Modeling Motion with Multi-Modal Features for Text-Based Video Segmentation

Wangbo Zhao[1,2,3] Kai Wang[1] Xiangxiang Chu[2] Fuzhao Xue[1] Xinchao Wang[1] Yang You[1*]
[1] National University of Singapore [2] Meituan Inc. [3] Northwestern Polytechnical University
wangbo.zhao96@gmail.com, kai.wang@comp.nus.edu.sg, chuxiangxiang@meituan.com,
f.xue@u.nus.edu, xinchao@nus.edu.sg, youy@comp.nus.edu.sg

Abstract

Text-based video segmentation aims to segment the target object in a video based on a describing sentence. Incorporating motion information from optical flow maps with appearance and linguistic modalities is crucial yet has been largely ignored by previous work. In this paper, we design a method to fuse and align appearance, motion, and linguistic features to achieve accurate segmentation. Specifically, we propose a multi-modal video transformer, which can fuse and aggregate multi-modal and temporal features between frames. Furthermore, we design a language-guided feature fusion module to progressively fuse appearance and motion features in each feature level with guidance from linguistic features. Finally, a multi-modal alignment loss is proposed to alleviate the semantic gap between features from different modalities. Extensive experiments on A2D Sentences and J-HMDB Sentences verify the performance and the generalization ability of our method compared to the state-of-the-art methods.

1. Introduction

Text-based video segmentation aims at locating and segmenting the object described by a language sentence in a video sequence. Unlike traditional tasks, which do prediction on video- or frame- level, *e.g.* text-to-video retrieval [27, 40, 52], video caption [32, 62], video question answering [22, 51], and language-queried video localization [1, 60], this task requires relatively more fine-grained multi-modal and temporal understanding for pixel-level segmentation. The challenge of this task can be thus summarized as: (1) how to reason between visual and linguistic modalities to locate the target object, and (2) how to leverage temporal information to enhance segmentation.

To solve the former problem, previous works adopt simple concatenation [18], generating dynamic filters [16, 48] and cross-modal attention modules [21, 49] to achieve interactions between two modalities. When it comes to the latter problem, they usually adopt 3D convolution neural networks (3D CNNs) *e.g.* I3D [8] to extract features from a video clip. However, all these methods ignore exploring the explicit motion information between frames for text-based video segmentation. In this task, the target object usually has action, and the corresponding text contains some words to describe its motion *e.g.* driving and jumping in Figure 1. This means that the motion information may help the model to find the target object. Despite the fact that some motion information between frames can be implicitly learned in 3D CNNs, it can not well interact with other modalities. Introducing motion information has been tried in some video tasks [9, 15, 24, 30, 53, 61, 63], but how to incorporate the motion information with appearance and linguistic features in text-based video segmentation is still challenging.

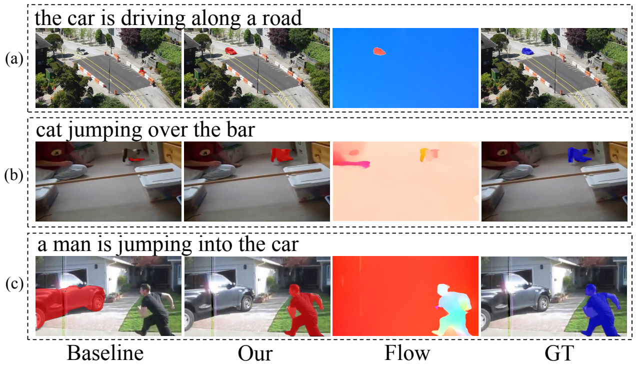
the car is driving along a road
(a)
cat jumping over the bar
(b)
a man is jumping into the car
(c)
Baseline
Our
Flow
GT

Figure 1. Comparison between baseline and our model. We adopt ”B” in 4.4 as the baseline model. Compared with the baseline model, our model can incorporate motion information from optical flow maps with appearance and linguistic features and generate better segmentation masks.

A common way to introduce explicit motion information is to extract features from flow maps generated from an optical flow estimation model. From flow maps in Figure 1, we can find that the target object with motion usually is distinctive and can be easily identified. This may promote the final performance. To leverage the motion information from optical flow, Gavrilyuk *et al.* [16] adopt two 3D CNNS with different parameters to generate masks from

[*] Corresponding author.
Our code is publicly available at: https://github.com/wangbo-zhao/2022CVPR-MMMMTBVS.

RGB frames and optical flow maps, respectively, then compute weighted averaged masks from them. However, such a simple fusion strategy ignores the interaction between motion modalities and appearance and linguistic features, leading to unsatisfactory improvement and huge computational overhead. Hence, designing a model to effectively incorporate the motion information from the optical flow with appearance features from RGB frames and linguistic features is necessary.

Motivated by observations above, we propose our multi-modal fusion and alignment network. First, since many previous works [3, 6, 9, 19] have demonstrated the superiority of transformers in reasoning and fusing multi-modal and temporal features, we build a multi-modal video transformer (MMVT) to model the interaction between appearance, motion, and linguistic features in different frames. Our transformer contains two attention modules in each layer: cross-modal attention and temporal attention. The former aims at fusing three modalities features, while the latter is adopted to aggregate fused features in the temporal dimension. By stacking them several layers, multi-modal information can flow and interact with each other between different frames. Benefiting from the multi-modal interaction between frames in MMVT, we do not rely on 3D CNNs to extract temporal information, which largely reduce the computational overhead.

Then, to fuse multi-modal features progressively, we propose the language-guided feature fusion (LGFF) module and insert it into each level to decode features. In each module, useful appearance and motion features will be selected by the linguistic feature, with the help of features from the higher level. By doing this, useful features can be gradually selected and fused. Moreover, since both appearance, motion, and language features are distinctive modalities features, which are generated from backbones separately pretrained on different datasets, the semantic gap between them would be large [20]. To alleviate this problem, we design a multi-modal alignment loss, which explicitly encourages the network to learn to align three modalities features in an embedding space, which further improves the performance of our model.

