Object-Agnostic Transformers for Video Referring Segmentation

Xu Yang, *Member, IEEE*, Hao Wang, De Xie, Cheng Deng, *Senior Member, IEEE*, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Video referring segmentation focuses on segmenting out the object in a video based on the corresponding textual description. Previous works have primarily tackled this task by devising two crucial parts, an intra-modal module for context modeling and an inter-modal module for heterogeneous alignment. However, there are two essential drawbacks of this approach: (1) it lacks joint learning of context modeling and heterogeneous alignment, leading to insufficient interactions among input elements; (2) both modules require task-specific expert knowledge to design, which severely limits the flexibility and generality of prior methods. To address these problems, we here propose a novel Object-Agnostic Transformer-based Network, called OATNet, that simultaneously conducts intra-modal and inter-modal learning for video referring segmentation, without the aid of object detection or category-specific pixel labeling. More specifically, we first directly feed the sequence of textual tokens and visual tokens (pixels rather than detected object bounding boxes) into a multi-modal encoder, where context and alignment are simultaneously and effectively explored. We then design a novel cascade segmentation network to decouple our task into coarse-grained segmentation and fine-grained refinement. Moreover, considering the difficulty of samples, a more balanced metric is provided to better diagnose the performance of the proposed method. Extensive experiments on two popular datasets, A2D Sentences and J-HMDB Sentences, demonstrate that our proposed approach noticeably outperforms state-of-the-art methods.**

***Index Terms*— Video referring segmentation, multi-modal learning, video grounding.**

I. Introduction

VIDEO has become one of the most popular forms of media today due to its ability to simultaneously characterize the static natural scene and the dynamic events it contains. With the explosive growth of video data over recent years, video understanding task has attracted ever-increasing attention in the computer vision community [1], [2]. However, traditional works tend to emphasize low-level vision understanding, such as video classification [3]–[7], action detection and localization [8], [9], and video object segmentation [10], [11], while failing to adequately interact with high-level semantics, e.g., human natural language. To understand the fine-grained actions performed by different actors in a video, Xu *et al.* [10] annotated the Actor-Action Dataset (A2D) with various actor-action pairs and introduced a challenging task of actor and action video segmentation. This task requires a comprehensive understanding of the static actors and dynamic actions in the video.

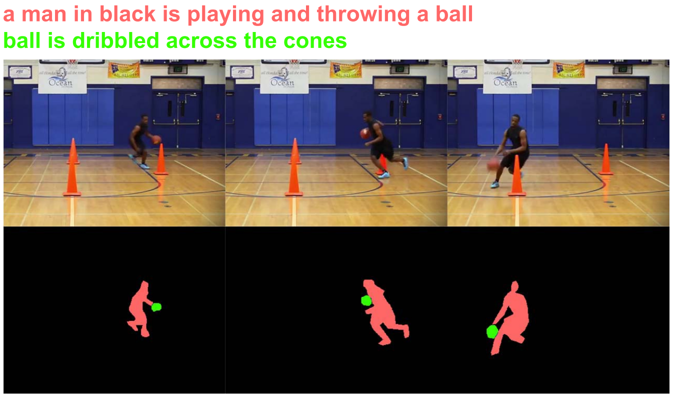

a man in black is playing and throwing a ball
ball is dribbled across the cones

Fig. 1. Given the textual description, video referring segmentation aims to generate a pixel-wise segmentation mask of the video object. The input sentence and colored mask share the same color if they are corresponding.

In an attempt to explore the intricate interactions between vision and language, Gavrilyuk *et al.* [12] augmented the video object segmentation dataset with corresponding queries and introduced the task of video referring segmentation, as illustrated in Figure 1. This task aims at selectively segmenting out the object related to the input textual description, which is extremely challenging, since it requires not only scene and language understanding, but also high-level semantic alignment among modalities. Moreover, video referring segmentation requires the joint learning of intra-modal context modeling and inter-modal alignment, which also does not rely on an object detector.

Previous works [12]–[17] have been explored for video referring segmentation. Following the paradigm of dynamic convolution in prior work [14], they learned the correlation

Manuscript received September 30, 2021; revised February 13, 2022; accepted March 8, 2022. Date of publication March 29, 2022; date of current version April 6, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0104100; in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, and Grant 62071361; in part by the Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03; and in part by the Fundamental Research Funds for the Central Universities under Grant ZDRC2102. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikolaos Mitianoudis. *(Corresponding author: Cheng Deng.)*

Xu Yang, Hao Wang, De Xie, and Cheng Deng are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: xuyang.xd@gmail.com; haowang.xidian@gmail.com; xiede.xidian@gmail.com; chdeng.xd@gmail.com).

Dacheng Tao is with the JD Explore Academy, Beijing 101111, China (e-mail: dacheng.tao@gmail.com).

Digital Object Identifier 10.1109/TIP.2022.3161832

between video objects and textual sentences by generating language-dependent convolutional filters for visual content. Wang *et al.* [15] adopted asymmetric cross-guided attention to aggregate visual context while Nin *et al.* [16] proposed a position-aware self-attention for this purpose, where linguistic information and the gating mechanism were taken into consideration simultaneously. Moreover, context modulated dynamic convolution [17] is proposed to incorporate spatial context during the interaction process. However, a majority of them [12]–[14], [17] simply utilize concatenation-convolution or dynamic filter [18] to align heterogeneous features, which neglects explicit intra-modal context modeling to help to capture complex linguistic knowledge and global visual relationships.

Recently, owning to the success of attention mechanisms [19] in natural language processing, some works have adopted gated self-attention to learn the intra-modal context on the visual modality with textual description but making less effort on language branch. We advocate that it requires joint intra-modal and inter-modal learning for vision and language tasks. For example, to segment the *blue car is parking on the left*, the pixels of the blue car on the left-not another color or location-should be focused on aggregating context information. Similarly, two words that are far from each other (e.g., *blue* and *left*) should be made a connection as they point to the same visual entity. Moreover, prior works adopt task-specific expert knowledge to devise intra-modal and inter-modal modules, an approach that greatly limits the flexibility and generality of their methods. For example, the convolutional operation for visual modality is inappropriate for social relationship data, meaning that necessary adaptation is required for the original intra-modal or inter-modal modules.

Multi-modal transformers have attracted an even-increasing attention in the context of vision and language tasks, e.g., visual question answering, visual grounding, and visual commonsense reasoning, as their powerful capability of long-range context modeling. Lu *et al.* [20] proposed ViLBERT by first applying transformers on texts and images independently to learn intra-modal interactions, then concatenated them to feed into another transformer to obtain inter-modal connections. Li *et al.* [21] devised a simple yet performant VisualBERT via utilizing a unified transformer for both vision and language. Unfortunately, most of these approaches heavily rely on the reliable object detector, which does not fit well with the practical setting in which the object categories are in an open set with new classes emerging.

In this paper, we propose a novel Object-Agnostic Transformer-based Network, named OATNet, for video referring segmentation without the object detection aids such as existing multi-modal transformers. It primarily comprises a multi-modal encoder and cascade segmentation network. More specifically, after extracting the features of the textual and visual tokens, the concatenation of them is fed into the multi-modal encoder to capture the intra-modal and inter-modal interactions, e.g., each element in the sequence can attend to the others of the same modality and different one. Then a cascade segmentation network is devised to decouple our task into coarse-grained segmentation and fine-grained refinement. Moreover, we propose a more balanced metric to analyze the experimental results by considering the difficulty of samples.

