MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment

Da Zhang[†], Xiyang Dai[‡], Xin Wang[†], Yuan-Fang Wang[†], and Larry S. Davis[§]

[†]University of California, Santa Barbara; [‡]Microsoft; [§]University of Maryland, College Park
{dazhang, xwang, yfwang}@cs.ucsb.edu, xiyang.dai@microsoft.com, lsd@umiacs.umd.edu

Abstract

This research strives for natural language moment retrieval in long, untrimmed video streams. The problem is not trivial especially when a video contains multiple moments of interests and the language describes complex temporal dependencies, which often happens in real scenarios. We identify two crucial challenges: semantic misalignment and structural misalignment. However, existing approaches treat different moments separately and do not explicitly model complex moment-wise temporal relations. In this paper, we present Moment Alignment Network (MAN), a novel framework that unifies the candidate moment encoding and temporal structural reasoning in a single-shot feed-forward network. MAN naturally assigns candidate moment representations aligned with language semantics over different temporal locations and scales. Most importantly, we propose to explicitly model moment-wise temporal relations as a structured graph and devise an iterative graph adjustment network to jointly learn the best structure in an end-to-end manner. We evaluate the proposed approach on two challenging public benchmarks DiDeMo and Charades-STA, where our MAN significantly outperforms the state-of-the-art by a large margin.

1. Introduction

Video understanding is a fundamental problem in computer vision and has drawn increasing interests over the past few years due to its vast potential applications in surveillance, robotics, etc. While fruitful progress [44, 49, 47, 4, 48, 51, 5, 52, 50, 28, 10, 41, 2, 61, 42, 53, 59] has been made on activity detection to recognize and localize temporal segments in videos, such approaches are limited to work on pre-defined lists of simple activities, such as playing basketball, drinking water, etc. This restrains us from moving towards real-world unconstrained activity detection. To solve this problem, we tackle the natural language moment retrieval task. Given a verbal description, our goal is to determine the start and end time (*i.e*. localization) of the temporal segment (*i.e*. moment) that best corresponds to this given query. While this formulation opens up great opportunities for better video perception, it is substantially more challenging as it needs to model not only the characteristics of sentence and video but also their complex relations.

Figure 1: We consider the natural language moment retrieval task in untrimmed videos. To properly localize the moment, the retrieval model must handle both *semantic misalignment (top)* with multiple moments of interests and *structural misalignment (bottom)* with complex temporal dependencies.

On one hand, a real-world video often contains multiple moments of interests. Consider a simple query like "The child touches the ground the second time", shown in Figure 1, a robust model needs to scan through the video and compare the video context to find the second occurrence of "child touches the ground". This raises the first challenge for our task: *semantic misalignment*. A simple ordinal number will result in searching from a whole video, where a naive sliding approach will fail. On the other hand, the language query usually describes complex temporal dependencies. Consider another query like "Child is running away af-

ter is closest to the camera", different from the sequence described in sentence, the "close to the camera" moment happens before "running away". This raises the second challenge for our task: *structural misalignment*. The language sequence is often misaligned with video sequence, where a naive matching without temporal reasoning will fail.

These two key challenges we identify: semantic misalignment and structural misalignment have not been solved in existing methods [18, 14] for the natural language moment retrieval task. Existing methods sample candidate moments by scanning videos with varying sliding windows, and compare the sentence with each moment individually in a multi-modal common space. Although simple and intuitive, this individualist representations of sentence and video make it hard to model semantic and structural relations among two modalities.

To address the above challenges, we propose an end-to-end Moment Alignment Network (MAN) for the natural language moment retrieval task. The proposed MAN model directly generates candidate moment representations aligned with language semantics, and explicitly model temporal relationships among different moments in a graph-structured network. Specifically, we encode the entire video stream using a hierarchical convolutional network and naturally assign candidate moments over different temporal locations and scales. Language features are encoded as efficient dynamic filters and convolved with input visual representations to deal with semantic misalignment. In addition, we propose an Iterative Graph Adjustment Network (IGAN) adopted from Graph Convolution Network (GCN) [26] to model relations among candidate moments in a structured graph. Our contributions are as follows:

- We propose a novel single-shot model for the natural language moment retrieval task, where language description is naturally integrated as dynamic filters into an end-to-end trainable fully convolutional network.
- To the best of our knowledge, we are the first to exploit graph-structured moment relations for temporal reasoning in videos, and we propose the IGAN model to explicitly model temporal structures and improve moment representation.
- We conduct extensive experiments on two challenging benchmarks: Charades-STA [14] and DiDeMo [18]. We demonstrate the effectiveness of each component and the proposed MAN significantly outperforms the state-of-the-art by a large margin.