In Figure 1, compared with the baseline model without motion information, our model can accurately locate the target object, obtain a more complete mask, and distinguish the target object from others. Our main contributions can be summarised as:

- To the best of our knowledge, we are the first to incorporate the motion information from optical flow maps with appearance and linguistic features for text-based video segmentation.
- We propose a transformer-based model to fuse multi-modal and temporal features and design a language-guided feature fusion module to progressively fuse multi-modal features from different feature levels.
- Noticing the semantic gap between different modal features, we propose a multi-modal alignment loss to explicitly align features from three different modalities, which further improve the performance of our method.
- Extensive experiments are conducted to verify the effectiveness of proposed methods. Our approach significantly surpasses existing state-of-the-art methods on most metrics on A2D Sentences and J-HMDB Sentences dataset with less computational overhead.

2. Related Work

Text-Based Image Segmentation Text-based image segmentation aims to segment the object in an image given a text describing its properties *e.g*. appearance and location. Hu *et al*. [18] are the first to propose this task, and they adopt the fully convolutional network to fuse extracted visual and linguistic features directly. Liu *et al*. [34] propose a multi-modal LSTM to force the word-visual interaction. Ye *et al*. [56] design a self-attention module to capture long-range relationships between two modalities. Luo *et al*. [37] propose a model to achieve joint learning of locating and segmentation since these two tasks can reinforce each other. Jing *et al*. [25] decouple this task into locating the target object position and accurately generating the segmentation mask. Yang *et al*. [54] represents the expression as a language graph and performs explainable visual reasoning to distinguish the target object from others. Ding *et al*. [13] introduce the encoder-decoder attention mechanism in transformer [46] and view the language expression as queries.

Unlike these works for images, which only need to focus on fusing features from the static RGB image and the language expression, we conduct multi-modal fusion between the RGB image, flow map, and text. In addition, we also consider the temporal information between adjacent frames.

Text-Based Video Segmentation For promoting comprehensive action understanding, Xu *et al*. [50] release a dataset named Actor-Action Dataset (A2D) containing a fixed vocabulary of actor and action pairs and pixel-level annotations. After that, Gavrilyuk *et al*. [16] further extend this dataset and propose text-based video segmentation. They generate dynamic filters from extracted text features and adopt them to convolve with vision features to obtain the final pixel-wise segmentation. They also try to average the masks from an optical flow map and an RGB frame to improve the performance further. Wang *et al*. [49] propose a cross-guided attention mechanism, where features from frames and the text can guide and promote each other. This design can reduce linguistic variation and incorporate query-focused visual features. Mcintosh *et al*. [38] propose a capsule-based network to encode and merge visual and textual features jointly. Wang *et al*. [48] introduce the idea of deformable convolution [11] into generating dynamic filters to address geometric deformation. Ning *et al.* [39] propose a polar positional encoding mechanism to measure the spatial relations in terms of direction and range, which is similar to natural language descriptions. Hui *et al.* [21] adopt 3D and 2D encoders to recognize the queried actions and accurately segment target object, respectively.

Different from [16], which ignores the interaction between motion information and other modalities, the motion information can be well fused and interact with appearance and linguistic features in our MMVT and LGFF.

Vision-Language Learning Tasks Owing to the development of NLP and CV tasks, more and more researchers are starting to explore the image-language, and video-language tasks [2, 28, 44, 47]. The latter is more related to our task since it requires exploring the information in the temporal dimension. Many attempts [29, 31, 42, 43] have been done on video-language following a pretraining then fine-tuning manner. They first adopt some proxy tasks to train the model in an self-supervised manner, categorized into completion, matching and ordering. Then, the well-learned representations should be transferred to downstream tasks, e.g. text-based video retrieval [58], action step localization [65], video question answering [44]. More details about vision-language learning tasks can be found in the survey [41].

Tasks mentioned above usually make video-level or frames-level prediction and do not require fine-grained features. In contrast, text-based video segmentation requires to predict on pixel-level. So pre-trained video-language models can not be directly applied to our task.

Vision Transformer Vaswani *et al.* [46] first propose the transformer, which shows its predominance in many Natual Language Processing (NLP) tasks. The main component of transformers is the self-attention mechanism, which can model long-range dependencies in the data. Computer vision community views this advantage and attempt to design transformer-based models for image classification [14, 36, 55, 57, 59], object detection [7, 64] and video understanding [3, 6]. Transformers have also been introduced into some multi-modal tasks. Hu *et al.* [19] propose a unified transformer model jointly trained on multiple tasks, including not only vision-only and language-only tasks but also vision-and-language reasoning. Chen *et al.* [9] adopt a multi-modal video transformer to collaboratively fuse appearance, motion and audio features for video action recognition. Liu *et al.* [35] adopt two transformers to extract appearance and depth information for saliency detection.

In this paper, we propose a transformer-based module that contains cross-modal attention and temporal attention. The former incorporates motion modalities with appearance and linguistic features, and the latter focuses on aggregating temporal information.

3. Method

The overall architecture of the proposed method is shown in Figure 2. For a video sequence, we have T frames, their corresponding flow maps, and the text, which describe the target object and its action. First, we adopt three encoders to extract appearance, motion, and language features, respectively. Then, the extracted three kinds of high-level features will be concatenated together and input into our multi-modal video transformer (MMVT) to fuse cross-modal features and build temporal relationships between frames. In the decoder, appearance and motion features from different levels will be progressively fused with language features in our language-guided feature fusion module (LGFF) and predict the final segmentation mask. During training, a multi-modal alignment loss is added to align features from different modalities. In the following paper, we will first simply introduce features extraction encoders in Section 3.1, then illustrate detailedly the proposed MMT, LGFF, MMAL in Section 3.2, 3.3, and 3.4, respectively.

3.1. Encoders

We adopt two visual backbones for a video clip with its flow maps to extract the multi-level appearance features $\mathcal{A}^i \in \mathbb{R}^{T\times C^i_{\mathcal{A}}\times H^i\times W^i}$ and motion features $\mathcal{M}^i \in \mathbb{R}^{T\times C^i_{\mathcal{M}}\times H^i\times W^i}$, where $i \in [1,4]$ denotes the ith stage from the backbone. Following [5], we leverage the bidirectional transformer model BERT [12] as the linguistic encoder to extract linguistic features. Specifically, we first tokenize the text and add the [CLS] and [SEP] tokens at the beginning and end of the tokenized sequence. Then we feed the tokens sequence into BERT and obtain the token representations as linguistic feature $\mathcal{L} \in \mathbb{R}^{L\times C_{\mathcal{L}}}$.