The main contributions of this work can be summarized as follows:

1) We propose an object-agnostic transformer-based network, in which the intra-modal and inter-modal interactions are simultaneously and effectively explored. Notably, while previous transformers used in image referring segmentation are based on object detection, ours operates directly on video pixels. Moreover, our method can easily be scaled to other modalities provided that they are processed as tokens.

2) We devise a novel cascade segmentation network of the multi-modal encoder, to decouple our task into coarse-grained segmentation and fine-grained refinement, which remarkably reduces the computational cost while maintaining acceptable performance.

3) Based on the difficulty of the samples, we present a novel metric for experimental results to help us analyze the performance from a more balanced perspective.

4) Experimental results on two popular video segmentation datasets demonstrate that our proposed approach significantly outperforms state-of-the-art methods.

II. Related Work

A. Actor and Action Video Segmentation

To understand the fine-grained actions performed by different actors in the video, Xu *et al.* [10] annotated the Actor-Action Dataset (A2D) with various actor-action pairs and introduced a challenging task of actor and action video segmentation. This task requires a comprehensive understanding of the static actors and dynamic actions in the video. Early works [10], [22] mainly adopted a graphical model to group the spatio-temporal information based on the supervoxel features. For example, Xu *et al.* [10] utilized separate classifiers for the actor, action and joint actor-action nodes to model the relationships among these nodes in the graph. An Interaction-Integrated Network [23], which contains a few Interaction-Integrated Cells, is designed to localize video clips according to a natural language description. Xu and Corso [22] adopted a grouping processing model to adaptively capture long-range interactions between video parts. Furthermore, Yan *et al.* [24] extended the task into the weakly supervised setting and proposed a robust multi-task ranking model to address it. Recently, deep learning has been successfully applied in many fields due to its powerful capability of feature extraction. Kalogeiton *et al.* [8] began to jointly learn the actor and its action detectors on top of the deep features of RGB and optical flow inputs, then segmentation on detected results was performed with existing methods. An efficient quantization parameter cascading technique [25] was also proposed for surveillance video coding. However, all these works above focus on low-level vision understanding, lacking interaction with high-level natural languages.

A Gaussian process embedded channel attention (GPCA) module [26] is proposed to model the correlations among the channels, which are assumed to be captured by beta distributed variables. Flow Edge-based Motion-Attentive Network (FEM-Net) [27] is designed to hallucinate edges of the ambiguous or missing region in the optical flow for the unsupervised video object segmentation problem. During the segmentation stage, the complementary temporal feature composed by the motion-attentive feature and flow edge is fed into a decoder to infer the salient foreground objects. In order to capture the temporal dependencies and gather information from multiple frames through bilateral temporal re-aggregation, Lin *et al.* [28] explored three schemes to build the aggregation, which can transfer the knowledge from a semi-supervised model to the weakly-supervised model without increasing the inference latency.

B. Video Referring Segmentation

To study the interaction between vision and language, Gavrilyuk *et al.* [12] collected corresponding sentences to describe the actor and its action in the video, and accordingly introduced the task of video referring segmentation. To overcome the limitations of traditional dynamic convolution, Wang *et al.* [17] proposed context modulated dynamic convolution to incorporate spatial context during the process of interaction. However, they neglected explicit intra-modal learning to capture complex linguistic knowledge and global visual relationship. Hence, inspired by the attention mechanism [19] in natural language processing, Yang *et al.* [29] adopted a two-stage paradigm to match the query and detected object bounding boxes. Different from the above works, our proposed approach takes advantage of a multi-modal transformer encoder to jointly capture intra-modal and inter-modal interactions, without using any object detection technique.

C. Multi-Modal Transformers

Based on the appealing performance of BERT [30], vision and language pre-training has become a nascent research area in computer vision community. Deng *et al.* [31] utilized a transformer to establish the multi-modal correspondence for visual-linguistic context interaction. Li and Sigal [32] proposed Referring Transformer to regress the bounding box and produce a segmentation mask simultaneously, which achieves superior performance. For video-language modalities, Sun *et al.* [33] proposed VideoBERT to extend the visual format to video data, which can be fine-tuned for downstream tasks, such as action classification and video captioning. Chen *et al.* [34] introduced UNITER mainly from the aspect of optimizing pre-training tasks, i.e., a conditional masking mechanism on masked language or masked region modeling. Toward learning more fine-grained cross-modal alignment, Huang *et al.* [35] proposed Pixel-BERT by changing the visual input from region-based image features to randomly selected (i.e., incomplete) pixel-level ones. By contrast, our proposed approach does not depend on object detection, which is end-to-end for training and evaluation. Moreover, it explores a complete (e.g., 1,024 v.s. 100) sequence of pixels and utilizes a novel cascade segmentation network to effectively address video referring segmentation. However, previous transformers used in vision and language tasks are based on object detection while ours operates on video pixels directly, which is non-trivial and needs more efforts to achieve the joint learning. Besides, the interactions, including intra-modal and inter-modal ones, among input elements become more fully explored when the number of stacked encoders increasing.

Our method simultaneously and effectively explores the joint intra-modal and inter-modal interactions, without the requirement of object detection like in UNITER [34] and other multi-modal transformers [36], [37], which can operate on video pixels directly and is naturally suitable for realistic scenarios. After extracting the features of the textual and visual tokens, the concatenation of them is fed into the multimodal encoder to capture the intra-modal and inter-modal interactions. In this way, each element in the sequence can attend to the others of the same modality and different one. Note that applying transformer-based methods to our task without the aid of object detection or category-specific pixel labeling is more challenging than most vision and language tasks.

III. Methodology

Given an input video $V = \{v_i\}_{i=1}^{T}$ with T frames and a corresponding natural language query $Q = \{q_i\}_{i=1}^{N}$ with N words, our approach is to segment out the object in the video described by the input textual query. The architecture of our method is illustrated in Figure 2, which consists of a multi-modal feature extractor, a multi-modal encoder and a cascade segmentation network.

A. Multi-Modal Feature Extractor

To effectively encode the static appearance and dynamic motion information of video data, we adopt a 3D convolution network to simultaneously capture video representations. With the emergence of many researches [4], [38], [39], the 3D convolution has been widely explored. Following previous works [12], [15], we utilize the Inflated 3D ConvNet (I3D) pre-trained on the Kinetics dataset [4] as a visual backbone, which is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn seamless spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters. Specifically, given a video clip $V \in \mathbb{R}^{3\times T\times H\times W}$, we obtain multi-scale outputs for subsequent multi-modal encoding and cascade segmentation, and formulate this procedure as follows:

$$V_S, V_M, V_L = \mathrm{Enc}_V(V; \theta_V), \tag{1}$$

where T, H and W represent the number of frames, the height and the width of each frame, respectively. Moreover, Enc_V denotes the video feature extractor parameterized with θ_V, while V_S, V_M and V_L are denoted as the outputs of small, medium and large scale. It is worth noting that the high-level but coarse-grained V_S is utilized for multi-modal interaction with textual features, which would introduce a far lower (i.e., $\frac{1}{256}$) computational cost when compared to directly adopting V_L.

