Asymmetric Cross-Guided Attention Network for Actor and Action Video Segmentation From Natural Language Query

Hao Wang[1], Cheng Deng[1,2]*, Junchi Yan[3], Dacheng Tao[4]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Tencent AI Lab, Shenzhen, China
[3]Department of CSE, and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
[4]UBTECH Sydney AI Centre, School of Computer Science, FEIT, University of Sydney, Australia
{haowang.xidian, chdeng.xd}@gmail.com, yanjunchi@sjtu.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Actor and action video segmentation from natural language query aims to selectively segment the actor and its action in a video based on an input textual description. Previous works mostly focus on learning simple correlation between two heterogeneous features of vision and language via dynamic convolution or fully convolutional classification. However, they ignore the linguistic variation of natural language query and have difficulty in modeling global visual context, which leads to unsatisfactory segmentation performance. To address these issues, we propose an asymmetric cross-guided attention network for actor and action video segmentation from natural language query. Specifically, we frame an asymmetric cross-guided attention network, which consists of vision guided language attention to reduce the linguistic variation of input query and language guided vision attention to incorporate query-focused global visual context simultaneously. Moreover, we adopt multi-resolution fusion scheme and weighted loss for foreground and background pixels to obtain further performance improvement. Extensive experiments on Actor-Action Dataset Sentences and J-HMDB Sentences show that our proposed approach notably outperforms state-of-the-art methods.

1. Introduction

With the explosive growth of video data in recent years, video understanding has attracted ever-increasing attention in computer vision community. However, traditional studies emphasize on video classification [23, 27, 29], action recognition and localization [35, 36, 37, 38, 40]. Both of them lack fine-grained analysis of video contents, such as pixel-level joint understanding of actors and their actions, which plays crucial role in human-robot interaction and autonomous driving. Attempting to understand actors and actions present in videos, Gavrilyuk *et al*. [6] introduced a challenging task of actor and action video segmentation from natural language query, as illustrated in Figure 1.

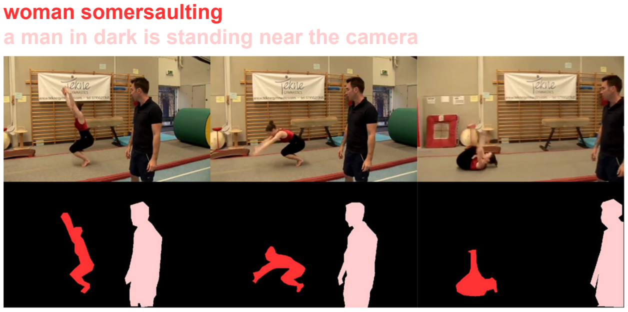

Figure 1. Based on the input natural language query, actor and action video segmentation aims at generating pixel-wise segmentation masks in a given video. The colored masks are corresponding to the sentences with the same color on the top of the video.

Recently, many approaches [7, 8, 16, 22, 39] have been exploited for semantic segmentation or object localization from natural language query. These approaches can be roughly divided into two categories. In the first category, dynamic convolution is utilized to adaptively segment or localize an object, where the generated dynamic convolutional filters vary with the input natural language query. However, the linguistic variation of input textual description would seriously impact sentence representation and subsequently make dynamic convolutional filters unstable, leading to inaccurate segmentation or localization. For example, "car in blue is parked on the grass" and "blue car standing on the grass" have the same meaning but different generated filters, resulting in unsatisfactory performance. In the second one, heterogeneous features from vision and language

*Corresponding author

modalities are concatenated firstly and then utilized for segmentation or localization via fully covolutional networks. Unfortunately, they are incapable of modeling global visual context, which is crucial for object segmentation or detection as verified in [21, 3]. Moreover, query-focused pixels should be devoted more efforts for context modeling to promote the correlation between visual information and language description. For example, to segment the "man on the chair", we need to take the pixels of the man on the chair not grass or floor into consideration for aggregating visual context.

In this paper, we propose a novel asymmetric cross-guided attention network to deal with actor and action video segmentation from natural language query. The network is structurally asymmetric and consists of two parallel attention modules: vision guided language attention module and language guided vision attention module. Specifically, to address the linguistic variation of natural language query, we devise a vision guided language attention module to obtain more robust sentence representation, which reduces the disturbance of noisy words and promotes the correlation between visual pixels and textual descriptions. Furthermore, to incorporate global visual context for segmentation, we elaborate a language guided vision attention module to aggregate query-focused visual context, leading to better segmentation performance. Additionally, we utilize the multi-resolution fusion for various grained segmentation masks and the weighted loss for foreground and background pixels to achieve extra performance improvement.

