You Only Infer Once: Cross-Modal Meta-Transfer for Referring Video Object Segmentation

Dezhuang Li[1*], Ruoqi Li[1*], Lijun Wang[1], Yifan Wang[1†], Jinqing Qi[1], Lu Zhang[1], Ting Liu[2], Qingquan Xu[2], Huchuan Lu[1]

[1] Dalian University of Technology, Dalian, China
[2] Meitu Inc., China
{Merci, dutlrq77}@mail.dlut.edu.cn, {ljwang, wyfan, jinqing}@dlut.edu.cn,
luzhangdut@gmail.com, {lt, xqq}@meitu.com, lhchuan@dlut.edu.cn

Abstract

We present YOFO (**Y**ou **O**nly in**F**er **O**nce), a new paradigm for referring video object segmentation (RVOS) that operates in an one-stage manner. Our key insight is that the language descriptor should serve as target-specific guidance to identify the target object, while a direct feature fusion of image and language can increase feature complexity and thus may be sub-optimal for RVOS. To this end, we propose a meta-transfer module, which is trained in a learning-to-learn fashion and aims to transfer the target-specific information from the language domain to the image domain, while discarding the uncorrelated complex variations of language description. To bridge the gap between the image and language domains, we develop a multi-scale cross-modal feature mining block that aggregates all the essential features required by RVOS from both domains and generates regression labels for the meta-transfer module. The whole system can be trained in an end-to-end manner and shows competitive performance against state-of-the-art two-stage approaches.

1 Introduction

Referring video object segmentation (RVOS) aims to segment target objects from a video sequence according to the language referring expressions. In contrast to semi-supervised VOS (Perazzi et al. 2016) that requires a per-pixel mask to initialize target location, RVOS identifies the target relying only on an abstract language query. For one thing, RVOS inherently provides a more convenient choice for human-computer interaction, and thus attracts wide attention from the community (Khoreva, Rohrbach, and Schiele 2018; Seo, Lee, and Han 2020). For another, RVOS is also more challenging as it requires simultaneous interpretation of both visual and language modalities.

Though still in its infancy, RVOS has witnessed significant research progress in recent years, where they are mostly addressed in a two-stage pipeline. For instance, in the seminal work (Khoreva, Rohrbach, and Schiele 2018), RVOS is decomposed into referring bounding box tracking task followed by a bounding box segmentation step. In a more recent study, Seo *et al.* (Seo, Lee, and Han 2020) propose a unified segmentation approach, which first obtains initial masks guided by the referring expressions, and then performs semi-supervised VOS to produce the final results by propagating the most confident initial masks and referring expressions. Compared to their one-stage counterparts, two-stage RVOS methods can effectively improve the segmentation accuracy at the cost of higher computational overhead. It is then natural to ask whether we can combine the merits of both frameworks, *i.e.*, achieving the high accuracy of two-stage approaches while enjoying the efficiency of one-stage methods.

We make the first attempt towards this goal by proposing YOFO (**Y**ou **O**nly in**F**er **O**nce), a one-stage RVOS method. As opposed to prior art (Bellver et al. 2020) that uses the combined feature of the image and referring expression for segmentation, we conjecture that the direct feature fusion of

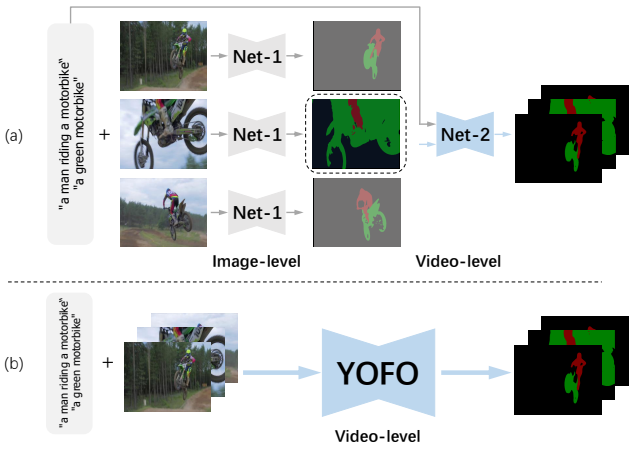

Figure 1: A comparison between (a) the two-stage method URVOS (Seo, Lee, and Han 2020) and (b) our one-stage YOFO. URVOS first obtains initial masks via referring image segmentation, and then produces the final segmentation results by conducting semi-supervised VOS. In contrast, our YOFO performs RVOS task in one stage and is end-to-end trainable.

*These authors contributed equally.
†Corresponding authors.

the two modalities may not be suitable for RVOS. Although language features contain target-specific cues, they also involve complex and diverse variations which is uncorrelated to the target. Directly combining the language and image features will inevitably introduce those extra noises from the language domain, leading to sub-optimal input features and thus inferior segmentation accuracy. Therefore, our essential philosophy is to extract target-specific cues from the referring expression which can then be transferred to the image feature domain, serving as guidance for target segmentation.