Recently, a collaborative spatial-temporal encoder-decoder framework [12] was proposed that contains a 3D temporal

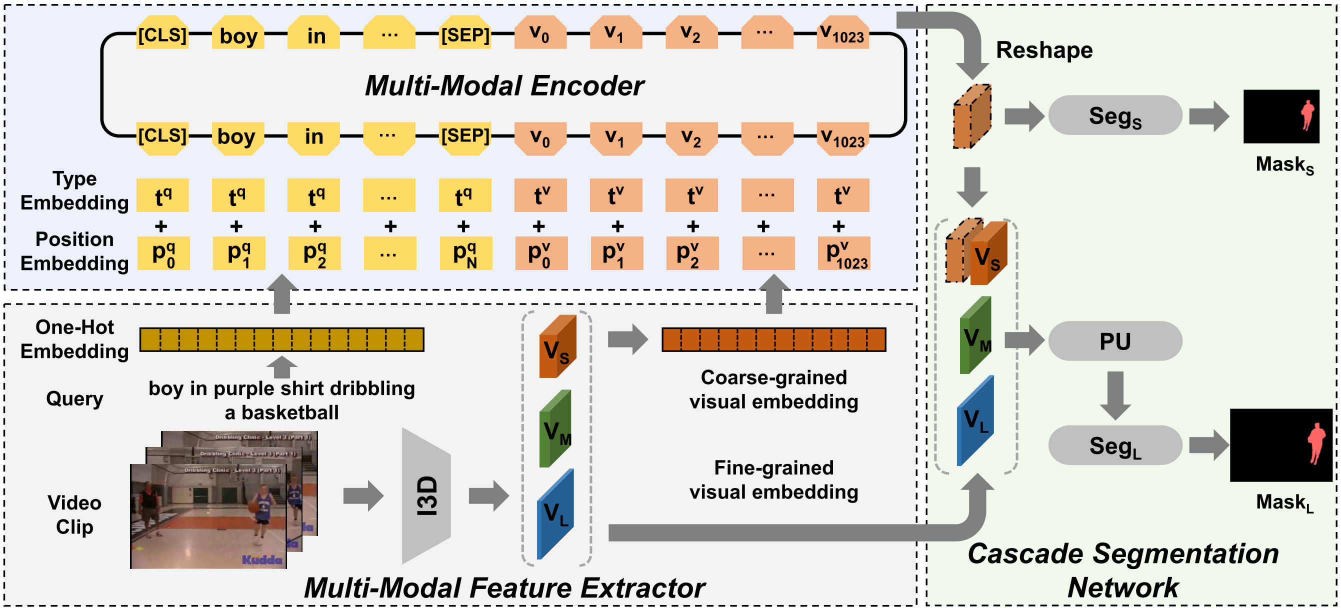


Fig. 2. The architecture of our proposed OATNet, which consists of multi-modal feature extractor, multi-modal encoder and cascade segmentation network. We first extract V_S (i.e., coarse-grained visual embedding), V_M, V_L (i.e., fine-grained visual embedding) and textual embedding (i.e., one-hot embedding) via multi-modal feature extractor. We then concatenate V_S and textual embedding to feed them into the multi-modal encoder to simultaneously capture the intra-modal and inter-modal interactions. Finally, we obtain the segmentation masks Mask_S by Seg_S and Mask_L through PU (i.e., progressive upsampling) and Seg_L.

encoder over the video clip that is propoed to recognize the queried actions, while a 2D spatial encoder over the target frame is utilized to accurately segment the queried actors. Following [12], we fix the input size as 512 × 512. We then select the temporally averaged outputs of I3D at the second, third and fourth stage as follows:

$$\begin{aligned} V_S &\in \mathbb{R}^{832\times 32\times 32}, \\ V_M &\in \mathbb{R}^{480\times 64\times 64}, \\ V_L &\in \mathbb{R}^{192\times 128\times 128}. \end{aligned} \tag{2}$$

To extract the word-level features in each sentence, we adopt a one-hot embedding method like that used in BERT rather than utilizing the vectors from the pre-trained word2vec model. More specifically, we tokenize the sentence into word-pieces [40], which consist of a deep LSTM network with eight encoder and eight decoder layers using residual connections as well as attention connections from the decoder network to the encoder, after which they employ an embedding matrix to embed each token into a vector. Formally, we can formulate the process as follows:

$$\widehat{Q} = \text{Enc}_Q(Q; \theta_Q), \tag{3}$$

where Enc_Q denotes the textual embedder parameterized with θ_Q, i.e., the learnable embedding matrix.

B. Multi-Modal Encoder

Before introducing the multi-modal encoder, we first revisit the architecture of the standard transformer [19] in natural language processing. The key operation of the transformer is self-attention, which is originally designed to capture the long-range relations of word tokens in each sentence. Concretely, given the input sequence $X \in \mathbb{R}^{N\times D}$, where N is the length of the sequence and D indicates its feature dimension. We first project input X into query X_Q, key X_K and value X_V by three matrices $W_Q \in \mathbb{R}^{D\times D}$, $W_K \in \mathbb{R}^{D\times D}$ and $W_V \in \mathbb{R}^{D\times D}$, respectively. The projection can be written as follows:

$$\begin{aligned} X_Q &= XW_Q, \\ X_K &= XW_K, \\ X_V &= XW_V. \end{aligned} \tag{4}$$

The attention output X_{att} is then calculated as follows:

$$X_{att} = \text{Softmax}(\frac{X_Q X_K^\top}{\sqrt{D}} + X_M)X_V, \tag{5}$$

where $X_M \in \mathbb{R}^{N\times N}$ is the self-attention mask, defined as:

$$(X_M)_{i,j} = \begin{cases} 0, & (X_Q)_i \text{ can attend to } (X_K)_j, \\ -\infty, & (X_Q)_i \text{ cannot attend to } (X_K)_j. \end{cases} \tag{6}$$

It is used to ignore the attention score calculated between the textual or visual token and the useless padding token.

We next adopt the transformer discribed above to process multiple modalities in our task. Specifically, we flatten the video features V_S as visual tokens and concatenate the textual ones $\widehat{Q}$ to feed them into the multi-modal transformer encoder, along with the positional encoding and type embedding as in BERT. This input sequence X can be defined as follows:

$$\begin{aligned} &\{[\text{CLS}], q_0^{in}, \cdots, q_N^{in}, [\text{SEP}], v_0^{in}, \cdots, v_{1023}^{in}\}, \\ &q_i^{in} = \text{LayerNorm}(\hat{q}_i + p_i^q + t^q), \\ &v_i^{in} = \text{LayerNorm}((v_s)_i + p_i^v + t^v), \end{aligned} \tag{7}$$