2. Related Work

Temporal Activity Detection. Temporal activity detection is the task to predict the start and end time (*i.e.*, localization) and the label (*i.e.*, classification) of activity instances in untrimmed videos. Earlier works on activity detection mainly used temporal sliding windows as candidates and trained activity classifier on hand-crafted features [35, 13, 23, 33, 46]. With the vast successes of deep learning methods, two-stream networks [44, 12, 49], 3D ConvNet [47] and other deep neural networks [4, 48, 38, 51, 9] have been proposed to model video sequences and significantly improved recognition performance. To better localize temporal boundaries, a large body of work incorporated deep networks into the detection framework and obtained improved performance [5, 28, 10, 41, 2, 61, 42, 53, 59]. Among these works, S-CNN [42] proposed a multi-stage CNN which adopted 3D ConvNet with multi-scale sliding window; R-C3D [53] proposed an end-to-end trainable activity detector based on Faster-RCNN [39]; S^3D [59] performed single-shot activity detection to get rid of explicit temporal proposals.

However, most of these methods have focused on detecting a fixed set of activity classes without language queries. In this paper, we propose to build a highly-integrated retrieval framework and adopt a similar single-shot encoding scheme inspired by the single-shot detectors [30, 59, 28].

Natural Language Moment Retrieval. The natural language moment retrieval is a new task introduced recently [18, 14]. The methods proposed in [18, 14] learn a common embedding space shared by video segment features and sentence representations and measure their similarities through sliding window [14] or handcrafted heuristics [18]. While simple and effective, these methods fail to consider the challenging alignment problems.

Until recently, several methods were proposed to closely integrate language and video representation [54, 6]: Xu *et al*. [54] proposed multilevel language and video feature fusion; TGN [6] applied frame-by-word interactions between video and language and obtained improved performance. Although these works share the same spirit with ours to better align semantic information, they fail to reason the complex cross-modal relations. Our work is the first to model both semantic and structural relations together in an unified network, allowing us to directly learn the complex temporal relations in an end-to-end manner.

Visual Relations and Graph Network. Reasoning about the pairwise relationships has been proven to be very helpful in a variety of computer vision tasks [16, 57, 58, 8]. Recently, visual relations have been combined with deep neural networks in areas such as object recognition [21, 11], visual question answering [40] and action recognition [31, 34]. A variety of papers have considered modeling spatial relations in natural images [7, 22, 37], and scene graph is widely used in the image retrieval tasks [24, 56]. In the field of natural language moment retrieval: Liu *et al*. [29] proposed to parse sentence structure as a dependency tree and construct a temporal modular network accordingly; Hendricks *et al*. [19] modeled video context as a latent variable