We adopt a 1D convolution layer for the linguistic feature to reduce its channel dimension to C, obtaining $z_{\mathcal{L}} \in \mathbb{R}^{L\times C}$. For the high-level appearance feature $\mathcal{A}^4$ and motion feature $\mathcal{M}^4$, we first respectively concatenate an 8-dimensional coordinate feature $PC^4 \in \mathbb{R}^{8\times H^4\times W^4}$ with them like [49] to encode the spatial location information. Then, two ASPP modules [10] are adopted to unify their channel dimensions to C, respectively. Finally, two features are flattened and reshaped, resulting $z_{\mathcal{A}} \in \mathbb{R}^{T\times H^4W^4\times C}$ and $z_{\mathcal{M}} \in \mathbb{R}^{T\times H^4W^4\times C}$, respectively.

3.2. Multi-Modal Video Transformer

As discussed in Section 1, to explore the rich multi-modal interaction and leverage temporal information in different frames, we propose our Multi-Modal Video Transformer (MMVT). From Figure 2, each layer of our MMVT contains three components: cross-modal attention (CMA), temporal attention (TA), and MLP. The multi-layer perceptron block (MLP) is a common component in transformers *e.g.* [14, 46] and we do not talk about it here.

In our cross-modal attention module, we aim to promote the interaction between different modalities in a single frame. Based on this, we first concatenate high-level

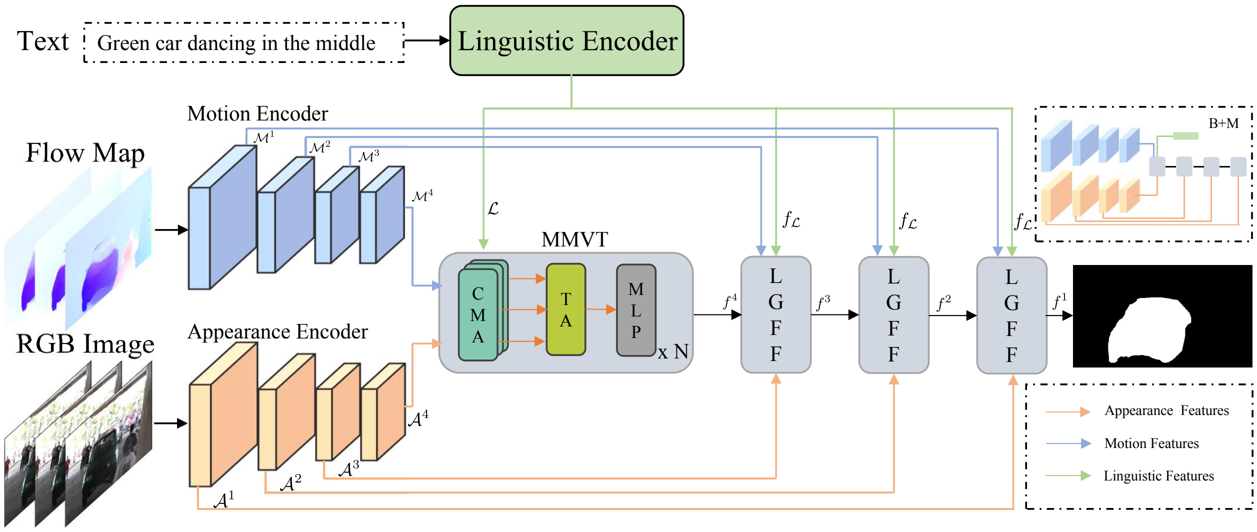

Figure 2. Overview of the proposed model. MMVT: Multi-modal video transformer. CMA: Cross-modal attention. TA: Temporal attention. LGFF: Language-guided feature fusion. "B+M" is the baseline model with motion information, details about which can be found in 4.4. Here, we do not show the multi-modal alignment loss for simplification.

features from three modalities and obtain feature $z \in \mathbb{R}^{T\times(2HW+L)\times C}$. Here, we omit the superscript of H and W for simplification. These can be formulated as:

$$z = Cat(z_\mathcal{A}, z_\mathcal{M}, z_\mathcal{L}). \tag{1}$$

We omit the broadcast operation along the temporal dimension for $z_\mathcal{L}$ here. Then, we pass it through a layer normalization (LN) [4] and we input z into the multi-head self attention (MSA) [46]. Note that, a residual connection is added here to improve robustness. Formally, this can be defined as:

$$z' = MSA(LN(z)) + z. \tag{2}$$

This process acts along the temporal dimension so that multi-model features in every frame can be well fused.

In our temporal attention module, the fused multi-modal features from different frames can interact with each other. First, we chunk z' into $z'_\mathcal{A} \in \mathbb{R}^{T\times HW\times C}$, $z'_\mathcal{M} \in \mathbb{R}^{T\times HW\times C}$, $z'_\mathcal{L} \in \mathbb{R}^{T\times L\times C}$. Here, $z'_\mathcal{A}$ can be considered as the appearance feature that has been enhanced by other modal features. To reduce the computational complexity, we only build the temporal relationships for $z'_\mathcal{A}$. Before feeding into MSA, we first flatten $z'_\mathcal{A}$ into $\mathbb{R}^{THW\times C}$, so that the information in temporal dimension can participate in the interaction. Then it can be formulated as follow:

$$z''_\mathcal{A} = MSA(LN(z'_\mathcal{A})) + z'_\mathcal{A} \tag{3}$$

After that, $z''_\mathcal{A}$ is reshaped back to $\mathbb{R}^{T\times HW\times C}$.