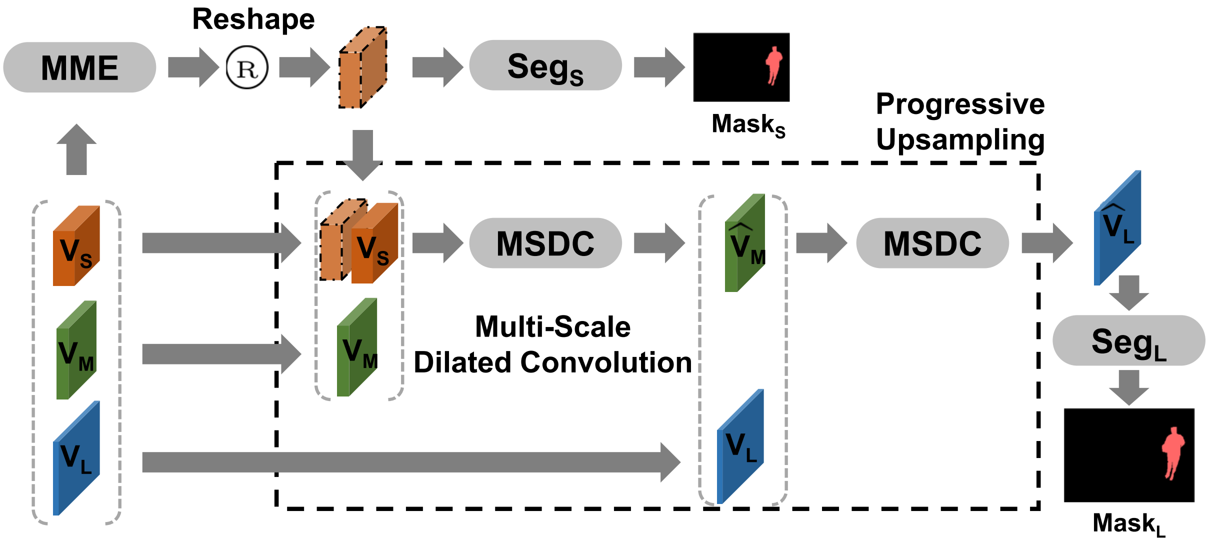


Fig. 3. The illustration of cascade segmentation network, which consists of a coarse-grained segmentation for heterogeneous interaction and a fine-grained refinement to capture more visual details. We first conduct the segmentation on high-level but coarse-grained V_S^{out}, which already learns interaction with corresponding language queries. Then, the progressive upsampling first upsamples the $\widehat{V_S}$ to the same size of V_M, then the concatenation of $\widehat{V_S}$ and V_M are fed into multi-scale dilated convolution (MSDC) to conduct fusion. Finally, $\widehat{V_L}$ is obtained with multi-modal interaction information and rich visual details for subsequent segmentation.

where [CLS] and [SEP] are special tokens for global cross-modal matching (not used in our task) and modality separation, respectively. p_i^q and p_i^v denote the positional embedding of textual and visual tokens. The former is the embedding indexed by the word order in a sentence, while the latter is the spatial location of each pixel, e.g., normalized coordinate (x, y) on a 2D feature map. Besides, t^q and t^v are type embeddings that explicitly indicate the modality category. Through stacking multiple encoders, intra-modal and inter-modal interactions are sufficiently explored. Finally, we obtain the output and formulate it as follows:

$$\{[\text{CLS}], q_0^{out}, \cdots, q_N^{out}, [\text{SEP}], v_0^{out}, \cdots, v_{1023}^{out}\}. \tag{8}$$

Since our task focuses on segmentation, we simply extract the visual part of the output and reshape it to the grid feature map. To simplify the description, we can define the multi-modal encoder as follows:

$$V_S^{out} = \text{MME}(V_S, \hat{Q}; \theta_{MME}), \tag{9}$$

where MME denotes the multi-modal encoder parameterized with θ_{MME}.

C. Cascade Segmentation Network

It is well known that the computation cost of self-attention scales with $\mathcal{O}(N^2)$, which means that V_L is not appropriate for multi-modal interaction with a textual query. However, V_L contains much visual details to significantly contribute the segmentation. Hence, we propose a novel cascade segmentation network, as illustrated in Figure 3, to decouple our task into coarse-grained segmentation and fine-grained refinement. Concretely, we first conduct segmentation on high-level but coarse-grained V_S^{out}, which already learns interaction with corresponding language queries. The segmentation mask M_S can be formulated as follows:

$$M_S = \sigma(\text{Seg}_S(V_S^{out})), \tag{10}$$

where Seg_S and σ denote the segmentation network and Sigmoid activation, respectively. To utilize visual details for better segmentation, we incorporate them via progressive upsampling and describe it as follows:

$$\begin{aligned}\widehat{V_S} &= [V_S, V_S^{out}],\\ \widehat{V_L} &= \text{PU}(\widehat{V_S}, V_M, V_L),\\ M_L &= \sigma(\text{Seg}_L(\widehat{V_L})),\end{aligned} \tag{11}$$

where Seg_L and PU$(\cdot)$ represent the refinement network and progressive upsampling, respectively. Specifically, the progressive upsampling first upsamples the $\widehat{V_S}$ to be the same size of V_M, after which the concatenation of $\widehat{V_S}$ and V_M are fed into multi-scale dilated convolution (MSDC) to conduct fusion. Finally, $\widehat{V_L}$ is obtained with multi-modal interaction information and rich visual details for subsequent segmentation.

For simplicity, we summarize the cascade segmentation network as follows:

$$M_S, M_L = \text{CSN}(V_S^{out}, V_S, V_M, V_L; \theta_{CSN}), \tag{12}$$

where CSN stands for the cascade segmentation network parameterized with θ_{CSN}.

D. Training and Inference

Given an input video clip V, corresponding natural language query Q and multi-scale ground-truth segmentation masks Y_S and Y_L, our proposed method generates the multi-scale predictions M_S and M_L. Therefore, we can formulate the overall loss as follows:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_L. \tag{13}$$

Specifically, each term outlined above is calculated as follows:

$$\mathcal{L}_r = -Y_r \log(\sigma(M_r)) - (1 - Y_r)\log(1 - \sigma(M_r)), \quad (14)$$

where $r \in \{S, L\}$, while σ denotes Sigmoid activation. During inference, we take a pixel as foreground when its prediction value exceeds half of the maximum value in the entire response map.

IV. Experiment

In this section, we first provide the dataset statistics and implementation details used in all experiments. Next, we compare our proposed approach with existing state-of-the-art methods to demonstrate the superiority of our method. Finally, we present quantitative analysis of the proposed model and visualization of experimental results.

A. Dataset Statistics

A2D Sentences [12]: is annotated and released by augmenting the original video dataset [10] with 6,655 natural language descriptions, for video referring segmentation. Statistically, there are 3,782 videos collected from YouTube, which includes 8 action classes and 7 actor classes. Specifically, the actions comprise climbing, crawling, eating, flying, jumping, rolling, running and walking. The actors include adult, baby, ball, bird, car, cat and dog. In each video, three to five frames are labeled with pixel-level masks for training and evaluation. Following [12], the dataset is divided into 3,017 training videos, 737 testing videos and 28 unlabeled videos.

J-HMDB Sentences: is also annotated through augmenting original dataset [41] with 928 natural language descriptions. The pixel-level ground-truth is a 2D articulated human puppet mask. It is worth noting that each sample of J-HMDB Sentences contains only one salient visual object, which means it is relatively easier to be segmented than the one of A2D Sentences.

Following prior work [12], [15], we adopt the popular criteria of Intersection-over-Union (IoU) and precision to evaluate the segmentation performance. In more detail, we utilize two kinds of IoU: mean IoU and overall IoU. The former first computes the IoU of each sample and then averages the results on the whole dataset. The latter is obtained by calculating the ratio of the total intersection area divided by the total union area on the entire dataset. Besides, the P@t computes the percentage of the testing samples (i.e., sentence-clip pairs) whose IoU scores are higher than threshold t, while the mean average precision (mAP) reports the averaged results over various thresholds from 0.5 to 0.95 with step 0.05.