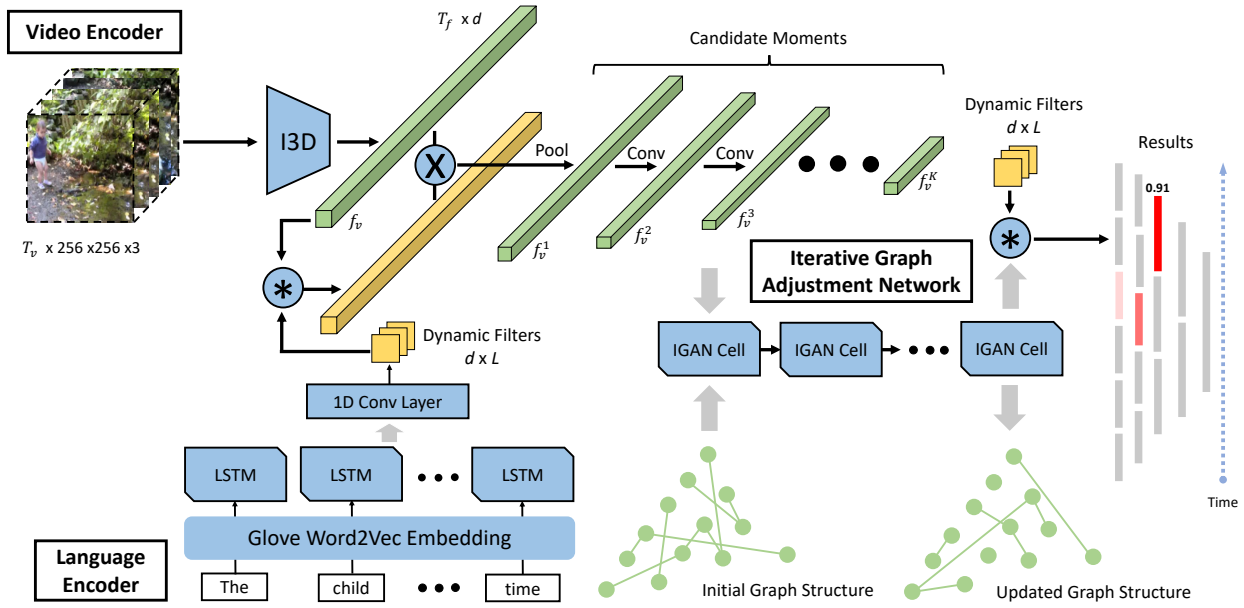

Figure 2: An overview of our end-to-end Moment Alignment Network (MAN) for natural language moment retrieval (best viewed in color). MAN consists three major components: (1) A language encoder to convert the input language query to dynamic convolutional filters through a single-layer LSTM. (2) A video encoder to produce multi-scale candidate moment representations in a hierarchical fully-convolutional network, where input visual features are aligned with language semantics by convolution. (3) An iterative graph adjustment network to directly model moment-wise temporal relations and update moment representations. Finally, the moments are retrieved by its matching scores with the language query.

to reason about the temporal relationships. However, their reasoning relies on a hand-coded structure, thus, fail to directly learn complex temporal relations.

Our work is inspired by the GCN [26] and other successful graph-based neural networks [32, 55]. While the original GCN is proposed to reason on a fixed graph structure, we modify the architecture to jointly optimize relations together. That is, instead of fixing the temporal relations, we learn it from the data.

3. Model

In this work, we address the natural language moment retrieval task. Given a video and a natural language description as a query, we aim to retrieve the best matching temporal segment (*i.e.*, moment) as specified by the query. To specifically handle the semantic and structural misalignment between video and language, we propose Moment Alignment Network (MAN), a novel framework combining both video and language information in a single-shot structure to directly output matching scores between moment and language query through temporal structure reasoning. As illustrated in Figure 2, our model consists of three main components: a language encoder, a video encoder and an iterative graph adjustment network. We introduce the details of each component and network training in this section.

3.1. Language Encoding as Dynamic Filters

Given an input of a natural language sentence as a query that describes the moment of interest, we aim to encode it so that we can effectively retrieve specific moment in video. Instead of encoding each word with a one-hot vector or learning word embeddings from scratch, we rely on word embeddings obtained from a large collection of text documents. Specifically, we use the Glove [36] word2vec model pre-trained on Wikipedia. It enables us to model complex linguistic relations and handle words beyond the ones in the training set. To capture language structure, we use a single-layer LSTM network [20] to encode input sentences. In addition, we leverage the LSTM outputs at all time steps to seek more fine-grained interactions between language and video. We also study the effects of using word-level or sentence-level encoding in our ablation study.

In more detail, a language encoder is a function $F_l(\omega)$ that maps a sequence of words $\omega = \{w_i\}_{i=1}^{L}$ to a semantic embedding vector $f_l \in \mathbb{R}^{L \times d}$, where L is the number of words in a sentence and d is the feature dimension, and F_l is parameterized by Glove and LSTM in our case.

Moreover, to transfer textual information to the visual domain, we rely on dynamic convolutional filters as earlier used in [27, 15]. Unlike static convolutional filters that are used in conventional neural networks, dynamic filters are

generated depending on the input, in our case on the encoded sentence representation. As a general convolutional layer, dynamic filters can be easily incorporated with the video encoder as an efficient building block.

Given a sentence representation $f_l \in \mathbb{R}^{L\times d}$, we generate a set of word-level dynamic filters $\{\Gamma_i\}_{i=1}^{L}$ with a single fully-connected layer:

$$\Gamma_i = tanh(W_\Gamma f_l^i + b_\Gamma) \tag{1}$$

where $f_l^i \in \mathbb{R}^d$ is the word-level representation at index i, and for simplicity, Γ_i is designed to have the same number of intput channels as f_l^i. Thus, by sharing the same transformation for all words, each sentence representation $f_l \in \mathbb{R}^{L\times d}$ can be converted to a dynamic filter $\Gamma \in \mathbb{R}^{d\times L}$ through a single 1D convolutional layer.

As illustrated in Figure 2, we convolve the dynamic filters with the input video features to produce a semantically-aligned visual representation, and also with the final moment-level features to compute the matching scores. We detail our usage in Section 3.2 and Section 3.3, respectively.

3.2. Single-Shot Video Encoder

Existing solutions for natural language moment retrieval heavily relies on handcrafted heuristics [18] or temporal sliding windows [14] to generate candidate segments. However, the temporal sliding windows are typically too dense and often times designed with multiple scales, resulting in a heavy computation cost. Processing each individual moment separately also fails to efficiently leverage semantic and structural relations between video and language.