The main contributions of this work are as follows:

- We frame an asymmetric cross-guided attention network, to simultaneously reduce the linguistic variation and incorporate query-focused global visual context, for more effective actor and action video segmentation;

- We devise a simple yet effective multi-resolution fusion scheme in addition with a weighted loss for foreground pixels, which can boost segmentation performance with negligible computation cost;

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

2. Related Work

2.1. Actor and Action Segmentation

For comprehensive action understanding, Xu *et al.* [31] collected and annotated the Actor-Action Dataset (A2D) with fixed actor and action pairs and introduced the challenging task of actor and action video segmentation. Existing methods can be mainly divided into two categories: methods based on supervoxels features and those based on deep features. In the first category, Xu *et al.* [31] proposed a trilayer approach to model the interaction of separate actor and action nodes with actor-action product nodes. Xu and Corso [30] proposed a grouping process to encourage adaptive and long-ranging interactions of video parts. Yan *et al.* [34] utilized robust multi-task ranking model to address weakly-supervised actor and action segmentation. In the second category, Kalogeiton *et al.* [10] jointly learned the detectors of object and its action in a video by taking advantage of deep features and then obtained segmentation results via existing segmentation methods. Recently, Gavrilyuk *et al.* [6] extended A2D with human annotated sentences and introduced the challenging task of actor and action video segmentation from natural language query. They adopted dynamic convolution, where the filters adaptively varied with different input textual descriptions. However, they not only ignored the linguistic variation of textual description but also solely tackled each pixel without considering the context information. Different from above works, our proposed asymmetric cross-guided attention network enables visual and linguistic features learn from each other, leading to better segmentation performance.

2.2. Actor and Action Localization from a Sentence

According to the tasks they are focused on, existing methods of actor and action localization from a sentence can be categorized into two classes: actor localization from a sentence and action localization from a sentence. In the first class, Li *et al.* [15] introduced an interesting task of person search with natural language description and proposed a recurrent neural network with a gated neural attention mechanism to calculate word-image affinity. Yamaguchi *et al.* [33] extracted candidate tubes and conducted relevance computation between text features and tube features for spatio-temporal person retrieval. In the second class, Gao *et al.* [5] proposed a multi-modal processing network to generate alignment scores and location offsets for temporal activity localization via language query. Hendricks *et al.* [1] integrated local and global video features to localize moments in video with natural language. Instead of generating bounding boxes around the human actor or performing action, we prefer a pixel-wise actor and action video segmentation from natural language query for the further understanding of video contents.

2.3. Attention Mechanism

To mimic how human vision works, attention mechanism has been exploited in many fields such as natural language processing [26], visual question answering [18], image caption [32], and video classification [28]. These methods can be divided into two categories according to the network architecture: self-attention based approaches and co-attention based approaches. The self-attention mechanism

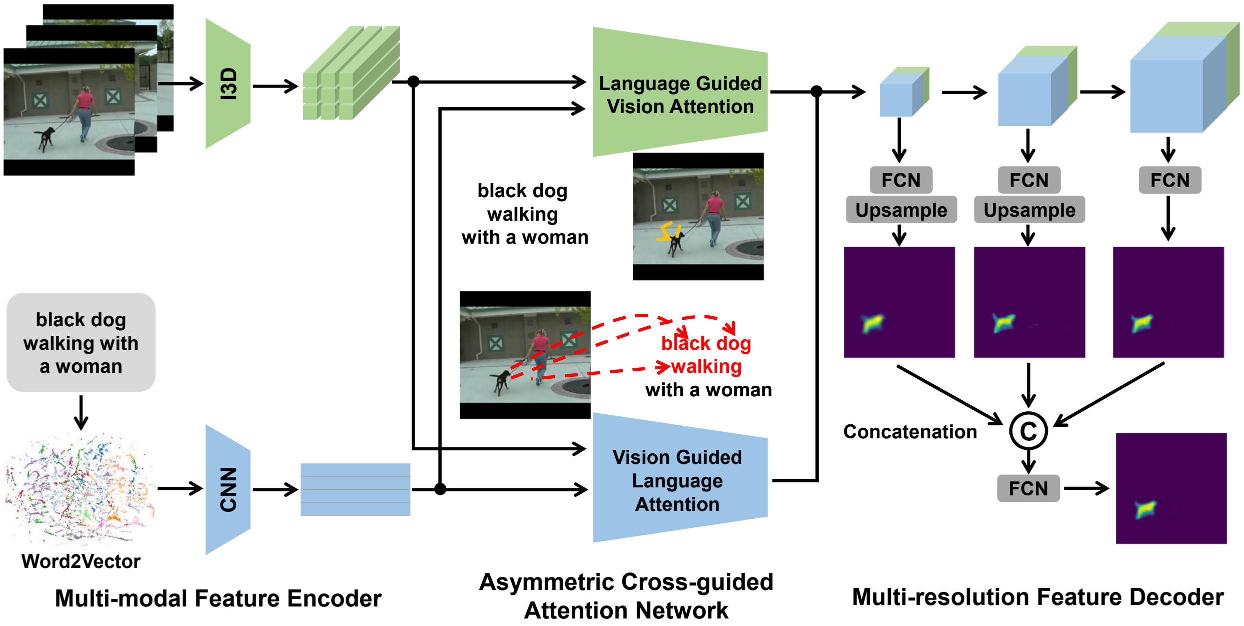
I3D
Language Guided Vision Attention
black dog walking with a woman
black dog walking with a woman
black dog
walking
with a woman
CNN
Word2Vector
Vision Guided Language Attention
FCN
Upsample
FCN
Upsample
FCN
Concatenation
C
FCN
Multi-modal Feature Encoder
Asymmetric Cross-guided Attention Network
Multi-resolution Feature Decoder

Figure 2. Our proposed asymmetric cross-guided attention network, which consists of multi-modal feature encoder, asymmetric cross-guided attention network and multi-resolution feature decoder. After extracting video and text features, the models learns most correlated language features for visual pixels, *e.g*., "black dog walking" for the dog, and incorporates query-focused context, *e.g*., the pixels of the exact described "dog". Finally, we concatenate the weighted vision and language features for segmentation.

