., N \tag{7}$$

During inference, we choose the mask with the highest score as the final result $\boldsymbol{m}$ for the target object. Notably, we empirically find that using more object queries leads to higher performance, despite that one object query is sufficient for the reference-based tasks.

3.5. Training and Inference

We jointly train UniRef on all related benchmarks of reference-based object segmentation tasks, *i.e.*, RIS, RVOS and VOS. The model with the *same weights* can perform different tasks at runtime by specifying the references.

Training. The network predicts N predictions of object scores, box coordinates and segmentation masks, where the object score indicates whether the object is visible in current frame. During training, we apply the set prediction loss [7, 100, 68] on these predictions. There is only one ground-truth for the reference-based object segmentation tasks. We assign multiple predictions to the ground-truth by selecting the top-k predictions with the least cost according to an optimal transport method [23, 24, 80]. The matching cost is formulated as:

$$\mathcal{C} = \lambda_{cls} \cdot \mathcal{C}_{cls} + \lambda_{L1} \cdot \mathcal{C}_{L1} + \lambda_{giou} \cdot \mathcal{C}_{giou} \tag{8}$$

where $\mathcal{C}_{cls}$ is the focal loss [45]. The box losses include the widely-used ℓ_1 loss and generalized IoU (GIoU) loss [64].

Table 1: Comparison with the state-of-the-art methods on three referring image segmentation (RIS) benchmarks. RN101 denotes ResNet-101 [25], WRN101 refers to Wide ResNet-101 [96], and DN53 denotes Darknet-53 [63].

	Method	Visual Backbone	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg	
				val	test A	test B	val	test A	test B	val-u	test-u
oIoU	CMSA [91]	RN101	LSTM	58.32	60.61	55.09	43.76	47.60	37.89	-	-
	STEP [8]	RN101	Bi-LSTM	60.04	63.46	57.97	48.19	52.33	40.41	-	-
	BRINet [28]	RN101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	-	-
	CMPC [29]	RN101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-
	LSCM [30]	RN101	LSTM	61.47	64.99	59.55	49.34	53.12	43.50	-	-
	CMPC+ [48]	RN101	LSTM	62.47	65.08	60.82	50.25	54.04	43.47	-	-
	MCN [54]	DN53	Bi-GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
	EFN [22]	WRN101	Bi-GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-
	LTS [32]	DN53	Bi-GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
	ReSTR [35]	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	-	-
	LAVT [86]	Swin-B	BERT-base	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
	UniRef-R50	RN50	RoBERTa-base	75.04	77.28	72.43	63.25	68.12	55.56	66.96	68.77
	UniRef-L	Swin-L	RoBERTa-base	**79.79**	**81.81**	**77.02**	**69.26**	**74.11**	**63.14**	**73.04**	**73.36**
mIoU	VLT [18]	DN53	Bi-GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
	CRIS [77]	RN101	GPT-2	70.47	73.18	66.10	62.27	68.06	53.68	59.87	60.36
	SeqTR [99]	DN53	Bi-GRU	71.70	73.31	69.82	63.04	66.73	58.97	64.69	65.74
	RefTr [41]	RN101	BERT-base	74.34	76.77	70.87	66.75	70.58	59.40	66.63	67.39
	LAVT [86]	Swin-B	BERT-base	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
	PolyFormer-L [47]	Swin-L	BERT-base	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17
	UniRef-R50	RN50	RoBERTa-base	78.14	80.09	75.94	69.09	73.64	62.62	71.76	73.10
	UniRef-L	Swin-L	RoBERTa-base	**81.90**	**83.03**	**79.61**	**73.81**	**78.30**	**68.33**	**76.65**	**77.09**

The top-k predictions with the least cost are assigned as positive samples and others as negatives. UniRef is optimized by minimizing the following loss function:

$$\mathcal{L} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{L1} \cdot \mathcal{L}_{L1} + \lambda_{giou} \cdot \mathcal{L}_{giou} + \lambda_{mask} \cdot \mathcal{L}_{mask} + \lambda_{dice} \cdot \mathcal{L}_{dice} \tag{9}$$

where the class loss and boxes losses are the same as those in Eq. 8. The mask-related losses contain the mask binary cross-entropy loss and DICE loss [58].

Inference. For RIS, we directly output the predicted mask of the query that has the highest score. For RVOS and VOS, our method infers the video in a frame-by-frame online fashion without the complex post-processing. Specifically, for the current frame, the network uses the corresponding references to produce the mask of target object. The mask would be output if its object score is higher than a pre-determined threshold σ. Otherwise the output mask values are all set to zeros. To handle the videos that contain multiple objects, we adopt the soft-aggregation method commonly used in prior works [59, 16].