Inspired by the single-shot object detector [30] and its successful applications in temporal activity detection [59, 28], we apply a hierarchical convolutional network to directly produce multi-scale candidate moments from the input video stream. Moreover, for the natural language moment retrieval task, the visual features itself undoubtedly play the major role in generating candidate moments, while the language features also help to distinguish the desired moment from others. As such, a novel feature alignment module is especially devised to filter out unrelated visual features from language perspective at an early stage. We do so by generating convolutional dynamic filters (Section 3.1) from the textual representation and convolving them with the visual representations. Similar to other single shot detectors, all these components are elegantly integrated into one feed-forward CNN, aiming at naturally generating variable-length candidate moments aligned with natural language semantics.

In more detail, given an input video, we first obtain a visual representation that summarizes spatial-temporal patterns from raw input frames into high-level visual semantics. Recently, Dai *et al*. proposed to decompose 3D convolutions into aggregation blocks to better exploit the spatial-temporal nature of video. We adopt the TAN [9] model to obtain a visual representation from video. As illustrated in Figure 2, an input video $V = \{v_t\}_{t=1}^{T_v}$ is encoded into a clip-level feature $f_v \in \mathbb{R}^{T_f\times d}$ where T_f is the total number of clips and d is the feature dimension. For simplicity, we set f_v and f_l to have the same number of channels. While f_v should be sufficient for building advanced recognition model [53, 28, 60], the crucial alignment information between language and vision is missing specifically for natural language moment retrieval.

As such, the dynamic convolutional filters are applied to fill the gap. We convolve the dynamic filter Γ with f_v to obtain a clip-wise response map M, and M is further normalized to augment the visual feature. Formally, the augmented feature f_v' is computed as:

$$\begin{aligned} M &= \Gamma * f_v \in \mathbb{R}^{T_v\times L} \\ M_{norm} &= softmax(sum(M)) \in \mathbb{R}^{T_v} \\ f_v' &= M_{norm} \odot f_v \in \mathbb{R}^{T_v\times d} \end{aligned} \tag{2}$$

where $\odot$ denotes matrix-vector multiplication.

To generate variable-length candidate moments, we follow similar design of other single-shot detectors [30, 59] to build a multi-scale feature hierarchy. Specifically, a temporal pooling layer is firstly devised on top of f_v' to reduce the temporal dimension of feature map and increase temporal receptive field, producing the output feature map of size $T_v/p \times d$ where p is the pooling stride. Then, we stack K more 1D convolutional layers (with appropriate pooling) to generate a sequence of feature maps that progressively decrease in temporal dimension which we denote as $\{f_v^k\}_{k=1}^{K}, f_v^k \in \mathbb{R}^{T_k\times d}$ where T_k is the temporal dimension of each layer. Thus each temporal feature cell is responsive to a particular location and length, and therefore corresponds to a specific candidate moment.

3.3. Iterative Graph Adjustment Network

To encode complex temporal dependencies, we propose to model moment-wise temporal relations in a graph to explicitly utilize the rich relational information among moments. Specifically, candidate moments are represented by nodes, and their relations are defined as edges. Since we gather $N = \sum_{k=1}^{K} T_k$ candidate moments in total each represented by a d-dimensional vector, we denote the node feature matrix as $f_m \in \mathbb{R}^{N\times d}$. To perform reasoning on the graph, we aim to apply the GCN proposed in [26]. Different from the standard convolutions which operate on a local regular grid, the graph convolutions allow us to compute the response of a node based on its neighbors defined by the graph relations. In the general form, one layer of graph convolutions is defined as:

$$H = ReLU(GXW) \tag{3}$$

where $G \in \mathbb{R}^{N\times N}$ is the adjacency matrix, $X \in \mathbb{R}^{N\times d}$ is the input features of all nodes, $W \in \mathbb{R}^{d\times d}$ is the weight matrix and $H \in \mathbb{R}^{N\times d}$ is the updated node representation.

However, one major limitation of the GCN applied in our scenario is that it can only reason on a fixed graph structure. To fix this issue, we introduce the Iterative Graph Adjustment Network (IGAN), a framework based on GCN but with a learnable adjacency matrix, that is able to simultaneously infer a graph by learning the weight of all edges and update each node representation accordingly. In more detail, we iteratively updates the adjacency matrix as well as node features in a recurrent manner. The IGAN model is fully differentiable thus can be efficiently learned from data in an end-to-end manner.

In order to jointly learn the node representation and graph structure together, we propose certain major modifications to the original GCN block: (1) Inspired by the successful residual network [17], we decompose the adjacency matrix into a preserving component and a residual component. (2) The residual component is produced from the node representation similar to a decomposed correlation [3]. (3) In a recurrent manner, we iteratively accumulate residual signals to update the adjacency matrix by feeding updated node representations. The overall architecture of a single IGAN cell is illustrated in the top half of Figure 3 and the transition function is formally defined as:

$$\begin{aligned} R_t &= norm(X_{t-1}W_t^r X_{t-1}^T) \\ G_t &= tanh(G_{t-1} + R_t) \\ X_t &= ReLU(G_t X_0 W_t^o) \end{aligned} \tag{4}$$

where $X_0 = f_m$ is the input candidate moment features, R_t is the residual component derived from the output of previous cell X_{t-1}, $norm()$ denotes a signed square root followed by a L2 normalization to normalize the features, and W_t^r and W_t^o are learnable weights. Note that the candidate moment features X_0 is the output of a hierarchical convolutional network combined with language information, thus can be jointly updated with the IGAN.