By doing the process above, the information contained in a frame can flow to other frames. After that, we concatenate the feature $z''_\mathcal{A}$, which has been enhanced by other frames, with $z'_\mathcal{M}$ and $z'_\mathcal{L}$, resulting in z''. Finally, we adopt the MLP to increase nonlinearity. These can be formulated as:

$$z'' = Cat(z''_\mathcal{A}, z'_\mathcal{M}, z'_\mathcal{L}), \tag{4}$$

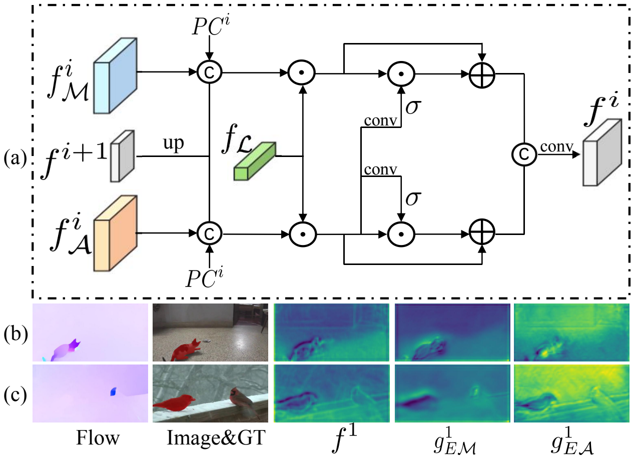

Figure 3. (a) Language-Guided Feature Fusion Module. "up": Upsample operation. PC^i: Coordinate feature for ith level. "C": Concatenation operation. $\odot$: Element-wise Multiplication. $\oplus$: Element-wise Addition. (b)(c) We visualize the feature map $g^1_{E\mathcal{M}}$, $g^1_{E\mathcal{A}}$ and f^1.

$$z''' = MLP(z'') + z''. \tag{5}$$

Since $z'_\mathcal{A}$ already contains information from other modalities, the multi-modal information can exchange and fuse between frames via the interaction of $z'_\mathcal{A}$ in the temporal attention module. Note that, these are all processes in one layer of MMVT. By stacking them for several layers, multi-modal features from different frames can be well fused and aggregated. Here, we set the number of layers to four by default.

3.3. Language-Guided Feature Fusion Module

Our language-guided feature fusion module (LGFF) aims at progressively fusing multi-modal features from different feature levels. As illustrated in Figure 3 (a), we first adopt two 1×1 convolution layers to reduce the channel number of appearance feature $f^i_{\mathcal{A}}$ and motion feature $f^i_{\mathcal{M}}$ to C. Then, each feature will be concatenated with the feature from the previous LGFF module f^{i+1} and the 8-dimensional coordinate feature, followed by a 3×3 convolution layer to fuse them. The feature f^{i+1} contains higher-level and semantically stronger information, while the coordinate feature can provide spatial location information.

Then, we need to emphasize the important region in the feature map with the guidance from linguistic features. Since [CLS] token in $\mathcal{L}$ has aggregated the representation of the whole sentence [12], we multiply it with two fused features, and obtain the enhanced appearance and motion feature, respectively. We can formulate this process as:

$$f^i_{E\mathcal{A}} = f_{\mathcal{L}} \odot Conv_3([PC^i, Up(f^{i+1}), f^i_{\mathcal{A}}]), \tag{6}$$

$$f^i_{E\mathcal{M}} = f_{\mathcal{L}} \odot Conv_3([PC^i, Up(f^{i+1}), f^i_{\mathcal{M}}]). \tag{7}$$

Here, $\odot$ denotes element-wise multiplication. $f_{\mathcal{L}}$ represents the [CLS] token in linguistic features $\mathcal{L}$. Through this process, the region related to the text in the feature will be selected and emphasized.

Since appearance usually contains more information than motion features, we adopt f^i_{EA} to generate two spatial-attention maps $att_{\mathcal{A}}$ and $att_{\mathcal{M}}$ through a 1×1 convolution layer followed by a sigmoid function to further emphasize the target region. Note that two convolution layers here do not share parameters. The residual connect is adopted here to avoid losing some meaningful information.

$$g^i_{E\mathcal{A}} = att_{\mathcal{A}} \odot f^i_{E\mathcal{A}} + f^i_{E\mathcal{A}}, \tag{8}$$

$$g^i_{E\mathcal{M}} = att_{\mathcal{M}} \odot f^i_{E\mathcal{M}} + f^i_{E\mathcal{M}}, \tag{9}$$

where $g^i_{E\mathcal{A}}$ and $g^i_{E\mathcal{M}}$ are two obtained features.

Finally, they are concatenated together and further fused with two 3×3 convolution layers with the ReLU function, resulting in f^i. We insert three LGFF modules into our network as the decoder, hence $i \in [1, 3]$. Note that we adopt $z'''_{\mathcal{A}}$ as the feature from the highest level f^4 in the first LGFF module since it has been enhanced by other modalities and temporal information from different frames in MMVT. From Figure 3 (b)(c), we can find that, no matter whether the flow map can highlight the target object or not, $g^i_{E\mathcal{M}}$ can distinguish the target object from other regions and incorporate well with $g^i_{E\mathcal{A}}$ to generate the final output feature f^1.

3.4. Multi-modal Alignment Loss

Although our model has achieved good performance with the two modules above, we notice that there may exist some semantic gap between the three modalities features since they are extracted from encoders that are pre-trained on different source data [20]. Based on this, we propose our multi-modal alignment loss so that three modalities features can be explicitly aligned.

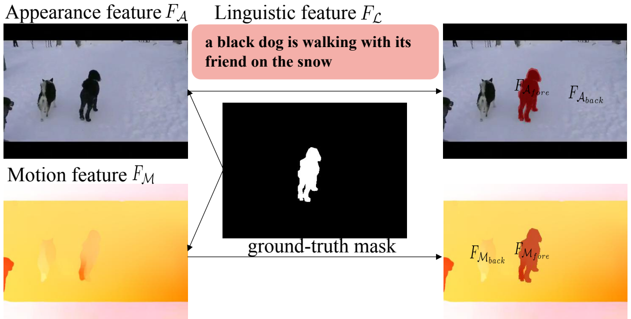
Appearance feature $F_{\mathcal{A}}$
Linguistic feature $F_{\mathcal{L}}$
a black dog is walking with its friend on the snow
$F_{\mathcal{A}_{fore}}$
$F_{\mathcal{A}_{back}}$
Motion feature $F_{\mathcal{M}}$
ground-truth mask
$F_{\mathcal{M}_{back}}$
$F_{\mathcal{M}_{fore}}$

Figure 4. We adopt the ground-truth mask to distinguish the features that belong to the foreground or the background in appearance features$f_{\mathcal{A}}$ and motion feature $f_{\mathcal{M}}$.