Motivated by the above observations, we design a Meta-Transfer (MT) module as the core component of our proposed YOFO method, which optimizes a parametric model to reconstruct the image and language features required for VOS using the image features solely. Since uncorrelated noises from language modality can hardly be restored using image features, they can be easily suppressed and only those essential for VOS are preserved and transferred to the output features. Since the above optimization process is also differentiable, the MT module can be learned in a learning-to-learn fashion, which ensures stronger generalization power across unseen scenarios and allows to aggregate the temporal information of video sequences, leading to more effective language feature transfer.

Another important design of this paper is a cross-modal feature mining module, which operates in a multi-scale manner and learns to enhance the essential cues from both image and language domains, providing the reconstruction target for the MT module. By incorporating the feature mining and MT module under our one-stage RVOS framework, the proposed YOFO method yields superior performance even compared to the two-stage counterparts.

The contribution of this paper can be summarized into three folds.

- We propose one of the first one-stage RVOS paradigm that can outperform state-of-the-art two-stage methods in accuracy.
- We present a meta-transfer module that ensures more effective target-specific feature learning and leverages temporal coherence for robust RVOS.
- We design a multi-scale cross-modal feature mining structure that is able to extract and integrate essential features required by RVOS from both image and language domains.

Experiments on two popular RVOS benchmarks have verified the effectiveness of our method.

2 Related Work

2.1 Referring Image Segmentation

Referring image segmentation (Hu, Rohrbach, and Darrell 2016; Yu et al. 2018; Ye et al. 2019; Huang et al. 2020) is to segment the object specified by the referring expression. Hu *et al.* (Hu, Rohrbach, and Darrell 2016) design a pioneering structure to extract visual and language features by CNN and LSTM independently. MattNet (Yu et al. 2018) introduces a language attention module and multi-modal sub-modules to extract the object information. CMPC (Huang et al. 2020) proposes a Cross-Modal Progressive Comprehension module and a Text-Guided Feature Exchange module to align features from both modalities. The recent work (Feng et al. 2021) develops an asymmetric co-attention (ACA) structure to fuse language features and image features, showing promising performance to understand the correlation of vision and language.

2.2 Video Object Segmentation

Video object segmentation (VOS) can be mainly categorized into unsupervised VOS and semi-supervised VOS. Unsupervised VOS aims to segment the most salient object in videos without any manual intervention (Li et al. 2018; Tokmakov, Alahari, and Schmid 2017; Zhou et al. 2020; Zhao et al. 2021). For the semi-supervised VOS (Caelles et al. 2017; Perazzi et al. 2017; Oh et al. 2019), the target objects to be segmented are given by the ground-truth mask in the first frame. Many recent works (Oh et al. 2019; Yang, Wei, and Yang 2020; Robinson et al. 2020; Bhat et al. 2020) deliver attractive performance in this field. For instance, STMNet (Oh et al. 2019) proposes a space-time memory network to utilize temporal information. CFBI (Yang, Wei, and Yang 2020) utilizes foreground-background matching and instance-level attention to reduce matching errors. (Bhat et al. 2020) proposes to cluster the object parts in the embedding space that is learned using meta-learning strategy, which is then used for segmentation prediction. The most related work to ours is LWLNet (Bhat et al. 2020). It captures the target information by integrating an optimization-based few-shot learner, which is learned by minimizing a regression error between the ground-truth information and features extracted by the image encoder. However, our work significantly differs LWLNet from at least two aspects. First, our Meta-Transfer module is designed to transfer the target-specific information from the language domain to the image domain and discard the uncorrelated variations of language description. Second, we propose a multi-scale cross-modal feature mining to integrate the target-specific information, which is unexplored in LWLNet.

2.3 Referring Video Object Segmentation

Referring video object segmentation (RVOS) (Bellver et al. 2020; Khoreva, Rohrbach, and Schiele 2018; Seo, Lee, and Han 2020) is a new sub-task of VOS, which introduces the language expression to specify the target object. Khoreva *et al.* develop the RVOS dataset called Ref-DAVIS (Khoreva, Rohrbach, and Schiele 2018) and design a two-stage approach for RVOS, which first performs the referring expression grounding and then utilizes the predicted bounding boxes to guide the pixel-wise segmentation. The recent work (Seo, Lee, and Han 2020) augments the Youtube-VOS dataset with referring expressions, and proposes a two-stage framework termed URVOS that first predicts the initial masks in image-level and then integrates the predicted masks and language into a semi-supervised VOS pipeline. In addition, there are some concurrent works (Gavrilyuk et al. 2018; Wang et al. 2019; Hui et al. 2021; McIntosh et al. 2020) to RVOS. They only focus on actors and actions in videos and are limited to a few target categories and

action-oriented descriptions. While the above methods deliver promising performance, they are mostly addressed in a two-stage pipeline (Khoreva, Rohrbach, and Schiele 2018; Seo, Lee, and Han 2020) at the cost of high computational overhead. In contrast, the proposed YOFO method performs RVOS in an one-stage paradigm and is able to outperform the state-of-the-art two-stage methods at a relatively lower computational complexity.