In our design, the initial adjacency matrix G_0 is set as a diagonal matrix to emphasize self-relations. we stack multiple IGAN cells as shown in the bottom half of Figure 3 to update the candidate moment representations as well as the moment-wise graph structure. Finally, we convolve the dynamic filter Γ with the final output X_T to compute the matching scores. We further study the effects of IGAN in our ablation study.

3.4. Training

Our training sample consists of an input video, an input language query and a ground truth best matching moment annotated with start and end time. During training, we need to determine which candidate moments correspond to a ground truth moment and train the network accordingly. Specifically, for each candidate moment, we compute the temporal IoU score with ground truth moment. If the temporal IoU is higher than 0.5, we regard the candidate moment as positive, otherwise negative. After matching each candidate moment with the ground truth, we derive a ground truth matching score s_i for each candidate moment.

For each training sample, the network is trained end-to-end with a binary classification loss using sigmoid cross-entropy. Rather than using a hard score, we use the temporal IoU score s_i as ground truth for each candidate moment. The loss is defined as:

$$\mathcal{L} = -\frac{1}{N_b}\sum_{i}^{N_b}(s_i \log(a_i) + (1-s_i)\log(1-a_i)) \tag{5}$$

where N_b is the number of total training candidate moments in a batch, a_i is the predicted score and s_i is the ground truth score.

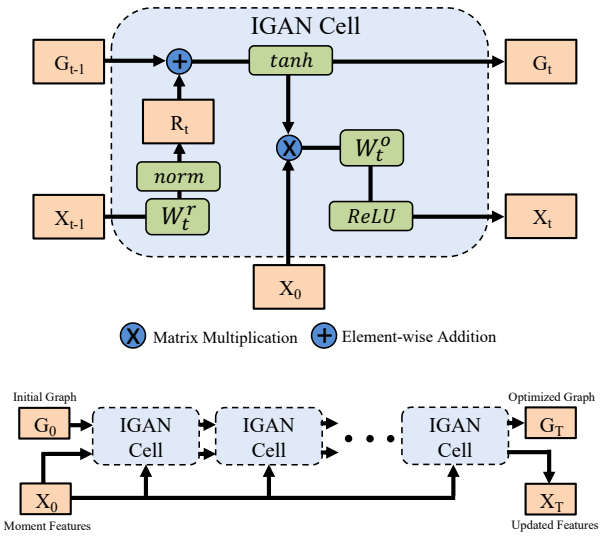

Figure 3: The structure of the proposed Iterative Graph Adjustment Network (IGAN). Top: In each IGAN cell, a residual component R_t is generated from the previous node representation X_{t-1} and aggregated with the preserving component G_{t-1} to produce the current adjacency matrix G_t. Node representations are updated according to Equation 3 with G_t, X_0 and W_t^o. Bottom: Multiple IGAN cells are connected to simultaneously update node representation and graph structure.

4. Experiments

We evaluate the proposed approach on two recent large-scale datasets for the natural language moment retrieval

task: DiDeMo [18] and Charades-STA [14]. In this section we first introduce these datasets and our implementation details and then compare the performance of MAN with other state-of-the-art approaches. Finally, we investigate the impact of different components via a set of ablation studies and provide visualization examples.

4.1. Datasets

DiDeMo The DiDeMo dataset was recently proposed in [18], specially for natural language moment retrieval in open-world videos. DiDeMo contains more than $10,000$ videos with $33,005$, $4,180$ and $4,021$ annotated moment-query pairs in the training, validation and testing datasets respectively. To annotate moment-query pairs, videos in DiDeMo are trimmed to a maximum of 30 seconds, divided into 6 segments of 5 seconds long each, and each moment contains one or more consecutive segments. Therefore, there are 21 candidate moments in each video and the task is to select the moment that best matches the query.

Following [18], we use Rank-1 accuracy (Rank@1), Rank-5 accuracy (Rank@5) and mean Intersection-over-Union (mIoU) as our evaluation metrics.