Specifically, we consider that the features belonging to the target object from appearance features are foreground features, and other features are background features. For motion features, it can also be categorized into the foreground and background features. Then, the feature alignment rules are defined as (1) the linguistic features should be close to foreground features from both appearance and motion features in embedding space, while far away from background features. (2) Appearance and motion features from the same category should be close to each other, meanwhile far away from other category features. Since the multi-modal alignment loss is defined per frame, we do not need to consider the temporal dimension here.

First, for each frame, we obtain a whole representation of appearance feature $F_{\mathcal{A}}$ by upsampling $f^2_{E\mathcal{A}}$, $f^3_{E\mathcal{A}}$ and concatenating them with $f^1_{E\mathcal{A}}$ together. $F_{\mathcal{M}}$ is obtained in the same way. Here we also adopt the [CLS] token $F_{\mathcal{L}}$ in $\mathcal{L}$ as the whole representation of the text. We employ three MLP functions to transform $F_{\mathcal{A}}$, $F_{\mathcal{M}}$, and $F_{\mathcal{L}}$ into the same embedding space with the same channel numbers c.

Now, we need to distinguish the feature belongs to the target object from other features in $F_{\mathcal{A}}$ and $F_{\mathcal{M}}$. This can be easily realized by leveraging the ground-truth mask. For example, in Figure 4, we can obtain foreground features $F_{\mathcal{A}_{fore}}$ and background features $F_{\mathcal{A}_{back}}$, where we know that $F_{\mathcal{A}_{fore}} \cup F_{\mathcal{A}_{back}} = F_{\mathcal{A}}$. We can obtain the alignment score $p_{\mathcal{AL}}$ between each element $f^i_{\mathcal{A}} \in F_{\mathcal{A}}$ and $F_{\mathcal{L}}$ by:

$$\hat{p}^i_{\mathcal{AL}} = \sigma(\tan(\frac{\pi}{2} sim(f^i_{\mathcal{A}}, F_{\mathcal{L}}))), \tag{10}$$

where sim represents the function to calculate the cosine similarity. If $f^i_{\mathcal{A}}$ is close to $F_{\mathcal{L}}$ in the embedding space, their cosine similarity will be close to 1 then the alignment score $\hat{p}^i_{\mathcal{AL}}$ will be close to 1, otherwise $\hat{p}^i_{\mathcal{AL}}$ will be close to 0. Based on this, we can define its label $p^i_{\mathcal{AL}}$ as: if $f^i_{\mathcal{A}} \in F_{\mathcal{A}_{fore}}$, $p^i_{\mathcal{AL}} = 1$ otherwise $p^i_{\mathcal{AL}} = 0$. Now the alignment loss $L_{\mathcal{AL}}$ between $F_{\mathcal{A}}$ and $F_{\mathcal{L}}$ can be defined as:

$$L_{\mathcal{AL}} = -\sum p^i_{\mathcal{AL}} \log \hat{p}^i_{\mathcal{AL}} + (1 - p^i_{\mathcal{AL}}) \log(1 - \hat{p}^i_{\mathcal{AL}}). \tag{11}$$

The alignment loss $L_{\mathcal{ML}}$ between $F_{\mathcal{M}}$ and $F_{\mathcal{L}}$ can also be defined in the same way.

For appearance features $f^i_{\mathcal{A}} \in F_{\mathcal{A}}$ and motion features $f^i_{\mathcal{M}} \in F_{\mathcal{M}}$, we can also align them together. The alignment score can be defined as:

$$\hat{p}^{i,j}_{\mathcal{AM}} = \sigma(\tan(\frac{\pi}{2} sim(f^i_{\mathcal{A}}, f^j_{\mathcal{M}}))). \tag{12}$$

When $f^i_{\mathcal{A}}$ and $F^j_{\mathcal{M}}$ belongs to the foreground or background at the same time, its label $p^{i,j}_{\mathcal{AM}} = 1$, otherwise $p^{i,j}_{\mathcal{AM}} = 0$. The alignment loss $L_{\mathcal{AM}}$ can be defined as:

$$L_{\mathcal{AM}} = -\sum p^{i,j}_{\mathcal{AM}} \log \hat{p}^{i,j}_{\mathcal{AM}} + (1 - p^{i,j}_{\mathcal{AM}}) \log(1 - \hat{p}^{i,j}_{\mathcal{AM}}). \tag{13}$$

Finally, we define the multi-modal alignment loss as:

$$L^{align} = L_{\mathcal{AL}} + L_{\mathcal{ML}} + L_{\mathcal{AM}}. \tag{14}$$

4. Experiments

4.1. Datasets and Evaluation Metrics

Following prior works, we conduct experiments on two popular text-based video segmentation datasets including **A2D Sentences** [16] and **J-HMDB** Sentences [16]. These two datasets are extended by Gavrilyuk *et al.* [16] via providing a referring language for each target object in Actor-Action Dataset (A2D) [50] and J-HMDB [23].

A2D Sentences contains 3,782 videos, which are split into 3,036 and 746 videos for training and testing, respectively. There are 3 to 5 frames with pixel-level annotations in each video for training and evaluating segmentation performance. Besides, there are 6,655 sentences to describe the actors and their actions in each video. **J-HMDB Sentences** contains 928 videos from 21 action classes with corresponding 928 sentences. All frames in it are annotated at the pixel level. Previous methods usually evaluate their generalization ability on this dataset.

Intersection-over-union (IoU) is the ratio of intersection area over union area between the ground-truth mask and prediction. Following prior works, we adopt **Overall IoU** and **Mean IoU** to evaluate the performance. The former treats ground-truth masks and predictions on the testing dataset as a whole, resulting in favor of larger objects, while the latter is the averaged IoU overall test samples. We also adopt **P@X** to measure the percentage of samples whose IoU are higher than the threshold X, where X $\in$ $[0.5, 0.6, 0.7, 0.8, 0.9]$. The mean average precision (**mAP**) over 0.5:0.95 is also adopted.