4. Experiments

In this section, we conduct comprehensive experiments on all reference-based tasks (RIS, RVOS and VOS) to evaluate the effectiveness of our proposed UniRef. The experimental settings will be first introduced in Sec. 4.1. We then compare UniRef with state-of-the-art methods on the prevalent benchmarks in Sec. 4.2. The ablation studies are presented in Sec. 4.3.

4.1. Experimental Setup

Datasets. We evaluate our UniRef on three tasks to verify its effectiveness. The specific datasets leveraged in this work for evaluation are presented in the following. **(i) RIS**: RefCOCO [94] consists of 142,209 language descriptions for 50,000 objects in 19,994 images. RefCOCO+ [94] has 141,564 expressions for 49,856 objects in 19,992 images. RefCOCOg [55] includes 85,474 referring expressions for 54,822 objects in 26,711 images. And we use the UMD split for RefCOCOg [55]. **(ii) RVOS**: Ref-YoutubeVOS [66] is a large-scale referring video object segmentation dataset which contains 3,978 videos with around 15k langauge descriptions. Ref-DAVIS17 [34] provides the referring expressions for each object in DAVIS17 [61]. It contains 90 videos in total. **(iii) VOS**: Youtube-VOS[1] [82] is the popular benchmark for video object segmentation. There are 474 and 507 videos in the validation set for 2018 and 2019 version, respectively. LVOS [26] is a long-term video object segmentation benchmark consisting of 220 videos. The

[1] Youtube-VOS and Ref-Youtube-VOS are evaluated using the official server https://youtube-vos.org/.

videos in LVOS have an average duration of 1.59 minutes, and the videos in Youtube-VOS last 6 seconds on average.

Implementation Details. We experiment with two prevalent backbones as our visual encoder: ResNet50 [25] and Swin Transformer-Large [50]. The text encoder is selected as RoBERTa [49] following the previous works [6, 78, 33] and we set the max length of sentences as 77. The Transformer architecture has 6 encoders and 6 decoders with the channel dimension of 256. The number of object queries is set as 300 by default. The loss coefficients in Eq. 9 are set as $\lambda_{cls} = 2.0$, $\lambda_{cls} = 2.0$, $\lambda_{L1} = 5.0$, $\lambda_{mask} = 2.0$ and $\lambda_{dice} = 5.0$, respectively.

The entire training process includes three sequential stages, in which pretrained weights from the previous stage are loaded and used for further training. **(i)** Visual Genome (VG) [37] pretraining. Due to the absence of mask annotations in VG dataset, we do not employ mask supervision in this training stage. **(ii)** Image-level training. We combine the training set of RefCOCO/+/g to train the network. **(iii)** Video-level training. At this stage, we randomly sample two frames from a video, where the first frame is considered as the reference frame. To avoid the knowledge forgetting for the RIS task, we also generate pseudo videos for RefCOCO/+/g. The network is jointly trained on all the related benchmarks, including RefCOCO/+/g [94, 55], Ref-YoutubeVOS [66], Ref-DAVIS17 [34], COCO [46], Youtube-VOS19 [82], OVIS [62] and LVOS [26].

In this work, we use Pytorch toolkit [60] to conduct all experiments on NVIDIA A100 GPUs. Unless otherwise stated, we use 4×8 A100 GPUs for the VG pretraining and 2×8 GPUs for the following image-level and video-level training. We adopt AdamW [52] as the optimizer and set the batch size as 2 for each GPU. We refer the readers to Appendix for more implementation details.

4.2. Quantitative Results

We employed ResNet50 [25] and Swin Transformer-Large [50] as visual backbones in our experiments, denoted as UniRef-R50 and UniRef-L, respectively. For each version, all results are computed with a single, jointly trained network which performs various tasks by simply altering the references during inference.

Referring Image Segmentation. We compare UniRef with state-of-the-art methods in Table 1. Following the previous works, we use both overall intersection-over-union (oIoU) and mean intersection-over-union (mIoU) as the evaluation metrics. It can be seen that UniRef with ResNet-50 backbone surpasses the previous methods on nearly all splits. When equipped with Swin-Large backbone, UniRef set new SOTA performance and outperforms the previous best results by a clear margin. For example, our UniRef-L has 4.96 mIoU performance gain over the recent SOTA method PolyFormer-L [47] on the validation split of RefCOCO .