[26, 28] first transforms the features into query, key and value features, and then calculates the self-attention matrix between query and key features via inner product. After normalization, final features are obtained through weighted summation of original features on the basis of self-attention matrix. Compared with the self-attention, the co-attention mechanism explicitly computes the interactions across two different modalities. Then features of each modality are aggregated as weighted summation of original features based on the co-attention matrix. Similarly, MRN [12] learns multimodal joint representation in a residual way and AVDLN [24] extends it with symmetric residual fusion and unidirectional attention. Our model offers a novel asymmetric cross-guided attention mechanism, which consists of vision guided language attention (*i.e*., co-attention) to reduce linguistic variation and language guided vision attention (*i.e*., gated self-attention) to aggregate query-focused global visual context.

3. Proposed Method

Given a video and a corresponding natural language query, our method is to segment the actor and its action in the video referred by the query. In this paper, we propose a novel asymmetric cross-guided attention network, which simultaneously reduces the linguistic variation of natural language query and incorporates query-focused global visual context, achieving significant improvement on segmentation performance. The architecture of our method is illustrated as Figure 2, which consists of three components: multi-modal feature encoder, asymmetric cross-guided attention network, and multi-resolution feature decoder.

3.1. Multi-modal Feature Encoder

To extract multi-modal features for segmentation, we introduce the text encoder and video encoder below.

We first obtain word vectors by using the word2vec model pre-trained on the Google News Dataset [20] instead of training word embedding model from scratch. It can not only simplify the training procedure of natural language model but also help to exploit similarity in descriptions across different datasets. Then temporal information of textual description is captured by a simple yet effective 1D convolutional neural network [13] atop word vectors instead of long-short term memory network like in [7, 16]. Specifically, each word is encoded as a 300-dimensional word vector and then the input sentence is composed by individual word representations. Subsequently, a single 1D convolutional layer with non-linear activation is utilized on input sentence matrix. We denote sentence matrix as $S \in \mathbb{R}^{N_T \times D_T}$, where N_T is the maximum length of words in the dataset and D_T is the feature dimension of word vector. Therefore, the text encoder can be formulated as

$$F_T = \mathrm{Enc}_T(S; \theta_T), \tag{1}$$

where Enc_T is text encoder parameterized with θ_T and

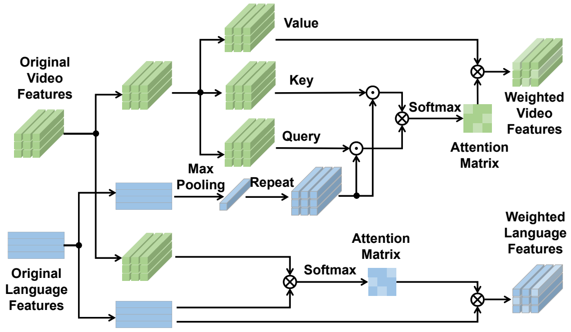

Figure 3. The proposed asymmetric cross-guided attention network consists of vision guided language attention module implemented with co-attention mechanism and language guided vision attention module implemented with gated self-attention mechanism. The "$\odot$" denotes element-wise multiplication while "$\otimes$" stands for inner production, respectively.

$F_T \in \mathbb{R}^{N_T \times D_T}$ is the extracted sentence representation.

To encode appearance information of actor and motion information of its corresponding action simultaneously, we adopt 3D convolution neural network to learn video representation. Different from 2D convolution neural network, Tran *et al*. [25] proposed C3D network and proved the effectiveness of 3D convolution and pooling in video processing. To better exploit the spatio-temporal property of video, Carreira and Zisserman [2] proposed Inflated 3D ConvNet (I3D) and achieved state-of-the-art performance on human action recognition. Here, the I3D model pre-trained on ImageNet [4] and Kinetics [2] datasets is adopted for video feature extraction. We feed a video clip into I3D model and extract the intermediate output before $\mathrm{maxpool3d\_5a}$ layer. Then clip features are obtained by conducting temporal average pooling and followed L_2 normalization on each pixel. Given an input video $V \in \mathbb{R}^{3 \times N_V \times H_V \times W_V}$, the video encoder can be formulated as

$$F_V = L_2(\mathrm{Avg}(\mathrm{Enc}_V(V; \theta_V))), \tag{2}$$

where Enc_V stands for the partial of pre-trained I3D model parameterized with θ_V, Avg and L_2 stand for temporal average pooling and L_2 normalization. N_V, H_V, and W_V denote the frame number, the height, and the width of input frames, respectively. To identify some words like "bottom" or "in the middle", we concatenate spatial coordinates features $C \in \mathbb{R}^{H_F \times W_F \times D_C}$ with video clip features $F_V \in \mathbb{R}^{H_F \times W_F \times D_F}$ along channel dimension. Here, H_F, W_F, D_F, and D_C denote the height, the width, the dimension of extracted video feature map, and the dimension of spatial coordinates features, respectively.