3 Method

3.1 Overview

As illustrated in Figure 2, the proposed YOFO segmentation method can be divided into five parts, including an image encoder, a language encoder, the multi-scale cross-modal feature mining (FM) module, the meta-transfer (MT) module and a decoder.

Given an input video sequence of T frames $\{\boldsymbol{I}^{(t)}\}_{t=1}^{T}$ and a language referring expression $\boldsymbol{Q}$ as query, our goal is to predict a sequence of segmentation masks $\{\boldsymbol{S}^{(t)}\}_{t=1}^{T}$ frame-by-frame to locate the referred target object. For each input frame, we first extract the image feature $\boldsymbol{X}$ and language feature $\boldsymbol{z}$ using the image and language encoders, respectively. The proposed cross-modal feature mining module integrates features from both modalities to produce a bi-modal representation $\boldsymbol{Y}$ for the current frame, which contains all the required information for target segmentation as well as uncorrelated noises from the language modality. Therefore, the MT module further distills the essential target cues from the bi-modal representation and transfers them to the image feature modality. Finally, the output feature $\tilde{\boldsymbol{Y}}$ from the MT module is fed into the decoder to generate the segmentation results. In the following, we present the details of the FM and MT modules in Section 3.2 and 3.3, respectively. Section 3.4 describes the training and inference details of our method.

3.2 Multi-Scale Cross-Modal Feature Mining

Our cross-modal feature mining (FM) module is designed to identify and integrate essential features from both language and image domains, which is then transferred using the MT module to the input features of the segmentation decoder. Though most existing methods for language and image feature fusion operate in a single scale, growing evidences (Ye et al. 2019) have shown that image feature fusion in a multi-scale manner is more beneficial in the sense of combining high-level coarse features with low-level fine-grained details. We hope our generated feature can also contain the multi-scale information for more accurate segmentation. Therefore, a baseline method is to firstly combine image and language features and then progressively fuse image features of multiple scales into the combined feature as shown in Figure 3(a). Nevertheless, this baseline only considers the scale variation of image features while failing to maintain the alignment between scale information conveyed by image features and the language. More importantly, the language information may be faded and overwhelmed by the continuously fused multi-scale image features.

Motivated by the above discussion, we propose to perform feature mining in a multi-scale manner, where image and language features are fused for each scale level, as illustrated in Figure 3(b). To this purpose, we adopt the outputs from the last three stages of ResNet50 (He et al. 2016) as the image feature representations which is denoted as $\{\boldsymbol{X}_i|i = 1, 2, 3\}$ with $i = 3$ indicating the coarsest-level feature. Meanwhile, we adopt the [CLS] representations $\{\boldsymbol{z}_i|i = 1, 2, 3\}$ generated by the last three Transformer decoder of BERT model (Vaswani et al. 2017; Devlin et al. 2018) as the corresponding language features. For the i-th scale, we first spatially tile the language feature vector $\boldsymbol{z}_i$ into a feature map $\boldsymbol{Z}_i$ that has the same spatial size as the corresponding image feature $\boldsymbol{X}_i$. The language and image features are aggregated with a fusion block $\mathcal{F}(\cdot, \cdot)$ to obtain a bi-modal feature representation $\boldsymbol{B}_i = \mathcal{F}(\boldsymbol{X}_i, \boldsymbol{Z}_i)$, which is then concatenated with feature map $\hat{\boldsymbol{B}}_{i-1}$ from the last scale. The concatenated feature is finally processed by a convolution layer with a stride of 2 to produce the output feature $\hat{\boldsymbol{B}}_i$ of the i-th scale, which is further sent to the next scale. Though a number of feature fusion strategies may be suitable for our purpose, we adopt the asymmetric co-attention mechanism proposed in (Feng et al. 2021) due to its simplicity and effectiveness. It first performs self-attention within each modality and then achieves cross-modal fusion through co-attention. We hope to explore other fusion mechanisms in our future work.

Given the combined feature $\hat{\boldsymbol{B}}_{\boldsymbol{3}}$ of the final scale, we use an additional convolution layer to produce the output bi-modal feature $\boldsymbol{Y}$. Since the output feature may inevitably contain noises, we employ a parallel branch with another convolution layer to infer a weight map $\boldsymbol{W}$ from $\hat{\boldsymbol{B}}_{\boldsymbol{3}}$, which has the same size of $\boldsymbol{Y}$ and attaches different weights to different positions and feature dimensions of $\boldsymbol{Y}$. As such, we can use the weight map to highlight target regions while suppressing noises.