Moreover, as illustrated in Figure 4, we observe that there exists easy examples in both datasets. The easy sample only contains one salient visual object while the hard sample contains more than one object. Till now, only A2D Sentences and J-HMDB Sentences have been publicly used for video referring segmentation. We advocate that extra metrics are required to better diagnosed the model performance, based on the sample difficulty. Hence, we propose the harmonic mean H@t metric to calculate the more balanced results of precision:

$$\mathrm{H@t} = \frac{2 * (\mathrm{P@}t)_{\mathrm{easy}} * (\mathrm{P@}t)_{\mathrm{hard}}}{(\mathrm{P@}t)_{\mathrm{easy}} + (\mathrm{P@}t)_{\mathrm{hard}}}, \quad (15)$$

which indicates that our goal is high precision value on both easy and hard samples.

B. Comparison Methods

We demonstrate the results of video referring segmentation compared with five approachs [12], [15]–[17], [29] that have adopted the same task.

1) Gavrilyuk *et al.* [12] collected corresponding sentences to describe the actor and its action in the video, and introduced the video referring segmentation task.
2) Wang *et al.* [15] adopted asymmetric cross-guided attention to aggregate visual context.
3) Nin *et al.* [16] proposed a position-aware self-attention to aggregate visual context, where linguistic information was simultaneously taken into consideration with the gating mechanism.
4) Wang *et al.* [17] proposed context modulated dynamic convolution to incorporate spatial context during the interaction process.
5) Yang *et al.* [29] adopted a two-stage paradigm to match the query and detect object bounding boxes.

C. Implementation Details

For multi-modal feature extractors, we adopt the I3D model, pre-trained on the Kinetics dataset, to extract video features by following [12] and vocabulary embedding matrix of pre-trained BERT [30] to obtain textual features. Concretely, V_S, V_M and V_L are temporally averaged outputs of I3D at the second, third and fourth stage, respectively. For the multi-modal encoder, we utilize the first 3 layers of the pre-trained 12-Layer BERT model to prevent the over-fitting on our datasets. For cascade segmentation network, we employ single fully convolutional layer for coarse-grained segmentation network and two fully convolutional layers (the hidden size is 128) for fine-grained refinement network. The approaches marked by †, $\diamond$ and $*$ fine-tune the layer mixed_4f, the layers from mixed_4b to mixed_4f and all layers of I3D, respectively.

We utilize PyTorch [42] package for all experiments in this paper. For optimizer, we adopt AdamW [43] with the learning rate 0.0005 and weight decay 0.01. The batch size is 10 and total training step is 20,000. We divide the learning rate by 10 after 14,000 steps. For input clip, the number of frames is 8 and the height or width are 512, respectively. The annotated frame are fixed in the middle of sampled clip. We only take RGB frames as input, instead of using RGB and optical flow clips like prior work [12].

D. Comparison With State-of-the-Art Methods

We demonstrate the results of video referring segmentation compared with several state-of-the-art methods in Table I. Previous works [13], [14] have two settings: methods are trained only on ReferIt dataset [44] without any fine-tuning on A2D Sentences and fine-tuning the models on the training samples of the target dataset. The methods of the former setting are shown in the first two rows, and the methods of the latter setting are marked with †. Obviously, the more layers

TABLE I

Segmentation Results on A2D Sentences. The Approaches Marked by †, ◇ and ∗ Fine-Tune the Layer mixed_4f, the Layers From mixed_4b to mixed_4f and All Layers of I3D on A2D Sentences, Respectively. The Method Marked by ○ Adopts Two-Stage Paradigm. It Should Be Noted That the Results of [12] Are Obtained on Two Streams - RGB and Optical Flow While the Others Only Take RGB Frames as Input

Method	Overlap					mAP	IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
Hu *et al.* [13]	7.7	3.9	0.8	0.0	0.0	2.0	21.3	12.8
Li *et al.* [14]	10.8	6.2	2.0	0.3	0.0	3.3	24.8	14.4
Hu *et al.* [13] †	34.8	23.6	13.3	3.3	0.1	13.2	47.4	35.0
Li *et al.* [14] †	38.7	29.0	17.5	6.6	0.1	16.3	51.5	35.4
Gavrilyuk *et al.* [12] †	50.0	37.6	23.1	9.4	0.4	21.5	55.1	42.6
Wang *et al.* [15] †	55.7	45.9	31.9	16.0	2.0	27.4	60.1	49.0
OATNet †	67.3	61.3	51.5	33.3	8.1	40.9	66.1	57.2
Gavrilyuk *et al.* [12] ◇	53.8	43.7	31.8	17.1	2.1	26.9	57.4	48.1
Wang *et al.* [17] ◇	60.7	52.5	40.5	23.5	4.5	33.3	62.3	53.1
OATNet ◇	66.7	61.3	51.4	32.9	8.5	40.8	65.3	57.2
Ning *et al.* [16] ∗	63.4	57.9	48.3	32.2	8.3	38.8	66.1	52.9
Yang *et al.* [29] ○	68.1	62.9	52.3	29.6	2.9	39.6	61.7	55.2
OATNet ∗	**69.3**	**63.7**	**54.9**	**38.4**	**11.3**	**43.9**	**67.4**	**59.0**

TABLE II

The Generalization Ability of Each Method on J-HMDB Sentences With the Model Trained on A2D Sentences. The Methods Marked by †, ◇ and ∗ Fine-Tune the Layer mixed_4f, the Layers From mixed_4b to mixed_4f and All Layers of I3D on A2D Sentences, Respectively. The Method Marked by ○ Adopts Two-Stage Paradigm

Method	Overlap					mAP	IoU	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
Hu *et al.* [13] †	63.3	35.0	8.5	0.2	0.0	17.8	54.6	52.8
Li *et al.* [14] †	57.8	33.5	10.3	0.6	0.0	17.3	52.9	49.1
Gavrilyuk *et al.* [12] †	69.9	46.0	17.3	1.4	0.0	23.3	54.1	54.2
Wang *et al.* [15] †	75.6	56.4	28.7	3.4	0.0	28.9	57.6	58.4
OATNet †	83.2	71.6	47.1	11.5	0.0	38.4	63.4	62.9
Gavrilyuk *et al.* [12] ◇	71.2	51.8	26.4	3.0	0.0	26.7	55.5	57.0
Wang *et al.* [17] ◇	74.2	58.7	31.6	4.7	0.0	30.1	55.4	57.6
OATNet ◇	**87.0**	74.3	49.4	12.0	0.0	40.0	64.5	65.0
Ning *et al.* [16] ∗	69.0	57.2	31.9	6.0	**0.1**	29.4	-	-
Yang *et al.* [29] ○	77.3	62.7	36.0	4.4	0.0	32.1	58.3	57.6
OATNet ∗	86.9	**75.7**	**53.0**	**13.2**	**0.1**	**41.2**	**65.2**	**65.3**

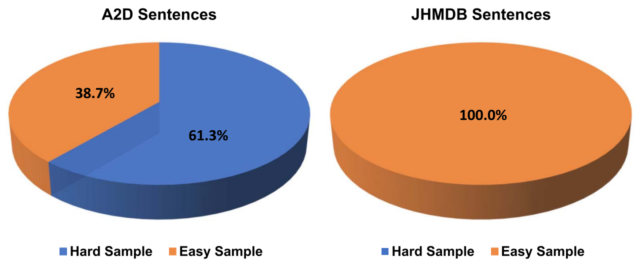


Fig. 4. The proportion of easy samples and hard ones on A2D Sentences and JHMDB Sentences respectively.