3.2. Asymmetric Cross-guided Attention Network

After the feature extraction of video and natural language query, heterogeneous features from two different modalities are concatenated along channel dimension to perform segmentation, as proposed in [7, 19]. However, they ignore the linguistic variation of textual description and solely tackle each pixel without considering the context information. To address these issues, we elaborate a novel asymmetric cross-guided attention network, consisting of vision guided language attention module to reduce the linguistic variation of input query and language guided vision attention module to aggregate query-focused global visual context. The architecture of the asymmetric cross-guided attention network is illustrated in Figure 3.

The vision guided language attention module captures pixel-wise interaction between vision and language modalities and then utilizes the calculated co-attention matrix followed by normalization to compute the weighted summation of original language features. The video features with spatial information, denoted as F_{VC}, are firstly aligned to the features with same dimension as language features,

$$F_{VC \to T} = \mathrm{Linear}(F_{VC}), \tag{3}$$

where $F_{VC \to T}$ is the aligned features and Linear stands for fully connected layer. By conducting co-attention, normalization, and weighted summation, we can formulate the process of vision guided language attention as

$$F_{TA} = \mathrm{softmax}\left(\frac{F_{VC \to T} F_T^\top}{\sqrt{D_T}}\right) F_T. \tag{4}$$

Then the weighted language features are concatenated with visual features along channel dimension. It means that, for each pixel of visual feature map, most related textual features are learned, which significantly reduces the linguistic variation and boosts the segmentation performance.

Recently, self-attention mechanism is proposed to capture long-ranging dependency and has achieved good results in natural language processing [26] and video classification [28]. However, the native self-attention only utilizes intra-modality information to estimate pixel-to-pixel importance, *i.e*., aggregating global context information. Inspired by the idea that relations between different pixels should be weighted differently according to input natural language query, we design a language guided vision attention module to incorporate query-focused global visual context. Firstly, we conduct temporal max pooling and spatial tile over textual features to align them with the same dimension as visual features, which can be defined as

$$F_{T \to VC} = \mathrm{Linear}(\mathrm{Repeat}(\mathrm{Max}(F_T))). \tag{5}$$

Then video features with spatial information (*i.e*., F_{VC}) are transformed into query, key, and value features via single fully connected layer,

$$\begin{aligned} F_{VCQ} &= \mathrm{Linear}(F_{VC}), \\ F_{VCK} &= \mathrm{Linear}(F_{VC}), \\ F_{VCV} &= \mathrm{Linear}(F_{VC}), \end{aligned} \tag{6}$$

where F_{VCQ}, F_{VCK}, and F_{VCV} are query, key, and value features, respectively. To introduce conditional information of natural language description, we obtain dynamic query features and dynamic key features by gating the original query and key features with textual information,

$$\begin{aligned}\tilde{F}_{VCQ} &= F_{VCQ} \odot F_{T \to VC}, \\ \tilde{F}_{VCK} &= F_{VCK} \odot F_{T \to VC},\end{aligned} \tag{7}$$

where $\odot$ is element-wise multiplication. Finally, the language guided vision attention can be described as

$$F_{VA} = \text{softmax}\left(\frac{\tilde{F}_{VCQ}\tilde{F}_{VCK}^{\top}}{\sqrt{D_V}}\right) F_{VCV}. \tag{8}$$

It can enhance the correlations among the pixels of the region related to the natural language query, leading to better segmentation by incorporating query-focused global visual context.

To simplify the description, we define asymmetric cross-guided attention network as

$$F_{TA}, F_{VA} = \text{Att}(F_{VC}, F_T; \theta_{Att}), \tag{9}$$

where Att is the attention network parameterized with θ_{Att}. It is implemented with standard components in neural networks and thus can be integrated into other tasks seamlessly like visual question answering and phrase referring.

3.3. Multi-resolution Feature Decoder

To obtain final segmentation results with the same resolution as the input video, we adopt multi-resolution (*i.e.*, 32×32, 128×128 and 512×512) feature decoders to upsample the feature map in a progressive manner. We concatenate weighted language features F_{TA}, weighted video features F_{VA}, and spatial features C along channel dimension, and then conduct segmentation through fully convolutional networks. The medium and large resolution video features are denoted as F_V^M and F_V^L, respectively. We formulate them as

$$\begin{aligned}F_V^M &= \text{Deconv}(F_{VA}), \\ F_V^L &= \text{Deconv}(F_V^M),\end{aligned} \tag{10}$$

where Deconv stands for deconvolutional network, consisting of one deconvolutional layer and one followed convolutional layer.