Compared to the baseline method in Figure 3(a), the proposed multi-scale cross-modal feature mining scheme in Figure 3(b) is able to identify the matched scale of the image features, which is best aligned to the language cues. Therefore, it is more robust to scale variations of the target. Besides, the proposed scheme fuses the language features into the image features within each scale level, maintaining an appropriate balance between the two modalities.

3.3 Meta-Transfer with Augmented Memory

Given the bi-modal feature $\boldsymbol{Y}$ produced by the FM module, a naïve solution is to directly send it to the segmentation decoder to achieve the segmentation result. However, our preliminary experiments suggest that this naïve solution fails to deliver satisfactory results. One possible reason might be attributed to the over-complexity of the bi-modal feature. Although the bi-modal feature captures both the image and language features that are informative for referred target segmentation, it also contains uncorrelated features from the language modality, which may involve complex linguistic variations and can hardly benefit target segmentation. As a result, the over-complexity of the input feature increases the

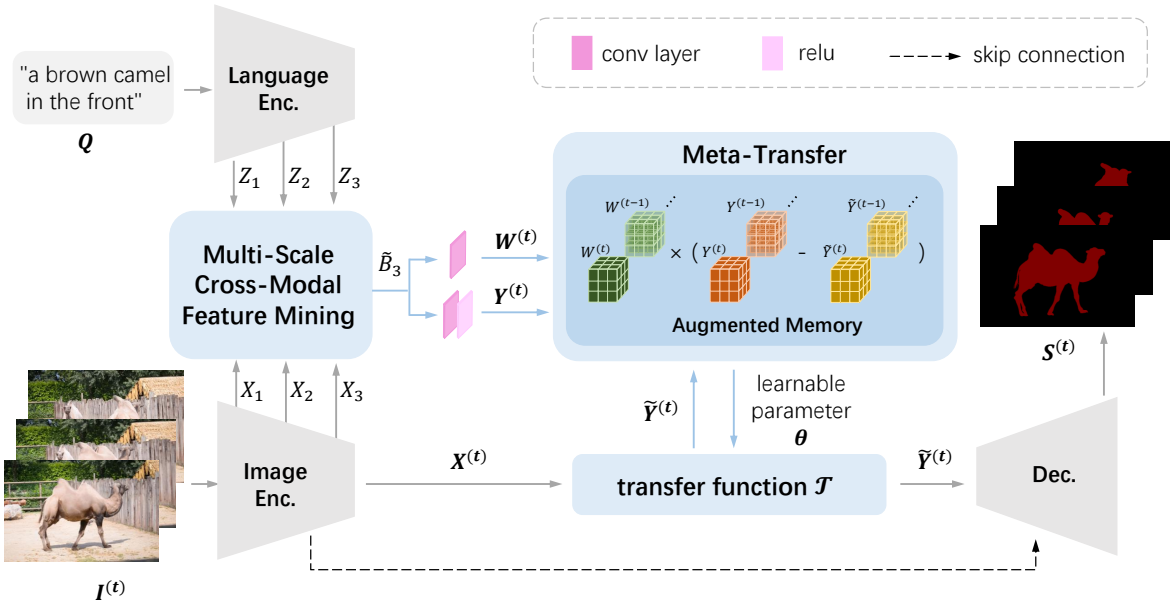

Figure 2: An overview of the proposed YOFO method. Given the referring language $\boldsymbol{Q}$ and the current frame $\boldsymbol{I}$, the multi-scale features of two domains are firstly extracted by the language encoder and image encoder, respectively, which is then fed into the proposed cross-modal feature mining module to generate the bi-modal feature $\boldsymbol{Y}^{(t)}$ and soft weight map $\boldsymbol{W}^{(t)}$. The Meta-Transfer module is learned to transfer the target-specific information from the bi-modal feature to the image domain, and the transferred image feature $\tilde{\boldsymbol{Y}}^{(t)}$ is sent to the decoder for the final prediction.

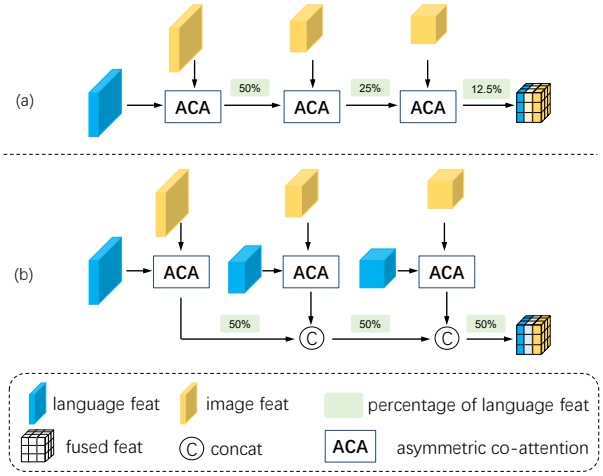

Figure 3: Illustration of two cross-modal feature mining approaches. (a) is the baseline method where the language information is gradually faded. (b) is our multi-scale cross-modal feature mining (FM) module that fuses image and language at multiple levels and avoids the imbalance between modalities. The percent numbers indicate the proportion of language information in the fused bi-modal features.

difficulty of decoder learning and also hinders its generalization ability.