4.2. Implementation Details

Following [30], we adopt the ResNet-101 and ResNet-34 [17] as the appearance and motion encoders to extract appearance and motion features. The stride of four stages in two encoders is set as 2, 2, 2, and 1, respectively. RAFT [45] is employed to generate optical flow maps. We adopt an Adam [26] optimizer with the learning rate 2×10^{-5} to train the whole network. The batch size is set to 8, and each batch contains a video clip with three frames. We set the maximum training step to 30,000, and the learning rate is divided by ten at 25,000 and 28,000, respectively. Following the settings in prior works, all frames are resized and padded to 320×320. The maximum length of each input sentence is 20. All experiments are conducted on 2 NVIDIA Tesla V100 GPUs.

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

A2D Sentences We employ the training and testing set of A2D Sentences to train and evaluate our model, respectively. As shown in Table 1, our method surpass over state-of-the-art method CSTM by 0.8% , 2.6% , 4.2% on Precision @0.6, @0.7 and @0.8, respectively. This means that, when the metric is more stricter, our model can surpass previous methods by a larger margin. It is noteworthy that, our method achieve 13.0 % on the most challenging metric Precision @0.9, which means that our method can generate particularly accurate segmentation masks. The mAP and Overall IoU can also be further improved by 2.0% and 1.1%, respectively. We also notice that our model is lower than CSTM [21] by 0.9% on Precision@0.5, which is because our model tend to generate more accurate and confident results, while some not accurate results from CSTM [21] can still considered to be True, since the threshold in Precision@0.5 is low. Furthermore, since CSTM generates masks on the feature map with original size while our model predicts on the feature map with $1/4$ original size, they may perform better on small objects. Hence the performance of our method is slightly lower than its on IoU Mean, which treats small objects equally.

J-HMDB Sentences Like previous works, we adopt the J-HMDB Sentences to verify the generalization ability of our method. Following [21, 49], we employ the model that achieves the best performance on A2D Sentences to directly evaluate on the test set of J-HMDB Sentences, which is split by [16]. As illustrated in Table 2, our method outperform all previous methods on all metrics. It is easy to find that our model can surpass other methods by a large margin, especially when the metric is strict *e.g.* Precision@0.6, @0.7 and @0.8. This phenomenon is similar to that in A2D Sentences, which means that our model shows more robust performance with the help of well-fused multi-modal information. Note that, like other methods, our approach can not achieve good results on Precision@0.9 (lower than 1 %), since all methods are not trained or finetuned on J-HMDB Sentences.

4.4. Ablation Study

Following previous works, we conduct ablation experiments on A2D Sentences to thoroughly analyze and verify

Table 1. Comparison with state-of-the-art methods on A2D Sentences testing set. † denotes adopting additional optical flow input.

Methods	Venue	Precision					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.* [18]	ECCV2016	34.8	23.6	13.3	3.3	0.1	13.2	47.4	35.0
Li *et al.* [33]	CVPR2017	38.7	29.0	17.5	6.6	0.1	16.3	51.5	35.4
Gavrilyuk *et al.* [16]	CVPR2018	47.5	34.7	21.1	8.0	0.2	19.8	53.6	42.1
Gavrilyuk *et al.* † [16]	CVPR2018	50.0	37.6	23.1	9.4	0.4	21.5	55.1	42.6
ACGA [49]	ICCV2019	55.7	45.9	31.9	16.0	2.0	27.4	60.1	49.0
VT-Capsule [38]	CVPR2020	52.6	45.0	34.5	20.7	3.6	30.3	56.8	46.0
CMDY [48]	AAAI2020	60.7	52.5	40.5	23.5	4.5	33.3	62.3	53.1
PRPE [39]	IJCAI2020	63.4	57.9	48.3	32.2	8.3	38.8	66.1	52.9
CSTM [21]	CVPR2021	**65.4**	58.9	49.7	33.3	9.1	39.9	66.2	**56.1**
Our †	–	64.5	**59.7**	**52.3**	**37.5**	**13.0**	**41.9**	**67.3**	55.8

Table 2. Comparison with state-of-the-art methods on J-HMDB Sentences testing set. All methods adopt the best model trained on A2D Sentences to directly eval on J-HMDB Sentences without finetuning. † denotes adopting additional optical flow input.

Methods	Venue	Precision					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.* [18]	ECCV2016	63.3	35.0	8.5	0.2	0.0	17.8	54.6	52.8
Li *et al.* [33]	CVPR2017	57.8	33.5	10.3	0.6	0.0	17.3	52.9	49.1
Gavrilyuk *et al.* [16]	CVPR2018	69.9	46.0	17.3	1.4	0.0	23.3	54.1	54.2
ACGA [49]	ICCV2019	75.6	56.4	28.7	3.4	0.0	28.9	57.6	58.4
VT-Capsule [38]	CVPR2020	67.7	51.3	28.3	5.1	0.0	26.1	53.5	55.0
CMDY [48]	AAAI2020	74.2	58.7	31.6	4.7	0.0	30.1	55.4	57.6
PRPE [39]	IJCAI2020	69.1	57.2	31.9	6.0	**0.1**	29.4	-	-
CSTM [21]	CVPR2021	78.3	63.9	37.8	7.6	0.0	33.5	59.8	60.4
Our †	–	**79.9**	**71.4**	**49.0**	**12.6**	**0.1**	**38.6**	**61.9**	**61.3**

Table 3. Quantitative results of each component in our model. Appearance: with appearance feature; Motion: with motion feature; MMVT: Multi-Modal Video Transformer; LGFF: Language-Guided Feature Fusion Module; Align: Multi-modal Alignment Loss.