The multi-resolution segmentation responses are obtained as

$$\begin{aligned}R^S &= \text{FCN}([F_{VA}, C, F_{TA}]), \\ R^M &= \text{FCN}([F_V^M, \text{Interp}(C), \text{Interp}(F_{TA})]), \\ R^L &= \text{FCN}([F_V^L, \text{Interp}(C), \text{Interp}(F_{TA})]),\end{aligned} \tag{11}$$

where R^S, R^M, and R^L are small, medium, and large resolution segmentation responses, respectively. Interp denotes bilinear interpolation operation and FCN denotes fully convolutional network. Furthermore, we elaborate a multi-resolution fusion scheme to take advantage of various grained segmentation responses and obtain the final response as

$$R^L = \text{FCN}([\text{Interp}(R^S), \text{Interp}(R^M), R^L]). \tag{12}$$

In summary, the multi-resolution feature decoder can be expressed as

$$R^S, R^M, R^L = \text{Dec}(F_{VA}, C, F_{TA}; \theta_D), \tag{13}$$

where Dec represents the feature decoder parameterized with θ_D. Unlike the static interpolation of segmentation response, the trainable deconvolution on feature map would make the model exploit more accurate segmentation results. In addition, multi-resolution structure can not only utilize various grained information for segmentation but also provide sufficient gradients for training the whole model better.

3.4. Training and Inference

Our proposed approach takes video clips V, natural language queries S, and binary ground-truth segmentation masks Y as inputs and generates selective segmentation mask related to textual description.For each resolution $r \in \{S, M, L\}$, the segmentation loss $\mathcal{L}^r$ between the response R^r and the ground-truth Y^r is calculated as

$$\mathcal{L}^r = \frac{1}{H^r W^r} \sum_{i=1}^{H^r} \sum_{j=1}^{W^r} \ell(R_{ij}^r, Y_{ij}^r), \tag{14}$$

where ℓ is weighted binary cross entropy, and H^r and W^r are the height and the width of ground-truth masks Y^r, respectively. Given coefficient P for foreground pixels, the weighted loss can be formulated as

$$\begin{aligned}\ell(R_{ij}^r, Y_{ij}^r) = &- PY_{ij}^r \log(\sigma(R_{ij}^r)) \\ &- (1 - Y_{ij}^r) \log(1 - \sigma(R_{ij}^r)),\end{aligned} \tag{15}$$

where σ is sigmoid function. The intermediate ground-truths Y^S and Y^M are acquired through the bilinear interpolation of Y^L. Finally, the loss of our proposed approach can be formulated as

$$\mathcal{L} = \lambda_1 \mathcal{L}^S + \lambda_2 \mathcal{L}^M + \lambda_3 \mathcal{L}^L, \tag{16}$$

where λ_1, λ_2, and λ_3 are weights for different resolutions.

During inference, we segment a pixel as foreground when its response value is higher than 50% of the max value in the response map. It should be noticed that we map the final mask back to their original frame size for evaluation.

Table 1. Segmentation results on A2D Sentences. The approaches marked by "*" are fine-tuned on the A2D Sentences. Our proposed model significantly outperforms the state-of-the-arts even only takes multiple RGB frames as inputs.

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.* [7]	7.7	3.9	0.8	0.0	0.0	2.0	21.3	12.8
Li *et al.* [16]	10.8	6.2	2.0	0.3	0.0	3.3	24.8	14.4
Hu *et al.* [7] *	34.8	23.6	13.3	3.3	0.1	13.2	47.4	35.0
Li *et al.* [16] *	38.7	29.0	17.5	6.6	0.1	16.3	51.5	35.4
Gavrilyuk *et al.* [6] (RGB)	47.5	34.7	21.1	8.0	0.2	19.8	53.6	42.1
Gavrilyuk *et al.* [6] (RGB + Flow)	50.0	37.6	23.1	9.4	0.4	21.5	55.1	42.6
Ours (RGB)	**55.7**	**45.9**	**31.9**	**16.0**	**2.0**	**27.4**	**60.1**	**49.0**

Table 2. We evaluate the generalization ability on J-HMDB Sentences with the best model trained on A2D Sentences.

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.* [7]	63.3	35.0	8.5	0.2	0.0	17.8	54.6	52.8
Li *et al.* [16]	57.8	33.5	10.3	0.6	0.0	17.3	52.9	49.1
Gavrilyuk *et al.* [6] (RGB + Flow)	69.9	46.0	17.3	1.4	0.0	23.3	54.1	54.2
Ours (RGB)	**75.6**	**56.4**	**28.7**	**3.4**	0.0	**28.9**	**57.6**	**58.4**

Table 3. Segmentation results on A2D Sentences for ablation studies. Multi-resolution Fusion, Weighted Binary Cross Entropy with logits, Attention model are abbreviated as "MRF", "WBCE" and "ATT", respectively.