To alleviate the above issue, we present the Meta-Transfer (MT) module, which aims to transfer the essential cues from the bi-modal feature to the image feature domain, while keeping the image feature under a reasonable complexity. Our assumption is that the image feature can represent the essential cues for referred target segmentation but can hardly characterize the redundant information from the language modality. Based on this assumption, our basic idea to achieve the above goal is to transfer the essential cues from the bi-modal feature through reconstruction using image feature solely. As such, the informative features will be preserved and uncorrelated or noisy ones will be discarded by the reconstructed feature.

To be specific, the feature transfer process can be formulated as a mapping function $\tilde{\boldsymbol{Y}} = \mathcal{T}(\boldsymbol{X}, \boldsymbol{\theta})$ with $\boldsymbol{X}$ denoting the input image feature generated by the image encoder and $\boldsymbol{\theta}$ the learnable parameter. We learn to reconstruct the bi-modal feature using image feature through the following objective function.

$$\arg\min_{\boldsymbol{\theta}} \sum_{k \in \mathcal{M}} \left\| \boldsymbol{W}^{(k)} \left(\boldsymbol{Y}^{(k)} - \mathcal{T}\left(\boldsymbol{X}^{(k)}, \boldsymbol{\theta} \right) \right) \right\|^2 + \lambda \left\| \boldsymbol{\theta} \right\|^2, \tag{1}$$

where $\mathcal{M}$ denotes an augmented memory comprising all the training frames, k denotes the frame index belonging to the memory, and λ indicates a hyper-parameter to balance the regularization term. $\boldsymbol{Y}^{(k)}$ and $\boldsymbol{W}^{(k)}$ represent the

bi-modal feature to be reconstructed and its corresponding weight map, respectively, both of which are generated by our cross-modal feature mining module. We leverage the weight map $\boldsymbol{W}^{(k)}$ to attach different importance weights to different elements of $\boldsymbol{Y}^{(k)}$. Given the learned parameter $\boldsymbol{\theta}$ and the current image feature $\boldsymbol{X}$, we compute the transferred image feature $\tilde{\boldsymbol{Y}} = \mathcal{T}(\boldsymbol{X}, \boldsymbol{\theta})$ and send it to the decoder to obtain the segmentation result.

Due to its simplicity, we adopt a linear convolution as the transfer function $\mathcal{T}$ with the parameter $\boldsymbol{\theta}$ being the convolutional kernel. One may learn the parameter $\boldsymbol{\theta}$ during offline training and fix $\boldsymbol{\theta}$ for online testing.

However, since different videos or language queries correspond to different targets, the essential cues for target segmentation varies across videos, requiring the transfer function also to be target-specific. Inspired by this fact, we adopt the steepest-descent method derived by (Bhat et al. 2020) to solve (1) and implement the learning process of $\boldsymbol{\theta}$ using differentiable operations. Learning $\boldsymbol{\theta}$ from (1) then becomes equivalent to the forward propagation of a differential function mapping the training samples $\{\boldsymbol{X}^{(k)}, \boldsymbol{Y}^{(k)}, \boldsymbol{W}^{(k)} | k \in \mathcal{M}\}$ to the optimal $\boldsymbol{\theta}$, which can be jointly optimized with the entire system. As a result, we are able to learn $\boldsymbol{\theta}$ for each input video during online inference, while the learning process of $\boldsymbol{\theta}$ can be learned during offline training, forming a meta-learning scheme, and hence the name Meta-Transfer.

Our Meta-Transfer module achieved through the above learning-to-learn strategy has two unique advantages. First, it permits the entire system to be end-to-end trainable, which ensures the cross-modal feature mining module to identify and produce essential features for more accurate referred target segmentation. Second, the Meta-Transfer parameters can be further adapted for the input video and language query during online learning, which delivers target-specific feature transfer and improves the generalization ability. In addition, the memory in (1) can be augmented with historical frames for online adapting the Meta-Transfer module, yielding more temporally consistent segmentation results. Our experiments show that the above advantages of our Meta-Transfer module can significantly benefit RVOS.

3.4 Implement Details

We empirically set the hyper-parameter λ in (1) to 0.01. The augmented memory $\mathcal{M}$ stores the recent N frames for learning the Meta-Transfer module. Our experiments show that a larger size of memory generally yields higher segmentation accuracy. However, the accuracy gain is marginal when the memory size is larger than 4. For both efficiency and effectiveness, we set the memory size N to 3.