fine-tuned, the better segmentation performance. For our proposed approach, we observe that fine-tuning on the same stage (mixed_4f or from mixed_4b to mixed_4f) obtains similar performance, which demonstrates that low-level video features (i.e., mixed_3b, mixed_3c and etc) play more crucial role in pixel-wise semantic segmentation. Our proposed approach achieves remarkable improvement at higher IoU thresholds, such as precision metrics ‘P@0.8’ and ‘P@0.9’, which reflects the advantages of our method compared with existing state-of-the-art approaches [12], [15]–[17]. Moreover, we bring 6.1% absolute improvement in Mean IoU, 1.3% in Overall IoU, and 5.1% in mAP over state-of-the-arts, respectively. Our approach only takes RGB frames as video inputs, without using any additional motion information (i.e., optical flow as in [12]). Although the visual backbone (i.e., I3D v.s. Faster R-CNN) is different from the two-stage work [29], our approach beats it on all metrics, showing the effectiveness and potential of the proposed method.

To evaluate the generalization ability of our proposed approach, we adopt the model pre-trained on A2D Sentences

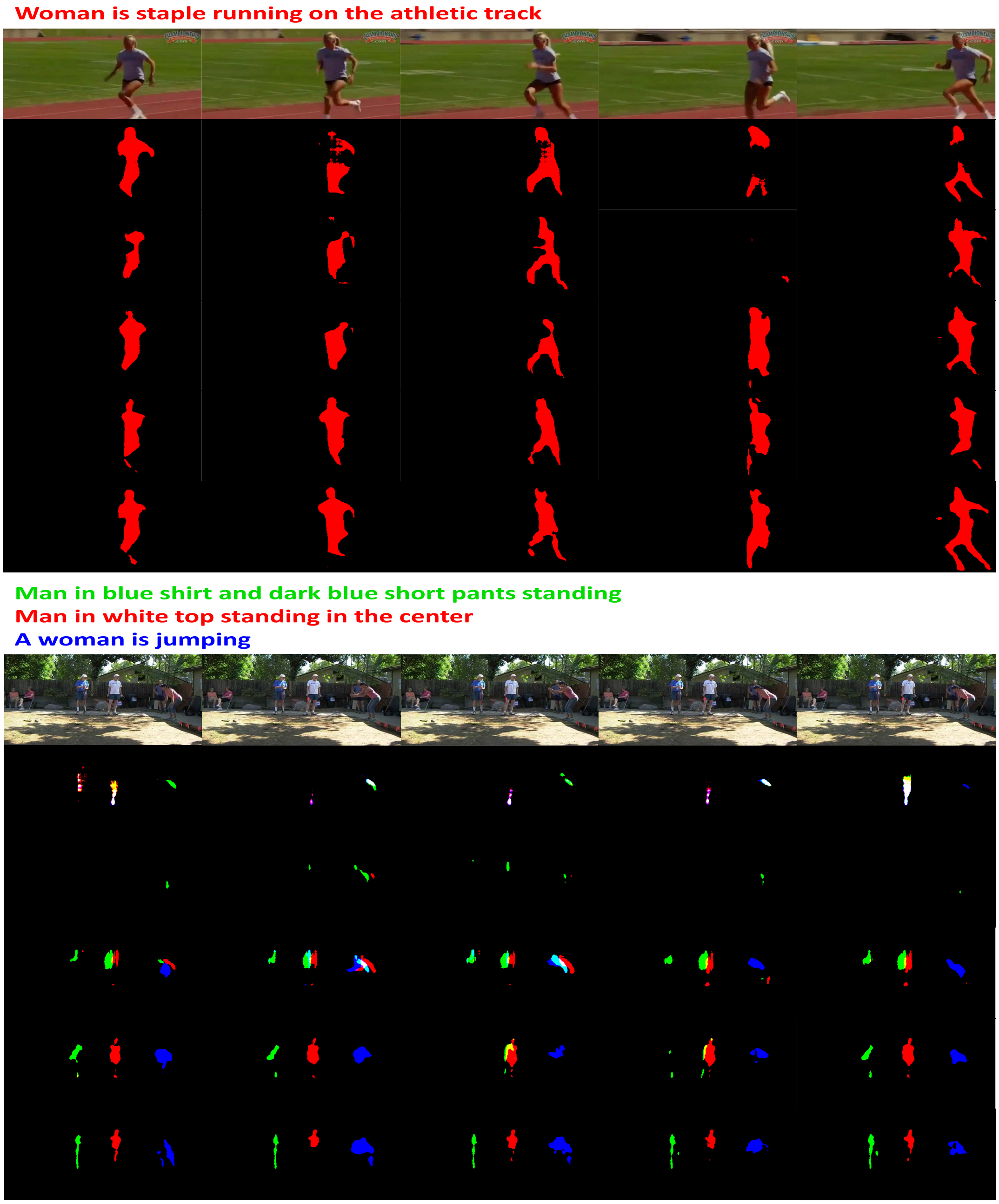

Woman is staple running on the athletic track
Man in blue shirt and dark blue short pants standing
Man in white top standing in the center
A woman is jumping

Fig. 5. Qualitative results on A2D Sentences. The first and second rows show the input textual descriptions and the frames of input videos. The third row illustrates the segmentation results from **Baseline**. The fourth row and the fifth row are the segmentation outputs of **Baseline + MME (scratch)** and **Baseline + MME (pre-trained)**, respectively. The last two rows are the results of pre-trained MME with CSN, where the top row is our model without V_L and the bottom is our full model. Both of them are trained on RGB frames for a fair comparison. The colored masks correspond to the sentences with the same color on the top of each video. Some overlaps are a mixture of colors.

to segment all samples on J-HMDB Sentences without any additional fine-tuning. During the evaluation, we uniformly sample 3 frames of each testing video as indicated in [12]. The segmentation results are reported in Table II. Our method obtains obvious improvement on most metrics, including the hardest on 'P@0.9'. Besides, the more layers fine-tuned on A2D Sentences, the better segmentation performance can be achieved.

In addition, we observe that the segmentation results of Ning *et al.* [16] on Overall IoU and P@t are high but low on Mean IoU. We suspect that their method prefers segmenting well on easy samples. Hence, for the first time, we present an extra analysis to determine the performance improvement whether from easy samples (i.e., only having one salient object in the video) or hard ones, as illustrated in Table III, where Easy and Hard represent the P@t of Easy and Hard samples compared with all testing samples. Moreover, we adopt a harmonic mean to compute the more balanced results of precision, which reveals the 'real' performance of the method.