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
Gavrilyuk *et al.* [6] (RGB)	47.5	34.7	21.1	8.0	0.2	19.8	53.6	42.1
Gavrilyuk *et al.* [6] (RGB + Flow)	50.0	37.6	23.1	9.4	0.4	21.5	55.1	42.6
Baseline (RGB)	48.9	36.1	21.8	9.2	0.3	20.6	52.8	44.1
Baseline + MRF (RGB)	50.1	38.4	26.2	13.0	1.1	23.1	57.7	45.5
Baseline + MRF + WBCE (RGB)	53.5	43.4	29.7	13.7	1.4	25.5	57.4	47.5
Baseline + MRF + WBCE + ATT (RGB)	**55.7**	**45.9**	**31.9**	**16.0**	**2.0**	**27.4**	**60.1**	**49.0**

4. Experiment

4.1. Datasets and Evaluation Criteria

A2D Sentences is extended on the Actor-Action Dataset (A2D) by Gavrilyuk *et al.* [6] via providing the textual descriptions for each video. It contains 3,782 videos collected from YouTube and includes 8 actions classes performed by 7 actors classes. There are 3 to 5 frames for each video with dense pixel-level actor and action annotations for evaluating segmentation performance. Besides, it contains 6,655 sentences to describe actors and their actions presented in the video. Following [6], we split the dataset into 3,017 training videos, 737 testing videos and 28 unlabeled videos.

J-HMDB Sentences contains 928 videos and corresponding 928 sentences, which is extended on the J-HMDB dataset [9]. The annotations include 2D articulated human puppet masks for dense segmentation labeling and natural language queries for describing what action the object is performing in each video.

We evaluate our proposed approach by using the criteria of Intersection-over-Union (IoU) and precision. Specifically, the overall IoU computes the ratio of the total intersection area divided by the total union area over all testing samples, which obviously favor large actors and objects. The mean IoU calculates the average IoU of all testing samples by treating samples of different size equally. The precision@K reports the percentage of testing samples whose IoU scores are higher than threshold K. We measure precision at 5 different IoU thresholds and average precision over 0.50:0.05:0.95 [17].

4.2. Implementation Details

According to [6], the multi-modal feature encoder adopts pre-trained I3D model to extract video clip features and pre-trained word2vector model to convert sentence into vector matrix. The maximum length of words is set to 20 and its

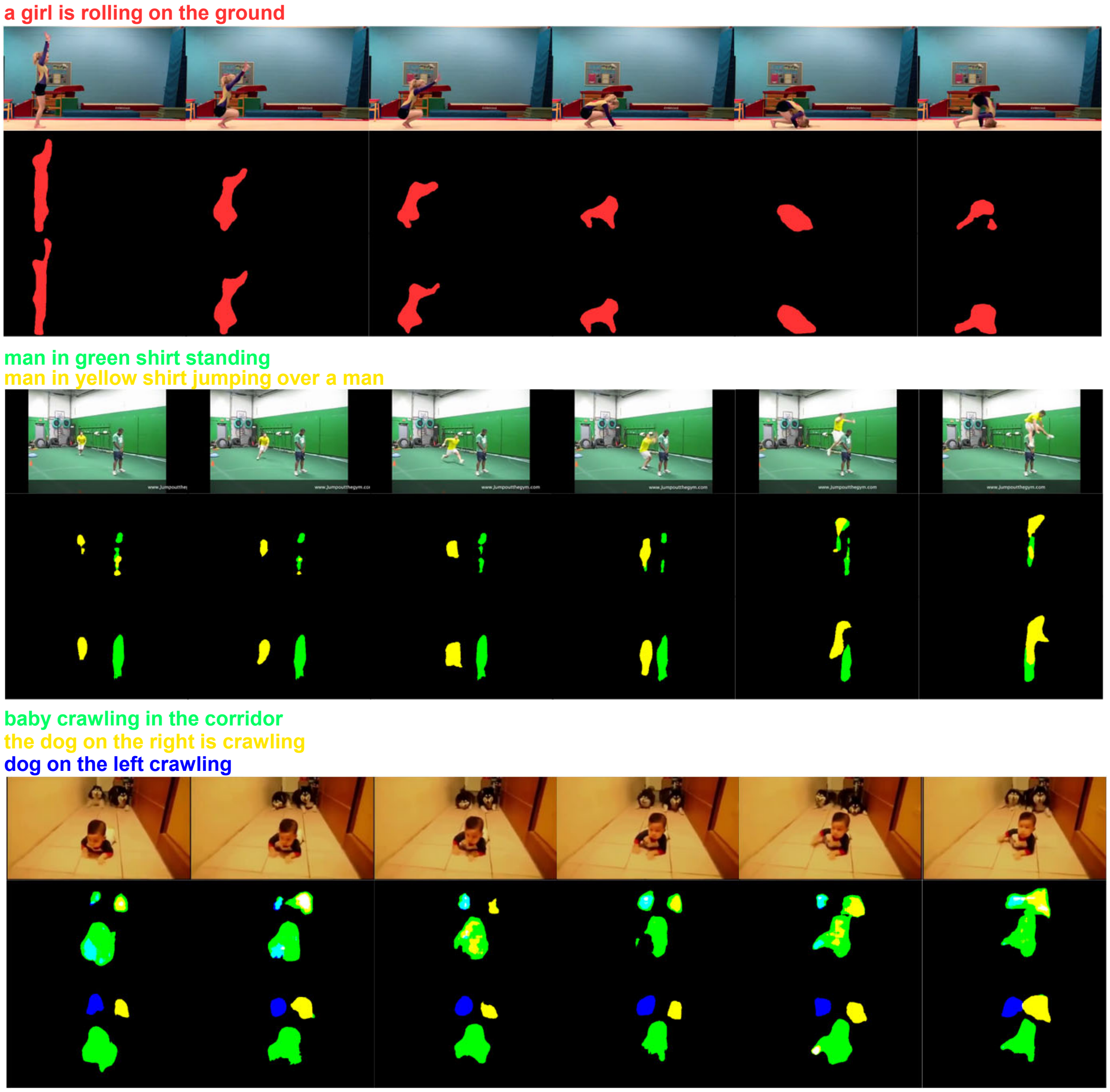
a girl is rolling on the ground
man in green shirt standing
man in yellow shirt jumping over a man
baby crawling in the corridor
the dog on the right is crawling
dog on the left crawling