We learn our method using the training sets of Refer-YouTube-VOS (Seo, Lee, and Han 2020), Refer-DAVIS2017 (Khoreva, Rohrbach, and Schiele 2018), and RefCOCO (Nagaraja, Morariu, and Davis 2016). Among them, RefCOCO is a referring image segmentation dataset and we use it to simulate video clips by random affine transformation. Ref-DAVIS2017 and Refer-Youtube-VOS are two popular RVOS datasets. Each object is annotated with two kinds of referring expressions for the *first* frame and *full* video. During training, we initialize the image encoder (*i.e.*, ResNet50 backbone) and the language encoder using the pretrained weights from (He et al. 2017) and (Devlin et al. 2018), respectively. At each iteration, we randomly sample 4 frames within a temporal window size of 100 from a training video, serving as the input to the network. Data augmentation techniques including color jittering, Gaussian blur and random erasing are also adopted to prevent overfitting. The whole network is end-to-end trained using the Lovasz segmentation loss (Berman, Triki, and Blaschko 2018). Adam optimizer (Kingma and Ba 2014) is adopted with a batch size of 4. We first train our network for 70 epochs by freezing the image and language encoders. The default learning rate is 2e-4 which decays by 0.2 in the 40^{th} epoch. Then the whole network is jointly trained for another 80 epochs. The default learning rate here is 2e-5 which decays by 0.2 in the 25^{th}, 75^{th} epoch. The proposed method runs at 10 FPS per object on NVIDIA 1080TI GPU, which has a good trade-off between efficiency and accuracy.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Referring expression: *first* frame			
Khoreva *et al.*	37.3	41.3	39.3
RefVOS	-	-	44.5
URVOS	47.29	55.96	51.63
YOFO	**50.14**	**58.74**	**54.44**
Referring expression: *full* video			
RefVOS	-	-	45.1
YOFO	**47.53**	**56.78**	**52.16**

Table 1: The quantitative evaluation of Refer-DAVIS2017 validation set.

4 Experiments

We first perform an overall comparison with state-of-the-art methods on the RVOS benchmark datasets, followed by the ablative studies to verify our main contributions.

4.1 State-of-the-art Comparisons

The proposed YOFO is compared with three recent state-of-the-art RVOS methods, including RefVOS (Bellver et al. 2020), URVOS (Seo, Lee, and Han 2020), and the method of Khoreva *et al.* (Khoreva, Rohrbach, and Schiele 2018) on both Ref-DAVIS2017 and Refer-Youtube-VOS benchmark datasets. Among them, RefVOS (Bellver et al. 2020) can be seen as a referring image segmentation method which performs object segmentation frame-by-frame without using temporal information. Both URVOS (Seo, Lee, and Han 2020) and Khoreva *et al.* (Khoreva, Rohrbach, and Schiele 2018) are two-stage methods.

The standard evaluation metrics (Perazzi et al. 2016) for VOS tasks are adopted, *i.e.*, region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and the average of $\mathcal{J}$ and $\mathcal{F}$ ($\mathcal{J}\&\mathcal{F}$).

Results on Ref-DAVIS2017. Table 1 shows the comparison results on Refer-DAVIS2017 validation set. The performance is evaluated by using the referring expressions of *first* frame and *full* video, respectively. Among the prior methods,

Figure 4: Qualitative comparison between refVOS (Bellver et al. 2020) and the proposed YOFO.