Name	Settings					Precision					mAP	IoU	
	Appearance	Motion	MMVT	LGFF	Align	P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	0.5:0.95	Overall	Mean
B	✓					55.1	50.7	44.2	31.7	9.5	35.3	61.9	48.2
B+M	✓	✓				56.8	51.9	45.0	32.3	10.0	36.3	63.5	49.5
B+T	✓		✓			59.2	54.1	46.1	32.2	9.8	37.2	64.4	51.3
B+M+T	✓	✓	✓			62.0	56.8	48.7	34.3	10.5	39.2	64.8	53.6
B+T+L	✓		✓	✓		62.0	57.4	49.6	36.2	11.6	40.1	65.5	54.0
B+M+T+L	✓	✓	✓	✓		63.1	58.5	51.2	37.1	12.6	41.1	66.8	54.8
B+M+T+L+A	✓	✓	✓	✓	✓	**64.5**	**59.7**	**52.3**	**37.5**	**13.0**	**41.9**	**67.3**	**55.8**

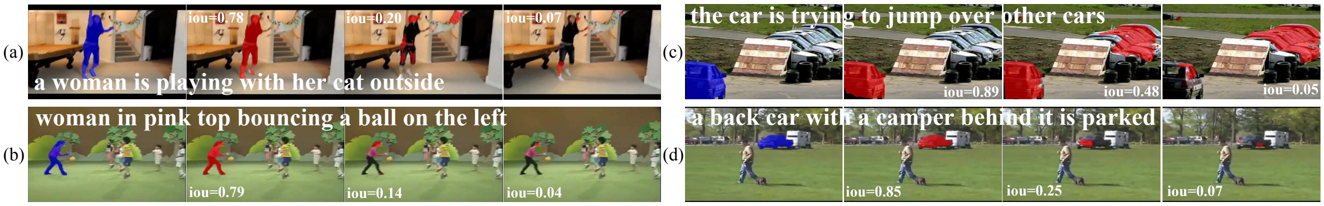
(a)
a woman is playing with her cat outside
iou=0.78
iou=0.20
iou=0.07
(b)
woman in pink top bouncing a ball on the left
iou=0.79
iou=0.14
iou=0.04
(c)
the car is trying to jump over other cars
iou=0.89
iou=0.48
iou=0.05
(d)
a back car with a camper behind it is parked
iou=0.85
iou=0.25
iou=0.07

Figure 5. Qualitative results comparison. From left to right in (a), (b), (c) and (d): ground-truth, ”B+M+T+L+A”, ”B+M”, and ”B”.

the effectiveness of the proposed method.

Effectiveness of Each Component. We first verify each component in our model in Table 3. “B+M” is the baseline model shown in Figure 2, which only adopt concatenation and convolutional layers to fuse multi-modal. In addition, only appearance features are fused in the decoder in ”B+M”. ”B” is the same as ”B+M” except without motion branch. By comparing them, we can find that introducing the explicit motion information from optical flow maps can effectively improve the performance. To verify the effectiveness of multi-modal interaction between frames, we replace the concatenation operation in ”B+M” with the proposed MMVT and obtain ”B+M+T”. We find that the performance is significantly improved, especially in mAP and Mean IoU, improved by 2.9% and 4.1%, respectively. This benefits from the powerful capacity of fusing multi-modal features between frames in MMVT. Then, we replace all simple concatenation operations in every level of the decoder in ”B+M+T” with the proposed LGFF and obtain ”B+M+T+L”. This demonstrates a remarkable improvement on all metrics, especially in rigorous metrics Precision@0.7, @0.8, and @0.9, which are improved by 2.5%, 2.7%, and 2.1%, respectively. This means the decoder with our LGFF can progressively fuse multi-modal features from different levels and gradually recover the resolution of the feature map, leading to more accurate segmentation masks. Finally, we add the proposed multi-modal alignment loss into ”B+M+T+L+A” and the results demonstrate that explicitly aligning multi-modal features can obtain better performance. To further verify the generalization of proposed components, we gradually add our MMVT and LGFF to ”B”, resulting in”B+T” and ”B+T+L”. The results show that only fusing appearance and linguistic features with our MMVT and LGFF can also improve the performance.

Table 4. Comparison of using different MMVT setting.

Name	Setting		mAP	IoU	
	CMA	TA	0.5:0.95	Overall	Mean
B+M			36.3	63.5	49.5
+CMA	✓		36.8	64.1	50.6
+CAT+TA		✓	38.6	64.8	53.2
B+M+T	✓	✓	39.2	64.8	53.6

Table 5. Comparison of using different decoder settings.

Name	Setting		mAP	IoU	
	CAT	LGFF	0.5:0.95	Overall	Mean
CAT	✓		37.6	63.5	51.6
LGFF		✓	41.1	66.8	54.8

Hence, our MMVT and LGFF can work well in different settings instead of only handling the setting with motion.

MMVT Settings. There are two attention modules in our MMVT, named cross model attention (CMA) and temporal attention (TA), respectively. Here, we conduct experiments to verify their effectiveness in Table 4. We remove all TA modules in MMVT from ”B+M+T”, which is denoted as ”+CMA”. We can see that the performance drop significantly, which verifies the usefulness of fusing information from different frames in TA. When compared with ”B+M”, ”+CAM” achieves better results, which shows its capacity of fusing multi-modal features. Furthermore, we try to remove all CMA modules in TA and only adopt a concatenation with a convolutional layer to fuse multi-modal features before the MMVT, which is denoted as ”+CAT+TA”. This results in worse performance than ”B+M+T”, which verify that it is useful and necessary to combine CMA and TA in our MMVT.

Decoder Settings. In Table 5, we conduct experiments to explore the effectiveness of our LGFF. We adopt the concatenation operation followed by a convolution layer to fuse appearance, motion, linguistic features as well as features from higher level as the baseline, which is denoted as ”CAT”. We find that such a simple fusion strategy degrades the performance obviously, compared with our model ”LGFF”. This means that it is necessary to design the LGFF to effectively fuse multi-modal features from different levels.

Effectiveness of Multi-modal Alignment Loss We conduct experiments to verify the effectiveness of our multi-modal alignment loss. ”l2am” denotes adopting $L_{\mathcal{AL}}$ and $L_{\mathcal{ML}}$ to align linguistic features with appearance and motion features, while ”a2m” represents employing $L_{\mathcal{AM}}$ to align appearance and motion features. We also try to add two traditional binary cross-entropy losses to ”B+M+T+L” for appearance and motion branch, respectively, which is denoted as ”+bce”. From Table 6, we can find that ”+bce” can not bring obvious improvement to the performance. When we add ”l2am” and ”a2m” to ”B+M+T+L”, the performance is improved gradually, which verifies the effectiveness of ”l2am” and ”a2m”.