Table 2: Comparison with the state-of-the-art methods for referring video object segmentation (RVOS). † and ‡ denote the model uses the `tiny` and `base` version of Video Swin Transformer [51] as visual encoders, respectively.

Method	Visual Encoder	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{J}$
Ref-Youtube-VOS				
CMSA [91]		36.4	34.8	38.1
URVOS [66]		47.2	45.3	49.2
YOFO [38]	ResNet-50	48.6	47.5	49.7
ReferFormer [78]		58.7	57.4	60.1
UniRef-R50		**60.6**	**59.0**	**62.3**
PMINet + CFBI [21]	Ensemble	54.2	53.0	55.5
CITD [42]		61.4	60.0	62.7
MTTR† [6]		55.3	54.0	56.6
VLT‡ [19]	Video-Swin	63.8	61.9	65.6
ReferFormer‡ [78]		64.9	62.8	67.0
ReferFormer [78]	Swin-L	64.2	62.3	66.2
UniRef-L		**67.4**	**65.5**	**69.2**
Ref-DAVIS17				
CMSA [91]		40.2	36.9	43.5
URVOS [66]		51.5	47.3	56.0
YOFO [38]	ResNet-50	54.4	50.1	58.7
ReferFormer [78]		58.5	55.8	61.3
UniRef-R50		**63.5**	**60.0**	**67.0**
VLT‡ [19]	Video-Swin	61.6	58.9	64.3
ReferFormer‡ [78]		61.1	58.1	64.1
PolyFormer-L [47]		61.5	57.2	65.8
ReferFormer [78]	Swin-L	60.5	57.6	63.4
UniRef-L		**66.3**	**62.9**	**69.7**

Referring Video Object Segmentation. For the RVOS task, we use the region jaccard $\mathcal{J}$, boundary accuracy $\mathcal{F}$ and the average score $\mathcal{J}\&\mathcal{F}$ as the evaluation metrics. The comparison of UniRef and state-of-the-art methods are presented in Table 2. It indicates that UniRef with the ResNet-50 visual encoder has significant improvement over all the previous methods on the two datasets. Especially, on Ref-DAVIS17, UniRef-R50 achieves notable 5.0 $\mathcal{J}\&\mathcal{F}$ gain over ReferFormer [78]. When using the Swin-Large as visual encoder, UniRef significantly outperforms the previous state-of-the-art ReferFormer [78] by a margin of 3.2 $\mathcal{J}\&\mathcal{F}$ on Ref-Youtube-VOS. UniRef-L also shows 3.6 $\mathcal{J}\&\mathcal{F}$ improvement over the image-level method VLT [18].

Video Object Segmentation. To evaluate the performance on Youtube-VOS [82], region jaccard $\mathcal{J}$ and countour accuracy $\mathcal{F}$ are computed for "seen" and "unseen" classes separately, denoted by subscripts s and u. $\mathcal{G}$ is the average $\mathcal{J}\&\mathcal{F}$ for both seen and unseen classes. For LVOS [26], $\mathcal{J}\&\mathcal{F}$, $\mathcal{J}$ and $\mathcal{F}$ are adopted as the evaluation metrics. We provide

Table 3: Comparison with the state-of-the-art methods on three video object segmentation (VOS) benchmarks.

Method	Youtube-VOS 2018 val					Youtube-VOS 2019 val					LVOS val		
	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Memory-based Methods													
STM [59]	79.4	79.7	84.2	72.8	80.9	-	-	-	-	-	-	-	-
AFB-URR [43]	79.6	78.8	83.1	74.1	82.6	-	-	-	-	-	34.8	31.3	38.2
CFBI [87]	81.4	81.1	85.8	75.3	83.4	81.0	80.6	85.1	75.2	83.0	50.0	45.0	55.1
RDE [40]	-	-	-	-	-	81.9	81.1	85.5	76.2	84.8	53.7	48.3	59.2
STCN [16]	83.0	81.9	86.5	77.9	85.7	82.7	81.1	85.4	78.2	85.9	45.8	41.1	50.5
AOT-B [88]	83.5	82.6	87.5	77.7	86.0	83.3	82.4	87.1	77.8	86.0	56.9	51.8	61.9
AOT-L [88]	83.8	82.9	87.9	77.7	86.5	83.7	82.8	87.5	78.0	86.7	59.4	53.6	65.2
XMem [15]	85.7	84.6	89.3	80.2	88.7	85.5	84.3	88.6	80.3	88.6	50.0	45.5	54.4
DeAOT [90]	86.0	84.9	89.9	80.4	88.7	85.9	84.6	89.4	80.8	88.9	-	-	-
Non-memory Methods													
FRTM [65]	72.1	72.3	76.2	65.9	74.1	-	-	-	-	-	-	-	-
LWL [4]	81.5	80.4	84.9	**76.4**	**84.4**	81.0	79.6	83.8	76.4	**84.2**	54.1	49.6	58.6
UniRef-R50	81.4	81.6	85.9	75.6	82.4	81.2	80.8	84.9	76.2	83.0	55.7	51.5	60.0
UniRef-L	**82.6**	**83.2**	**87.5**	76.2	83.7	**82.7**	**82.9**	**86.9**	**76.8**	84.1	**60.9**	**57.2**	**64.6**