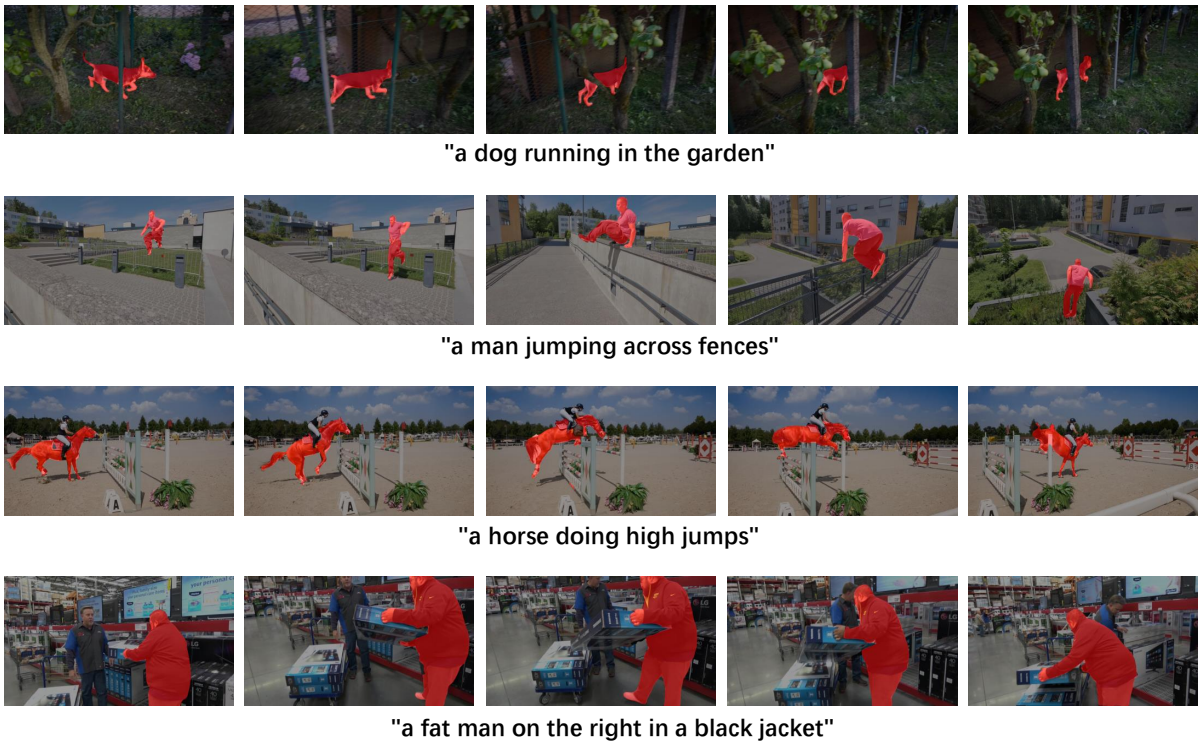

Figure 5: Some qualitative results of the proposed YOFO.

URVOS (Seo, Lee, and Han 2020) achieves the best scores. Nevertheless, the proposed one-stage method YOFO outperforms the two-stage URVOS by 2.81 in terms of $\mathcal{J}\&\mathcal{F}$.

Results on Refer-Youtube-VOS. Refer-Youtube-VOS is a recently developed dataset. We compare with URVOS, RefVOS and CMPC-V (Liu et al. 2021). We also report results of the first stage of URVOS (denoted as "URVOS-first") that predicts per-frame segmentation masks by using referring image segmentation pipeline. The evaluation results are provided in Table 2. The proposed YOFO shows the overwhelming superiority compared with URVOS-first and is competitive with the two-stage URVOS and CMPC-V.

Visualization. Figure 4 shows some visual comparison results. Obviously, the proposed YOFO can better distin-

Figure 6: The visualization of feature maps learned by YOFO with MT module (top) and the compared model "w/o MT" (bottom). Dlayer-i ($i \in [1, 2, 3, 4]$) denotes the input feature of the i-th layer of decoder. The target ground-truth masks are shown in the original images.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
RefVOS	39.5	-	-
URVOS-first	41.34	-	-
URVOS	45.27	49.19	47.23
CMPC-V	45.64	49.32	47.48
YOFO	**47.50**	**49.68**	**48.59**

Table 2: The quantitative evaluation of Refer-Youtube-VOS validation set.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
YOFO	**44.6**	**54.4**	**49.5**
w/ FM-baseline	37.4	53.0	45.2
w/ FM-SS	39.5	47.4	43.5
w/ FM-Cat	38.7	47.0	42.9
w/o MT	37.6	47.9	42.7
w/o memory	43.5	53.3	48.4

Table 3: Ablation experiments on Ref-DAVIS2017 with the referring expressions of *first* frame.

guish the target objects with various appearance, *e.g.* narrow shape (bike), fast motion (dancer), objects of the same category (dancer, judo), and small objects (gun). Besides, Figure 5 shows the temporal consistent segmentation results of YOFO for the occluded objects.

4.2 Ablation Experiments

We further analyze the impact of key components of the proposed YOFO by designing several ablation studies. To speed up the training procedure, we train the whole network of all the compared methods end-to-end without the pretraining stage. We summarize the comparison results in Table 3.

Cross-Modal Feature Mining: To validate the effectiveness of the proposed multi-scale cross-modal feature mining (FM) module, we compare it with three variants for the cross-modal feature mining. The baseline method named "w/ FM-baseline" is illustrated in Figure 3 (a), which firstly combines image and language features and then progressively fuses multi-scale image features into the combined one. The other variant is denoted as "w/ FM-SS", which only fuses the coarsest-level information of both image and language in a single scale. Besides, we also explore another feature fusion approach ("w/ FM-Cat"), which concatenates image and language features at each scale along the channel dimension. Table 3 shows that the proposed FM technique delivers the best segmentation performance.

Meta-Transfer module: One of our main contributions is that we design the Meta-Transfer (MT) module to transfer the essential cues from the bi-modal feature to the image feature domain. To verify its effectiveness, we remove the MT module and directly take the bi-modal feature as the input of the decoder. We denote this variant as "w/o MT". It shows a significant performance drop compared with YOFO. Figure 6 visualizes the feature maps learned by YOFO and "w/o MT". With the proposed meta-transfer strategy, YOFO can better identify the target regions. In contrast, when directly feeding the fused bi-modal feature for segmentation, the learned feature of "w/o MT" fails to distinguish the referred target from the distracting objects in background.