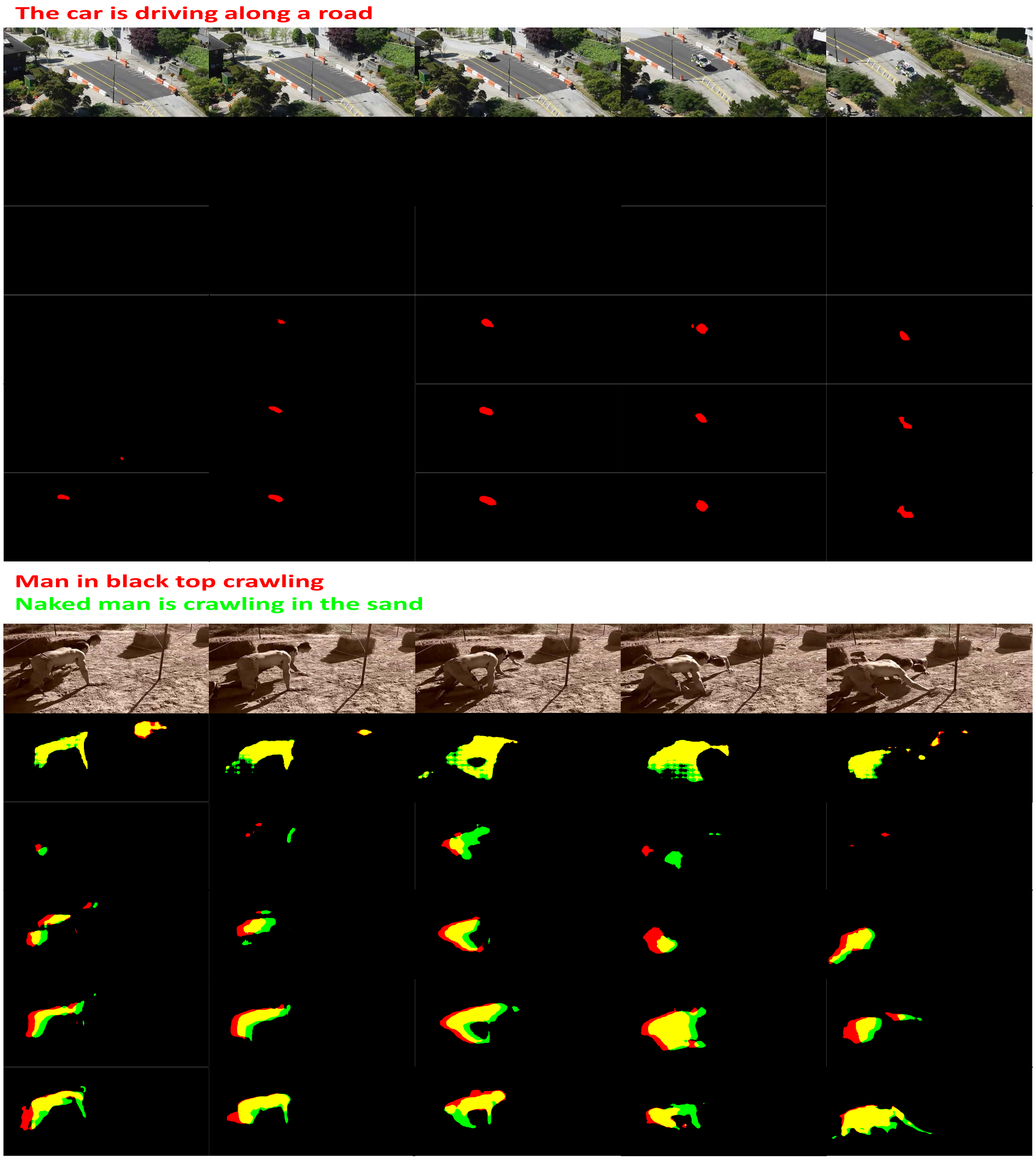

The car is driving along a road
Man in black top crawling
Naked man is crawling in the sand

Fig. 6. Qualitative results on A2D Sentences. Our method can also achieve competitive results when there are moving small objects or occlusions.

TABLE III

The Testing Samples Are Divided Into Easy Samples and Hard Ones. H@t Represents the Harmonic Mean Value of Them

	Overlap				
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9
Easy	35.6	34.1	30.6	23.3	7.9
Hard	33.7	29.6	24.3	15.1	3.4
H@t	34.6	31.6	27.0	18.3	4.7

We have tested the inference performance of our approach, which achieves 29.46 FPS via adopting the technique of mixed precision. Concretely, the model takes the raw sequence of video frames and sentences as input and outputs the segmentation result with the same size as the input frame.

E. Ablation Studies

To further verify the effectiveness of each component, we conduct ablation studies and show the results in Table IV.

Baseline: first utilizes I3D and GRU [45] to extract features and then follows the traditional way of concatenation-convolution to obtain the final segmentation result.

Baseline + MME: adopts a multi-modal encoder to extract linguistic features as well as perform a heterogeneous alignment. We can observe that merely replacing GRU with **MME (scratch)**, i.e., multi-modal encoder without pre-training, significantly improves the performance. Besides, further improvement can be observed when employing **MME (pre-trained)**, i.e., the encoder pre-trained on a text corpus. Here, the proposed multi-modal encoder obtains excellent performance on our task and is more flexible than prior methods since it only requires tokenizing the input of various modalities, e.g., natural language queries and pixels of the visual feature map.

Furthermore, our full model, i.e., **Baseline + MME (pre-trained) + CSN**, achieves state-of-the-art performance on both datasets. It is worth noting that with the addition of

TABLE IV

SEGMENTATION RESULTS ON A2D SENTENCES FOR ABLATION STUDIES. HERE, MME AND CSN DENOTE MULTI-MODAL ENCODER AND CASCADE SEGMENTATION NETWORK, RESPECTIVELY. THE PRE-TRAINED OR SCRATCH INDICATES WHETHER THE BERT IS TRAINED OR NOT ON TEXT CORPUS

Method	**Overlap**					**mAP**	**IoU**	
	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
Baseline	49.3	40.3	30.4	16.1	2.0	25.1	53.9	46.0
Baseline + MME (scratch)	63.6	56.8	45.1	27.0	5.7	36.4	63.4	54.1
Baseline + MME (pre-trained)	66.1	60.0	48.7	29.3	6.0	38.8	65.7	56.4
Baseline + MME (pre-trained) + CSN $w/o\ V_L$	67.5	62.2	52.1	35.3	9.8	41.9	67.1	57.9
Baseline + MME (pre-trained) + CSN	**69.3**	**63.7**	**54.9**	**38.4**	**11.3**	**43.9**	**67.4**	**59.0**

multi-scale features, the model performance steadily improves. It also demonstrates that modeling visual details greatly improves segmentation performance.

Qualitative results on A2D Sentences are presented in Figure 5, where the first and second rows show the input textual descriptions and the frames of input videos. The fourth row and the fifth row are the segmentation outputs of **Baseline + MME (scratch)** and **Baseline + MME (pre-trained)**, respectively. The last two rows are the results of pre-trained MME with CSN, where the top row is **Baseline + MME (pre-trained) + CSN** $w/o\ V_L$ and the bottom is our full model. Both of them are trained on RGB frames for a fair comparison. The colored masks correspond to the sentences with the same color on the top of each video. Some overlaps are a mixture of colors. Concretely, the **Baseline** performs fairly on the simple sample with one actor, however, when there are multiple actors, the **Baseline** cannot even locate the correct target. The model's ability to locate the correct actors is significantly enhanced with **Baseline + MME (pre-trained)**, which is also an important reason for the significant improvement in our performance. Meanwhile, the results have demonstrated the multi-modal encoder can effectively capture the intra-modal interactions for contextual feature learning and inter-modal ones for heterogeneous alignment.