Table 4: **Ablation experiments of UniRef**. We evaluate our model on the RefCOCO, Ref-Youtube-VOS and Youtube-VOS2018 validation set for the RIS, RVOS and VOS tasks, respectively. Our default settings are marked in gray.

Training	RIS (oIoU)	RVOS ($\mathcal{J}\&\mathcal{F}$)	VOS ($\mathcal{G}$)
Task-specific	73.25	57.8	82.2
Multi-task	73.68	60.1	82.2

(a) Task-specific Training.

Tasks	Levels	RIS (oIoU)	RVOS ($\mathcal{J}\&\mathcal{F}$)	VOS ($\mathcal{G}$)
✓		73.21	60.0	78.4
	✓	73.33	60.3	82.3
✓	✓	73.68	60.1	82.2

(b) Parameter-sharing for UniFusion.

Query	RIS (oIoU)	RVOS ($\mathcal{J}\&\mathcal{F}$)	VOS ($\mathcal{G}$)
1	70.64	56.7	79.3
100	73.63	60.0	82.1
300	73.68	60.1	82.2

(c) Query Number.

a comprehensive comparison of different methods on these three datasets in Table 3. Our proposed UniRef-L outperforms non-memory-based methods, achieving the best results with 82.6/82.7 $\mathcal{G}$ on the Youtube-VOS 2018/2019. Unlike memory-based methods [59, 16], our approach does not rely on predicted masks from past frames, making it more memory-efficient and suitable for handling long videos. As shown in last three columns of Table 3, UniRef-L reaches 60.9 $\mathcal{J}\&\mathcal{F}$, which demonstrates the state-of-the-art performance in the long-time LVOS dataset.

4.3. Ablation Study

The ablation experiments are evaluated on RefCOCO, Ref-Youtube-VOS and Youtube-VOS2018 to study UniRef in detail. Unless otherwise stated, we use the ResNet-50 as visual backbone and only go through the image-level training and video-level training for efficiency.

Task-specific Training. In Table 4a, we compare results of task-specific training and multi-task joint training. Task-specific models are trained on the corresponding datasets, while multi-task models are trained jointly on all datasets. The ablation results indicate that multi-task learning offers significant benefits for both RIS and RVOS. Specifically, Ref-Youtube-VOS achieves 60.1 $\mathcal{J}\&\mathcal{F}$, which is 2.3 points higher than the task-specific model. This can be attributed to the fact that the jointly trained model is better at learning mask propagation through VOS training. For VOS, the unified model produces the similar results to the task-specific model. Interestingly, the performance on Youtube-VOS2018 is higher than that in Table 3 (82.2 *v.s.* 81.4). This may be due to that the VG pretraining makes the network more suitable for RIS/RVOS tasks. In summary, multi-task joint training improves the performance of task-specific models and saves a significant number of parameters.

Parameter-sharing for UniFusion Module. Our UniRef leverages a parameter-sharing UniFusion module to fuse information from both mask and language references for multi-scale visual features. In Table 4b, we present ablation experiments on two weight-sharing variants of the UniFusion module. Firstly, we examine the use of different UniFusion modules for different visual levels. As shown in the first row of the table, we observe a consistent performance drop when compared with our original version. In particular, the $\mathcal{G}$ metric on Youtube-VOS2018 drops significantly