In our implementation, we employ the augmented memory $\mathcal{M}$ that stores the recent three frames for learning the Meta-Transfer module. We ablate this setting by removing the memory mechanism, meaning that we only use the current frame for meta-transfer learning. We name this variant as "w/o memory", whose performance is degenerated as expected. Nevertheless, it still surpasses the variant "w/o MT" by a large margin, which again demonstrates the effectiveness of our Meta-Transfer module.

5 Conclusion

We present a novel one-stage method YOFO for RVOS. We design a multi-scale cross-modal feature mining (FM) module to extract the essential information required by RVOS from both image and language domains. Our Meta-Transfer (MT) module is developed to transfer the target-specific feature from language domain through reconstruction using image feature solely and is trained in a learning-to-learn fashion. By integrating the proposed FM model and MT module into a unified one-stage framework, our YOFO can achieve outstanding results in the RVOS task.

6 Acknowledgments

The paper is supported by the National Key R&D Program of China under Grant No. 2018AAA0102001.

References

Bellver, M.; Ventura, C.; Silberer, C.; Kazakos, I.; Torres, J.; and Giro-i Nieto, X. 2020. Refvos: A closer look at referring expressions for video object segmentation. *arXiv preprint arXiv:2010.00263*.

Berman, M.; Triki, A. R.; and Blaschko, M. B. 2018. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4413–4421.

Bhat, G.; Lawin, F. J.; Danelljan, M.; Robinson, A.; Felsberg, M.; Van Gool, L.; and Timofte, R. 2020. Learning what to learn for video object segmentation. In *European Conference on Computer Vision*, 777–794. Springer.

Caelles, S.; Maninis, K.-K.; Pont-Tuset, J.; Leal-Taixé, L.; Cremers, D.; and Van Gool, L. 2017. One-shot video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 221–230.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, G.; Hu, Z.; Zhang, L.; and Lu, H. 2021. Encoder Fusion Network with Co-Attention Embedding for Referring Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15506–15515.

Gavrilyuk, K.; Ghodrati, A.; Li, Z.; and Snoek, C. G. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5958–5966.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from natural language expressions. In *European Conference on Computer Vision*, 108–124. Springer.

Huang, S.; Hui, T.; Liu, S.; Li, G.; Wei, Y.; Han, J.; Liu, L.; and Li, B. 2020. Referring image segmentation via cross-modal progressive comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10488–10497.

Hui, T.; Huang, S.; Liu, S.; Ding, Z.; Li, G.; Wang, W.; Han, J.; and Wang, F. 2021. Collaborative Spatial-Temporal Modeling for Language-Queried Video Actor Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4187–4196.

Khoreva, A.; Rohrbach, A.; and Schiele, B. 2018. Video Object Segmentation with Language Referring Expressions. In *ACCV*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S.; Seybold, B.; Vorobyov, A.; Lei, X.; and Kuo, C.-C. J. 2018. Unsupervised video object segmentation with motion-based bilateral networks. In *Proceedings of the European conference on computer vision (ECCV)*, 207–223.

Liu, S.; Hui, T.; Huang, S.; Wei, Y.; Li, B.; and Li, G. 2021. Cross-Modal Progressive Comprehension for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

McIntosh, B.; Duarte, K.; Rawat, Y. S.; and Shah, M. 2020. Visual-Textual Capsule Routing for Text-Based Video Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, 792–807. Springer.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9226–9235.

Perazzi, F.; Khoreva, A.; Benenson, R.; Schiele, B.; and Sorkine-Hornung, A. 2017. Learning video object segmentation from static images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2663–2672.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 724–732.

Robinson, A.; Lawin, F. J.; Danelljan, M.; Khan, F. S.; and Felsberg, M. 2020. Learning fast and robust target models for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7406–7415.

Seo, S.; Lee, J.-Y.; and Han, B. 2020. 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*, 208–223. Springer.

Tokmakov, P.; Alahari, K.; and Schmid, C. 2017. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, 4481–4490.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, H.; Deng, C.; Yan, J.; and Tao, D. 2019. 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*, 3939–3948.

Yang, Z.; Wei, Y.; and Yang, Y. 2020. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*, 332–348. Springer.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10502–10511.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1307–1315.

Zhao, X.; Pang, Y.; Yang, J.; Zhang, L.; and Lu, H. 2021. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In *ACM MM*, 2645–2653.

Zhou, T.; Wang, S.; Zhou, Y.; Yao, Y.; Li, J.; and Shao, L. 2020. Motion-attentive transition for zero-shot video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13066–13073.