More importantly, we can see that our model can generate fine-grained segmentation of actors or objects after the cascade segmentation network, i.e., legs and arms of the actor in Figure 5, which shows that the proposed network can effectively conduct coarse-grained segmentation and fine-grained refinement, especially for the scene with three actors. Finally, we also select some complex scenes to evaluate the proposed method and the results shown in Figure 6. For the small moving car and occluded targets, traditional methods cannot locate the targets accurately, but our method can still achieve satisfactory results.

V. CONCLUSION

In this paper, we have proposed a novel object-agnostic transformer-based network, OATNet, to address the emerging task of video referring segmentation. Our approach simultaneously handles the intra-modal context modeling and inter-modal heterogeneous alignment. As a result, it significantly improves segmentation performance with the less expert knowledge required to design for better flexibility and generality. Moreover, our method can be easily scaled to other modalities as long as they are processed as tokens. To reduce the computation cost while maintaining acceptable performance, we present a novel cascade segmentation network atop the multi-modal encoder to decouple our task into coarse-grained segmentation and fine-grained refinement. We evaluate our proposed approach on two popular video segmentation datasets, and the results demonstrate that it significantly outperforms state-of-the-art methods.

REFERENCES

[1] H. Zhang, H. Cao, X. Yang, C. Deng, and D. Tao, "Self-training with progressive representation enhancement for unsupervised cross-domain person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 5287–5298, 2021.

[2] J. Xie *et al.*, "Advanced dropout: A model-free methodology for Bayesian dropout optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, May 21, 2021, doi: 10.1109/TPAMI.2021.3083089.

[3] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 20–36.

[4] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[5] D. Xie, C. Deng, H. Wang, C. Li, and D. Tao, "Semantic adversarial network with multi-scale pyramid attention for video classification," in *Proc. AAAI*, 2019, pp. 9030–9037.

[6] P. Barra, S. Barra, C. Bisogni, M. De Marsico, and M. Nappi, "Web-shaped model for head pose estimation: An approach for best exemplar selection," *IEEE Trans. Image Process.*, vol. 29, pp. 5457–5468, 2020.

[7] X. Chen, C. Xu, X. Yang, and D. Tao, "Long-term video prediction via criticization and retrospection," *IEEE Trans. Image Process.*, vol. 29, pp. 7090–7103, 2020.

[8] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Joint learning of object and action detectors," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4163–4172.

[9] C. Gu *et al.*, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[10] C. Xu, S.-H. Hsieh, C. Xiong, and J. J. Corso, "Can humans fly? Action understanding with multiple classes of actors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2264–2273.

[11] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 724–732.

[12] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. M. Snoek, "Actor and action video segmentation from a sentence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5958–5966.

[13] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 108–124.

[14] Z. Li, R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Tracking by natural language specification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6495–6503.

[15] H. Wang, C. Deng, J. Yan, and D. Tao, "Asymmetric cross-guided attention network for actor and action video segmentation from natural language query," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3939–3948.

[16] K. Ning, L. Xie, F. Wu, and Q. Tian, "Polar relative positional encoding for video-language segmentation," in *Proc. IJCAI*, Jul. 2020, pp. 948–954.

[17] H. Wang, C. Deng, F. Ma, and Y. Yang, "Context modulated dynamic networks for actor and action video segmentation with language queries," in *Proc. AAAI*, 2020, pp. 12152–12159.

[18] J. Xu, D. B. Bert, T. Tinne, and G. V. Luc, "Dynamic filter networks," in *Adv. Neural Inf. Process. Syst.*, 2016, pp. 667–675.

[19] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[20] J. Lu, D. Batra, D. Parikh, and S. Lee, "VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[21] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[22] C. Xu and J. J. Corso, "Actor-action semantic segmentation with grouping process models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3083–3092.

[23] K. Ning, L. Xie, J. Liu, F. Wu, and Q. Tian, "Interaction-integrated network for natural language moment localization," *IEEE Trans. Image Process.*, vol. 30, pp. 2538–2548, 2021.

[24] Y. Yan, C. Xu, D. Cai, and J. J. Corso, "Weakly supervised actor-action segmentation via robust multi-task ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1298–1307.

[25] Y. Gong, K. Yang, Y. Liu, K.-P. Lim, N. Ling, and H. R. Wu, "Quantization parameter cascading for surveillance video coding considering all inter reference frames," *IEEE Trans. Image Process.*, vol. 30, pp. 5692–5707, 2021.

[26] J. Xie, Z. Ma, D. Chang, G. Zhang, and J. Guo, "GPCA: A probabilistic framework for Gaussian process embedded channel attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 10, 2021, doi: 10.1109/TPAMI.2021.3102955.

[27] Y. Zhou, X. Xu, F. Shen, X. Zhu, and H. T. Shen, "Flow-edge guided unsupervised video object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Feb. 8, 2021, doi: 10.1109/TCSVT.2021.3057872.

[28] F. Lin, H. Xie, C. Liu, and Y. Zhang, "Bilateral temporal re-aggregation for weakly-supervised video object segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 13, 2021, doi: 10.1109/TCSVT.2021.3127562.

[29] J. Yang, Y. Huang, K. Niu, Z. Ma, and L. Wang, "Actor and action modular network for text-based video segmentation," 2020, *arXiv:2011.00786*.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL*, 2019, pp. 4171–4186.

[31] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "TransVG: End-to-end visual grounding with transformers," 2021, *arXiv:2104.08541*.

[32] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," 2021, *arXiv:2106.03089*.

[33] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7464–7473.

[34] Y.-C. Chen *et al.*, "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.

[35] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers," 2020, *arXiv:2004.00849*.

[36] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9992–10002.

[37] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*. NIH Public Access, 2019, p. 6558. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/pmc7195022/

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[39] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[40] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[41] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.

[42] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[44] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. EMNLP*, 2014, pp. 787–798.

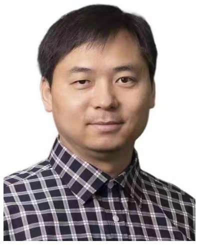
[45] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

Xu Yang (Member, IEEE) received the B.E. and Ph.D. degrees in electronic and information engineering from Xidian University, China, in 2016 and 2021, respectively. He is currently a Lecturer with the School of Electronic Engineering, Xidian University. His research interests lie primarily in computer vision and machine learning.

Hao Wang received the B.E. degree in electronic and information engineering from Hangzhou Dianzi University, Hangzhou, China, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China. His main research interests include human action recognition, video and language understanding, and zero-shot learning.

De Xie received the B.E. degree from the Xi'an University of Architecture and Technology, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree with the School of Electronic Engineering, Xidian University. His research interests focus on computer vision, natural language processing, and multi-modal analysis.

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He is the author and the coauthor of more than 100 scientific articles at top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, NeurIPS, ICML, CVPR, ICCV, AAAI, IJCAI, and KDD. His research interests include computer vision, pattern recognition, and information hiding.



Dacheng Tao (Fellow, IEEE) is currently the President of the JD Explore Academy and the Senior Vice President of JD.com. He is also an Advisor and a Chief Scientist of the Digital Science Institute, The University of Sydney. He mainly applies statistics and mathematics to artificial intelligence and data science. His research is detailed in one monograph and over 200 publications in prestigious journals and proceedings at leading conferences. He is a fellow of the Australian Academy of Science, AAAS, and ACM. He has received the 2015 Australian Scopus-Eureka Prize, the 2018 IEEE ICDM Research Contributions Award, and the 2021 IEEE Computer Society McCluskey Technical Achievement Award.