Table 6. Comparison of using different Multi-modal Alignment Loss settings.

Name	Setting		mAP	IoU	
	l2am	a2m	0.5:0.95	Overall	Mean
B+M+T+L			41.1	66.8	54.8
+bce			41.2	66.3	54.8
+l2am	✓		41.2	67.1	55.2
B+M+T+L+A	✓	✓	41.9	67.3	55.8

4.5. Qualitative Results Comparison

We visualize some representative samples generated from ”B”, ”B+M”, and ”B+M+T+L+A” in Figure 5. In some complex scenes like Figure 5 (a) and (b), there are multiple objects moving, leading to unsatisfying segmentation results from ”B+M”, which simply adopts concatenation to fuse multi-modal features. From Figure 5 (c), we can find that, when the motion information is adopted, although the model ”B+M” can find the car, it still misclassifies some pixels from other cars as foreground, while ”B+M+T+L+A” can generate more accurate mask. These examples show that our method can well incorporate and fuse appearance, motion and linguistic features together to locate the target object and generate more accurate masks. Figure 5 (d) demonstrates that our model can still accurately segment the target object without motion.

5. Conclusion

In this paper, we propose a method to fuse and align multi-modal features for text-based video segmentation. First, we introduce the explicit motion information from optical flow maps to incorporate with appearance and linguistic features. Then, we design the MMVT to fuse multi-modal features between frames. Furthermore, we propose the LGFF module to progressively fuse multi-modal features from different feature levels. Finally, the multi-modal alignment loss is adopted to explicitly align multi-modal features to reduce the semantic gap between them. Extensive experiments verify the effectiveness of each component in our method and demonstrate that our method can significantly outperform state-of-the-art methods on two popular datasets.

Acknowledgments: We thank Google TFRC for supporting us to get access to the Cloud TPUs. We thank CSCS (Swiss National Supercomputing Centre) for supporting us to get access to the Piz Daint supercomputer. We thank TACC (Texas Advanced Computing Center) for supporting us to get access to the Longhorn supercomputer and the Frontera supercomputer. We thank LuxProvide (Luxembourg national supercomputer HPC organization) for supporting us to get access to the MeluXina supercomputer. This research is supported by the National Research Foundation Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-008), NRF Centre for Advanced Robotics Technology Innovation (CARTIN) Project, and NUS Faculty Research Committee Grant (WBS: A-0009440-00-00).

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. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 3

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 2, 3

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[5] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. Refvos: A closer look at referring expressions for video object segmentation. *arXiv preprint arXiv:2010.00263*, 2020. 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 2, 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 3

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1

[9] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. *arXiv preprint arXiv:2108.09322*, 2021. 1, 2, 3

[10] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 5

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[15] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, pages 3664–3673, 2017. 1

[16] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, pages 5958–5966, 2018. 1, 2, 3, 6, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[18] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124. Springer, 2016. 1, 2, 7

[19] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021. 2, 3

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2, 5

[21] Tianrui Hui, Shaofei Huang, Si Liu, Zihan Ding, Guanbin Li, Wenguan Wang, Jizhong Han, and Fei Wang. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In *CVPR*, pages 4187–4196, 2021. 1, 3, 6, 7

[22] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017. 1

[23] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, pages 3192–3199, 2013. 6

[24] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. *arXiv preprint arXiv:2108.03151*, 2021. 1

[25] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 1

[28] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 3

[29] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 3

[30] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, pages 7274–7283, 2019. 1, 6

[31] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 3

[32] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *CVPR*, pages 4641–4650, 2016. 1

[33] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *CVPR*, pages 6495–6503, 2017. 7

[34] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring

image segmentation. In *ICCV*, pages 1271–1280, 2017. 2

[35] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, pages 4722–4732, 2021. 3

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 3

[37] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 2

[38] Bruce McIntosh, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. Visual-textual capsule routing for text-based video segmentation. In *CVPR*, pages 9942–9951, 2020. 2, 7

[39] Ke Ning, Lingxi Xie, Fei Wu, and Qi Tian. Polar relative positional encoding for video-language segmentation. In *IJCAI*, volume 9, page 10, 2020. 3, 7

[40] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, pages 3202–3212, 2015. 1

[41] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *arXiv preprint arXiv:2109.09920*, 2021. 3

[42] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019. 3

[43] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, 2021. 3

[44] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *CVPR*, pages 4631–4640, 2016. 3

[45] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419. Springer, 2020. 6

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 2, 3, 4

[47] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 3

[48] Hao Wang, Cheng Deng, Fan Ma, and Yi Yang. Context modulated dynamic networks for actor and action video segmentation with language queries. In *AAAI*, volume 34, pages 12152–12159, 2020. 1, 2, 7

[49] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *ICCV*, pages 3939–3948, 2019. 1, 2, 3, 6, 7

[50] 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. 2, 6

[51] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACMMM*, pages 1645–1653, 2017. 1

[52] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 1

[53] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, pages 7177–7188, 2021. 1

[54] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *CVPR*, pages 11266–11275, 2021. 2

[55] Yiding Yang, Jiayan Qiu, Mingli Song, Dacheng Tao, and Xinchao Wang. Distilling knowledge from graph convolutional networks. In *CVPR*, 2020. 3

[56] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019. 2

[57] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *arXiv:2111.11418*, 2021. 3

[58] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, pages 471–487, 2018. 3

[59] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 3

[60] H. Zhang, A. Sun, W. Jing, L. Zhen, J. T. Zhou, and R. S. M. Goh. Natural language video localization: A revisit in span-based question answering framework. *TPAMI*, 2021. 1

[61] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *CVPR*, pages 16826–16835, 2021. 1

[62] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1

[63] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, volume 34, pages 13066–13073, 2020. 1

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[65] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, pages 3537–3545, 2019. 3