Figure 4. Qualitative results on A2D Sentences. The first row shows the frames of input videos. The second row illustrates the segmentation results from [6] and the third row are the segmentation outputs of our method. Both of them are trained on RGB frames for fair comparison. The colored masks correspond to the sentences with the same color on the top of each video. Some overlaps are mixture of colors.

feature dimension is 300. We fine-tune the last inception block before maxpool3d_5a layer of video encoder only on A2D Sentences. The FCN in deconvolutional network consists of three fully convolutional layers, where the kernel size is 3×3 for the first two layers and 1×1 for the remaining layer. For FCN in multi-resolution fusion, there are only one fully convolutional layer with kernel size 3×3.

All experiments in this paper are implemented with PyTorch package. We use an Adam [14] optimizer with the learning rate 5×10^{-4}. The batch size and maximum number of training epochs are 4 and 12, respectively. The learning rate is divided by 10 every 8 epochs. The loss weights of λ_1, λ_2, and λ_3 are fixed as 1 across all the experiments. The coefficient of weighted binary cross entropy loss for foreground pixels is set to 1.5. We only take the number of 16 RGB frames as video inputs for our proposed approach. The frame annotated with ground-truth mask is in the middle of video clips.

4.3. Comparison with State-of-the-art Methods

We show results of actor and action video segmentation from natural language query compared with one approach [6] of the same task and two approaches [7, 16] of image segmentation from a sentence in Table 1. There are two

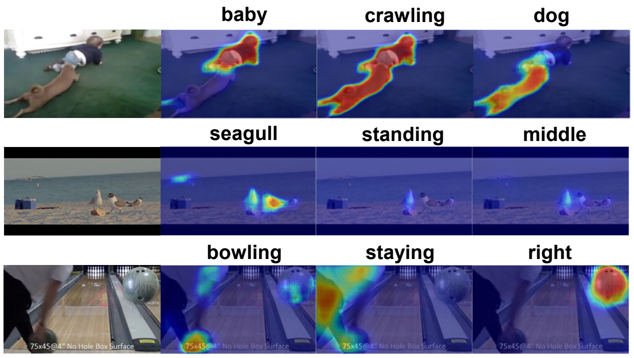

Figure 5. Visualization on the attended region with the word as input above each image.

training setting for prior works [7, 16] . In the first setting, they are trained solely on ReferIt dataset [11] without any fine-tuning on A2D sentences and their results are showed in the first two rows. In the second setting, we fine-tune the models on the training samples of A2D sentence and mark the approaches with "*". We observe that the same approach fine-tuned on A2D Sentences significantly improves the segmentation performance, which demonstrates that dataset-specific video features play 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 demonstrates the advantages of our method compared with existing state-of-the-art method [6]. Moreover, we bring 5.0% absolute improvement on Overall IoU, 6.4% in Mean IoU, and 5.9% in mAP over state-of-the-arts, respectively. It should be noticed that our proposed approach only takes RGB frames as video inputs without using any additional information (*i.e.*, optical flow computed from adjacent frames as in [6]). Furthermore, qualitative results on A2D Sentences are presented in Figure 4. We observe that our method can produce more fine-grained and separated masks than [6]. Specifically, our model can generate fine-grained segmentation of actors or objects, *e.g.*, hands of the girl in the first video. The model can tackle the background interference, *e.g.*, "main in green" in the second video. Besides, our model can generate better responses for spatial qualifiers, *e.g.*, in the third video. In Figure 5, we also provide the visualization of attention region for individual word to understand the correlation between visual and linguistic features. We find that the model can learn the correlations between the nouns, verbs, spatial qualifiers and their corresponding visual parts.