Charades-STA The Charades-STA [14] was another recently collected dataset for natural language moment retrieval in indoor videos. Charades-STA is built upon the original Charades [43] dataset. While Charades only provides video-level paragraph description, Charades-STA applies sentence decomposition and keyword matching to generate moment-query annotation: language query with start and end time. Each moment-query pair is further verified by human annotators. In total, there are $12,408$ and $3,720$ moment-query pairs in the training and testing datasets respectively. Since there is no pre-segmented moments, the task is to localize a moment with predicted start and end time that best matches the query.

We follow the evaluation setup in [14] to compute "R@n, IoU@m", defined as the percentage of language queries having at least one correct retrieval (temporal IoU with ground truth moment is larger than m) in the top-n retrieved moments. Following standard practice, we use $n \in \{1, 5\}$ and $m \in \{0.5, 0.7\}$.

4.2. Implementation Details

We train the whole MAN model in an end-to-end manner, with raw video frames and natural language query as input. For *language encoder*, each word is encoded as a 300-dimensional Glove word2vec embedding. All the word embeddings are fixed without fine-tuning and each sentence is truncated to have a maximum length of 15 words. A single-layer LSTM with $d = 512$ hidden units is applied to obtain the sentence representation. For *video encoder*, TAN [9] is used for feature extraction. The model takes as input a clip of 8 RGB frames with spatial size 256×256 and extracts a 2048-dimensional representation as output of an average pooling layer. We add another 1D convolutional layer to reduce the feature dimension to $d = 512$. Each video is decoded at 30 FPS and clips are uniformly sampled among the whole video. On Charades, we sample $T_f = 256$ clips and set the pooling stride $p = 16$ and apply a sequence of 1D convolutional filters (pooling stride 2) to produce a set of $\{16, 8, 4, 2, 1\}$ candidate moments, resulting in 31 candidate moments in total. Similarly, on DiDeMo, in order to match with the pre-defined temporal boundary, we sample $T_f = 240$ clips and set pooling stride $p = 40$ with a sequence of 1D convolutional filters (pooling is adjusted accordingly) to produce a set of $\{6, 5, 4, 3, 2, 1\}$ candidate moments, resulting in 21 candidate moments in total. For both datasets, we apply 3 IGAN cells. We implement our MAN on TensorFlow [1]. The whole system is trained by Adam [25] optimizer with learning rate 0.0001.

Method	Rank@1	Rank@5	mIoU
TMN [29]	18.71	72.97	30.14
TGN [6]	24.28	71.43	38.62
MCN [18]	24.42	75.40	37.39
MAN(ours)	**27.02**	**81.70**	**41.16**

Table 1: Natural language moment retrieval results on DiDeMo dataset. MAN outperforms previous state-of-the-art mehtods by $\sim 3\%$ among all metrics.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Random [14]	8.51	3.03	37.12	14.06
CTRL [14]	21.42	7.15	59.11	26.91
Xu *et al*. [54]	35.60	15.80	79.40	45.40
MAN(ours)	**46.53**	**22.72**	**86.23**	**53.72**

Table 2: Natural language moment retrieval results on Charades-STA dataset. MAN significantly outperforms previous state-of-the-art methods by a large margin.

4.3. Comparison with State-of-the-art

We compare our MAN with other state-of-the-art methods on DiDeMo [18] and Charades-STA [14]. Note that the video content and language queries differ a lot among two different datasets. Hence, strong adaptivity is required to perform consistently well on both datasets. Since our MAN only takes raw RGB frames as input and doesn't rely on external motion features such as optical flow, for a fair comparison, all compared methods use RGB features only.

DiDeMo Table 1 shows our natural language moment retrieval results on the DiDeMo dataset. We compare with state-of-the-art methods published recently including the

Method	Rank@1	Rank@5	mIoU
Base	23.56	77.66	36.36
Base+FA(1)	24.45	78.69	37.72
Base+FA(L)	25.10	79.57	38.78
Base+FA+IGANx1	25.67	79.36	39.13
Base+FA+IGANx2	26.10	80.08	40.21
Base+FA+IGANx3	**27.02**	**81.70**	**41.16**

Table 3: Ablation study for effectiveness of MAN components: Top: Advantage of a single-shot video encoder. Mid: Effectiveness of the feature alignment. Bottom: Importance of the IGAN.

methods that use temporal modular network [29], fine-grained frame-by-word attentions [6] and temporal contextual encoding [18]. Among all three evaluation metrics, the proposed method outperforms previous state-of-the-art methods by around 3% in absolute values.

Charades-STA We also compare our method with the recent state-of-the-art methods on Charades-STA dataset. The results are shown in Table 2, where CTRL [14] applies a cross-modal regression localizer to adjust temporal boundaries and Xu *et al*. [54] even boosts the performance with more closely multilevel language and vision integration. Our model tops all the methods among all evaluation metrics and significantly improves R@1, IoU=0.5 by over 10% in absolute values.

4.4. Ablation Studies

To understand the proposed MAN better, we evaluate our network with different variants to study their effects.