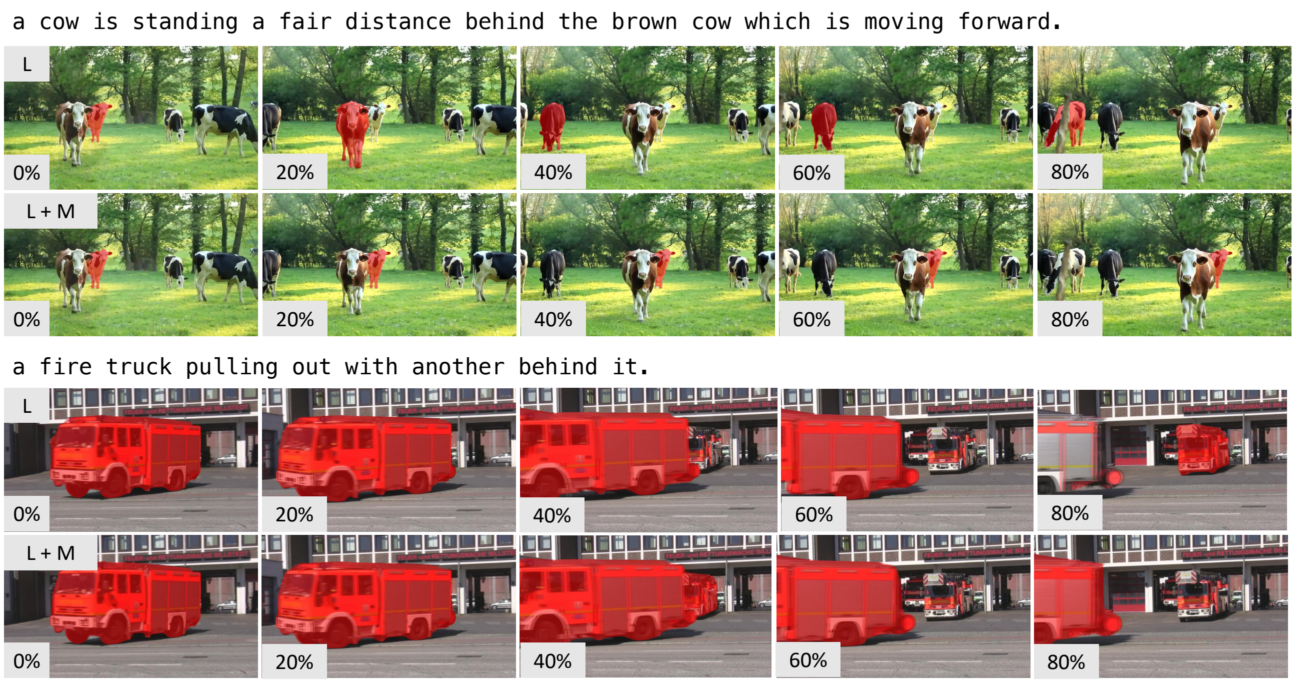


Figure 3: **Comparison of the use of mask references for RVOS**. The red masks highlight the predicted objects. The percentages indicate the relative temporal position of each frame in the video. 'L' and 'M' represent language and mask references, respectively.

from 82.2 to 78.4. This suggests that a single UniFusion module is more effective in learning frame similarities for multi-scale visual features. Secondly, we explore the use of task-specific UniFusion modules to handle different references (mask or language). By comparing the last two rows, we can observe that they obtain the similar results.

Query Number. Reference-based tasks have one specific target for each reference, making it possible to complete the tasks with just one query. In Table 4c, we present an ablation study on the number of queries to investigate its impact on performance. As observed, increasing the number of queries leads to higher performance. This is reasonable as the model can have more candidates to find the target object, which is particularly helpful in complex scenes where many similar objects co-exist.

Does Mask Reference Help RVOS? For the RVOS task, UniRef processes the videos in an online-fashion. Specifically, we not only use the language reference as guidance, but also leverage the predicted masks in the previous frames for mask propagation. To study the effectiveness of the mask reference, we use our final version models and provide the ablation results in Table 5. As illustrated in the table, when additionally using the mask references, the model gets 0.9 and 1.1 $\mathcal{J}\&\mathcal{F}$ improvement for UniRef-R50 and UniRef-L, respectively. This evidently proves that the mask propagation helps the model to achieve temporal consistency for the target object.

Table 5: **Ablation on the references used for RVOS**. In the table, 'Lang' means language. Results are evaluated on Ref-Youtube-VOS validation set.

Reference	ResNet-50			Swin-L		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Lang	59.7	58.5	60.8	66.3	64.5	68.1
Lang + Mask	60.6	59.0	62.3	67.4	65.5	69.2

4.4. Qualitative Results

In order to demonstrate the effectiveness of employing mask references in RVOS, we present the qualitative results in Figure 3. As depicted in the first example, utilizing language references alone struggles in identifying the referred object in a complex scene with multiple similar objects. By integrating mask references, the network can leverage mask propagation to accurately track the target object. This figure illustrates the efficacy of incorporating mask references in RVOS for improving temporal consistency of target object.