To further evaluate the generalization ability of our proposed approach, we use the model pre-trained on A2D Sentences to segment all samples in J-HMDB Sentences without any additional fine-tuning. During evaluation, we uniformly sample 3 frames of each testing video as indicated in [6]. The segmentation results are reported in Table 2. In spite of obtaining obvious improvement on most metrics, we still get poor precision performance at the threshold of 0.9. We guess the video encoder trained on A2D Sentences can not produce features for accurate segmentation without any fine-tuning on J-HMDB Sentences. More detailed analyses will be included in supplementary material.

4.4. Ablation Studies

In order to verify the effectiveness of each component in our proposed approach, we conduct ablation studies and their results are illustrated in Table 3.

Baseline only replaces the dynamic convolution with fully convolutional network to model the complex correlations of concatenated heterogeneous features. When using RGB frames as video inputs, it obviously beats the state-of-the-art method [6] in most cases under different metrics.

Baseline+MRF improves segmentation performance by a large margin through fusing multi-resolution segmentation responses. It reflects the great potential of fusing various grained results for final segmentation.

Baseline+MRF+WBCE achieves similar performance on Overall IoU but more higher performance on Mean IoU contrast to Baseline+MRF, which means the weighted loss is beneficial to segment out much more foreground pixels.

Baseline+MRF+WBCE+ATT obtains remarkable improvement on all metrics, which shows that the asymmetric cross-guided attention network can significantly reduce linguistic variation and incorporate query-focused global visual context.

5. Conclusion

In this paper, we have proposed an asymmetric cross-guided attention network to handle the linguistic variation of natural language query, which also incorporates query-focused global visual context. Our approach achieves notable improvement on segmentation performance. It can be seamlessly integrated into other tasks such as visual question answering and phrase referring. In the future, we should devote more efforts on the generalization ability of segmentation model to have more in-depth understanding of the underlying mechanism.

6. Acknowledgement

Our work was supported by the National Natural Science Foundation of China under Grant 61572388 and 61703327, the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02, 2017ZDCXL-GY-05-02 and 2018ZDXM-GY-176, the National Key R&D Program of China under Grant 2017YFE0104100, and the ARC FL-170100117, DP-180103424.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[3] Zhe Chen, Shaoli Huang, and Dacheng Tao. Context refinement for object detection. In *ECCV*, pages 74–89, 2018.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[5] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: temporal activity localization via language query. In *ICCV*, pages 5267–5275, 2017.

[6] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018.

[7] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124, 2016.

[8] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, pages 4555–4564, 2016.

[9] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013.

[10] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Joint learning of object and action detectors. In *ICCV*, pages 4163–4172, 2017.

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.

[12] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *NeurIPS*, pages 361–369, 2016.

[13] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv:1408.5882*, 2014.

[14] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv:1412.6980*, 2014.

[15] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017.

[16] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *CVPR*, pages 6495–6503, 2017.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: common objects in context. In *ECCV*, pages 740–755, 2014.

[18] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, pages 289–297, 2016.

[19] Bruce McIntosh, , Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Multi-modal capsule routing for actor and action video segmentation conditioned on natural language queries. *arXiv:1812.00303*, 2018.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013.

[21] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *CVPR*, pages 4353–4361, 2017.

[22] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, pages 38–54, 2018.

[23] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, pages 568–576, 2014.

[24] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018.

[25] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[27] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, pages 20–36, 2016.

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018.

[29] De Xie, Cheng Deng, Hao Wang, Chao Li, and Dapeng Tao. Semantic adversarial network with multi-scale pyramid attention for video classification. In *AAAI*, pages 9030–9037, 2019.

[30] Chenliang Xu and Jason J Corso. Actor-action semantic segmentation with grouping process models. In *CVPR*, pages 3083–3092, 2016.

[31] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, pages 2264–2273, 2015.

[32] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

[33] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *ICCV*, pages 1453–1462, 2017.

[34] Yan Yan, Chenliang Xu, Dawen Cai, and Jason J Corso. Weakly supervised actor-action segmentation via robust multi-task ranking. In *CVPR*, pages 1298–1307, 2017.

[35] Yanhua Yang, Cheng Deng, Shangqian Gao, Wei Liu, Dapeng Tao, and Xinbo Gao. Discriminative multi-instance multitask learning for 3d action recognition. *IEEE Trans. Multimedia*, 19(3):519–529, 2016.

[36] Yanhua Yang, Cheng Deng, Dapeng Tao, Shaoting Zhang, Wei Liu, and Xinbo Gao. Latent max-margin multitask learning with skelets for 3-d action recognition. *IEEE Trans. Cybern.*, 47(2):439–448, 2016.

[37] Yanhua Yang, Ruishan Liu, Cheng Deng, and Xinbo Gao. Multi-task human action recognition via exploring super-category. *IEEE Trans. Signal Process.*, 124:36–44, 2016.

[38] Zhenheng Yang, Jiyang Gao, and Ram Nevatia. Spatio-temporal action detection with cascade proposal and location anticipation. *arXiv:1708.00042*, 2017.

[39] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018.

[40] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, pages 2914–2923, 2017.