Network Components. On DiDeMo dataset, we perform ablation studies to investigate the effect of each individual component we proposed in this paper: single-shot video encoder, feature alignment with language query and iterative graph adjustment network.

Single-shot video encoder. In this work, we introduced a single-shot video encoder using hierarchical convolutional network for the natural language moment retrieval task. To study the effect of this architecture alone, we build a **Base** model which is the same as we described in Section 3.2 except for two modifications: (1) We remove the feature alignment component (Equation 2) and directly use the visual feature f_v to construct the network. (2) We remove all IGAN cells on top and directly feed f_m to compute matching scores. The result is reported in the top line in Table 3, even with only a single-shot encoding scheme, we achieve 23.56% on Rank@1 and 77.66% on Rank@5 which is better or competitive with other state-of-the-art methods.

Dynamic filter. We further validate our design to augment the input clip-level features with dynamic filters. The results are shown in the middle part in Table 3. On top of the Base model, we study two different variants: (1) Construct a sentence-level dynamic filter where only the last LSTM hidden state is used for feature alignment, denoted as **Base+FA(1)**. (2) Construct word-level dynamic filters where all LSTM hidden states are converted to a multi-channel filter for feature alignment, denoted as **Base+FA(L)**. We observe that Base+FA(1) already improves the accuracy compared to the base model, which indicates the importance of adding feature alignment in our model. Moreover, adding more fine-grained word-level interactions between video and language can further improve the performance.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Xu *et al*. [54]	35.60	15.80	79.40	45.40
MAN-VGG	41.24	20.54	83.21	51.85
MAN-TAN	**46.53**	**22.72**	**86.23**	**53.72**

Table 4: Ablation study on different visual features. MAN with VGG-16 features already outperforms state-of-the-art method, and TAN features further boost the performance.

Iterative graph adjustment network. A major contribution of MAN is using the IGAN cell to iteratively update graph structure and learned representation. We measure the contribution of this component to the retrieval performance in the bottom section in Table 3, where **Base+FA+IGANxn** denotes our full model with n IGAN cells. The result shows a decrease in performance with fewer IGAN cells, dropping monotonically from 27.02% to 25.67% on Rank@1. This is because the temporal relations represented in a moment graph structure can be iteratively optimized thus more IGAN cells result in better representation for each candidate moment. Despite the performance gain, we also notice that Base+FA+IGANx3 converges faster and generalizes better with smaller variance.

Visual Features. We conduct experiments to study the effect of different visual features on Charades-STA dataset. We consider two different visual features: (1) Two-stream RGB features [44] from the original Charades dataset, which is a frame-level feature from VGG-16 [45] network, we denote the model as **MAN-VGG**. (2) TAN features as described in the paper, which is a clip-level feature from aggregation blocks, we denote the model as **MAN-TAN**. The results are summarized in Table 4. It can be seen that TAN features outperform VGG-16 features among all evaluation metrics, this is consistent with the fact that better base network leads to better overall performance. But more interestingly, while the overall performance using only VGG visual features is noticeably lower than using TAN features, our **MAN-VGG** model already significantly outperforms the state-of-the-art method. Since frame-level VGG-16 net-

Figure 4: Qualitative visualization of the natural language moment retrieval results (Rank@1) by MAN (best viewed in color). First example is from Charade-STA dataset, and second example is from DiDeMo dataset. Ground truth moments are marked in black and retrieved moments are marked in green.

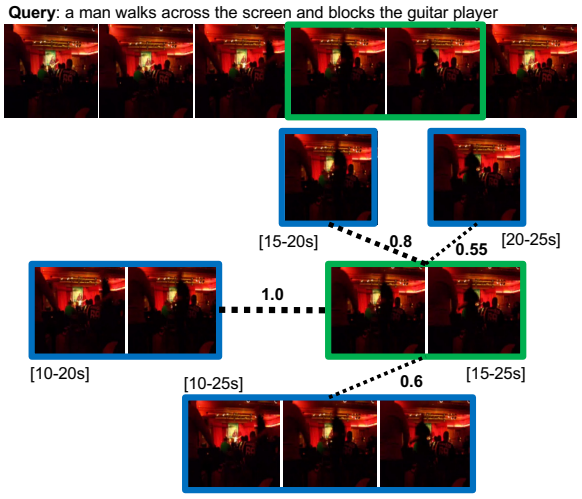

Figure 5: Qualitative example of MAN evaluated on a video-query pair (best viewed in color). The final moment-wise graph structure with top related edges and their corresponding moments is visualized. The retrieved moment is marked in green and other moments are marked in blue. The dashed line indicates the strength of each edge with the highest one normalized to 1.0.

work provides no motion information when extracting features, this superity highlights MAN's strong ability to perform semantic alignment and temporal structure reasoning.