5. Conclusion

We present UniRef, a unified model for three reference-based object segmentation tasks (RIS, RVOS and VOS). By introducing a UniFusion module to incorporate different types of references, our model can flexibly perform multi-tasks at runtime by specifying the corresponding references and achieves superior performance with a single network.

Acknowledgements

This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17200622. The paper is supported in part by the National Natural Science Foundation of China under grant No.62293540, 62293542, U1903215 and the Fundamental Research Funds for the Central Universities No.DUT22ZD210.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified approach for target-based video segmentation. *arXiv preprint arXiv:2301.02657*, 2023. 2

[4] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 777–794. Springer, 2020. 7

[5] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: an end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12031–12041, 2022. 2

[6] 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. 2, 6

[7] 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. 4

[8] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019. 2, 5

[9] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 2

[10] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022. 2

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[12] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9384–9393, 2020. 3

[13] 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, 4

[14] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 4

[15] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 640–658. Springer, 2022. 1, 3, 7

[16] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 1, 2, 3, 5, 7

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[18] 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, 5, 6

[19] 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. 6

[20] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973, 2022. 2

[21] 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. 6

[22] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for

referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021. 5

[23] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 4

[24] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 4

[25] 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. 3, 5, 6

[26] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. *arXiv preprint arXiv:2211.10181*, 2022. 5, 6

[27] 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. 2

[28] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4424–4433, 2020. 2, 5

[29] 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. 1, 2, 5

[30] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 59–75. Springer, 2020. 5

[31] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. *arXiv preprint arXiv:2211.06220*, 2022. 2

[32] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9858–9867, 2021. 2, 5

[33] 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, 6

[34] 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. 1, 5, 6

[35] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022. 2, 5

[36] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. *arXiv preprint arXiv:2205.10337*, 2022. 2

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 6

[38] Dezhuang Li, Ruoqi Li, Lijun Wang, Yifan Wang, Jinqing Qi, Lu Zhang, Ting Liu, Qingquan Xu, and Huchuan Lu. You only infer once: Cross-modal meta-transfer for referring video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1297–1305, 2022. 2, 6

[39] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[40] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1332–1341, 2022. 3, 7

[41] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 2, 5

[42] 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. 6

[43] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020. 3, 7

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[45] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[47] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. *arXiv e-prints*, pages arXiv–2302, 2023. 5, 6

[48] 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. 5

[49] 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, 6

[50] 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*, pages 10012–10022, 2021. 6

[51] 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. 6

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[53] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2

[54] 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, 5

[55] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5, 6

[56] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9670–9679, 2021. 3

[57] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9942–9951, 2020. 2

[58] 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. 5

[59] 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. 1, 2, 3, 5, 7

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[61] 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. 1, 5

[62] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022. 6

[63] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[64] 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

[65] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7406–7415, 2020. 3, 7

[66] 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, 5, 6

[67] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12889–12898, 2021. 3

[68] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 4

[69] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–*

ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 282–298. Springer, 2020. 4

[70] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9481–9490, 2019. 3

[71] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6578–6588, 2020. 3

[72] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12152–12159, 2020. 2

[73] 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. 2

[74] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2

[75] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019. 3

[76] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[77] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2, 5

[78] 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. 2, 6

[79] Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021. 4

[80] 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. 4

[81] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1286–1295, 2021. 3

[82] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 5, 6

[83] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2946–2954, 2022. 3

[84] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 733–751. Springer, 2022. 2

[85] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 2

[86] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 1, 2, 5

[87] 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. 3, 7

[88] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 1, 3, 7

[89] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021. 3

[90] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *arXiv preprint arXiv:2210.09782*, 2022. 3, 7

[91] 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. 1, 2, 5, 6

[92] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[93] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 2

[94] 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. 1, 5, 6

[95] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[96] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 5

[97] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022. 2

[98] Wangbo Zhao, Kai Wang, Xiangxiang Chu, Fuzhao Xue, Xinchao Wang, and Yang You. Modeling motion with multi-modal features for text-based video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11737–11746, 2022. 2

[99] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 598–615. Springer, 2022. 2, 5

[100] 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. 4

[101] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815, 2022. 2

[102] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. *arXiv preprint arXiv:2212.11270*, 2022. 2