Visualization. *Qualitative Results.* We provide qualitative results to demonstrate the effectiveness and robustness of the proposed MAN framework. As shown in Figure 4, MAN is capable of retrieving a diverse set of moments including the one requiring strong temporal dependencies to identify "woman shows her face for the first time". The advantage of MAN is best pronounced for tasks that rely on reasoning complex temporal relations.

Graph Visualization. An advantage of a graph structure is its interpretability. Figure 5 visualizes the final moment-wise graph structure learned by our model. In more detail, Figure 5 displays a 30-second video where "man walks" from 10 to 30 seconds and "blocks the guitar player" from 15 to 25 seconds. MAN is able to concentrate on those moments with visual information related to "man walks across the screen". It also reasons among multiple similar moments including some incomplete moments (15-20s, 20-25s) and some other moments partially related to "blocks the guitar player" (10-20s, 10-25s) to retrieve the one best matching result (15-25s).

5. Conclusion

We have presented MAN, a Moment Alignment Network that unifies candidate moment encoding and temporal structural reasoning in a single-shot structure for natural language moment retrieval. Particularly, we identify two key challenges (*i.e.* semantic misalignment and structural misalignment) and study how to handle such challenges in a deep learning framework. To verify our claim, we propose a fully convolutional network to force cross-modal alignments and an iterative graph adjustment network is devised to model moment-wise temporal relations in an end-to-end manner. With this framework, We achieved state-of-the-art performance on two challenging benchmarks Charades-STA and DiDeMo.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[2] S Buch, V Escorcia, B Ghanem, L Fei-Fei, and JC Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[3] Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*, pages 430–443. Springer, 2012.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017.

[5] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018.

[6] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 162–171, 2018.

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3298–3308. IEEE, 2017.

[8] Xiyang Dai, Joe Yue-Hei Ng, and Larry S Davis. Fason: First and second order information fusion network for texture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7352–7360, 2017.

[9] Xiyang Dai, Bharat Signh, Joe Yue-Hei Ng, and Larry S. Davis. Tan: Temporal aggregation network for dense multi-label action recognition. In *WACV*, 2018.

[10] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5727–5736. IEEE, 2017.

[11] Georgia Gkioxari Ross Girshick Piotr Dollár and Kaiming He. Detecting and recognizing human-object interactions. *CVPR*, 2018.

[12] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[13] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2782–2795, 2013.

[14] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275, 2017.

[15] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966, 2018.

[16] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017.

[19] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018.

[22] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4418–4427. IEEE, 2017.

[23] Mihir Jain, Jan Van Gemert, Hervé Jégou, Patrick Bouthemy, and Cees GM Snoek. Action localization with tubelets from motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 740–747, 2014.

[24] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[26] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[27] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, Arnold WM Smeulders, et al. Tracking by natural language specification. In *CVPR*, volume 1, page 5, 2017.

[28] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 988–996. ACM, 2017.

[29] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular

networks for retrieving complex compositional activities in videos. In *European Conference on Computer Vision*, pages 569–586. Springer, 2018.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[31] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. *CVPR*, 2018.

[32] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20–28. IEEE, 2017.

[33] Pascal Mettes, Jan C van Gemert, Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. Bag-of-fragments: Selecting and encoding video fragments for event detection and recounting. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 427–434. ACM, 2015.

[34] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2016.

[35] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Proceedings of the IEEE international conference on computer vision*, pages 1817–1824, 2013.

[36] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[37] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *ICCV 2017-International Conference on Computer Vision 2017*, 2017.

[38] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5534–5542. IEEE, 2017.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[40] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[41] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1417–1426. IEEE, 2017.

[42] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[43] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016.

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[46] Kevin Tang, Bangpeng Yao, Li Fei-Fei, and Daphne Koller. Combining the right features for complex event recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2696–2703, 2013.

[47] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[48] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[49] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.

[50] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[51] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018.

[52] Xin Wang, Jiawei Wu, Da Zhang, Yu Su, and William Yang Wang. Learning to compose topic-aware mixture of experts for zero-shot video captioning. *arXiv preprint arXiv:1811.02765*, 2018.

[53] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 6, page 8, 2017.

[54] Huijuan Xu, Kun He, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Multilevel language and vision integration for text-to-clip retrieval. AAAI, 2019.

[55] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018.

[56] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. *ECCV*, 2018.

[57] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 17–24. IEEE, 2010.

[58] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 702–709. IEEE, 2012.

[59] Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. S3d: Single shot multi-span detector via fully 3d convolutional network. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[60] Da Zhang, Xiyang Dai, and Yuan-Fang Wang. Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. *ACCV*, 2018.

[61] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2933–2942. IEEE, 2017.