Neurocomputing 552 (2023) 126523

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

MPP-net: Multi-perspective perception network for dense video captioning

Yiwei Wei [a,*], Shaozu Yuan [b,1], Meng Chen [b], Xin Shen [b], Longbiao Wang [a], Lei Shen [b], Zhiling Yan [b]

[a] College of Intelligence and Computing, Tianjin University, Tianjin, China
[b] JD AI Research, Beijing, China

ARTICLE INFO

Communicated by Zidong Wang

Keywords:
Dense video captioning
Deformable transformer
Object detection
Paragraph video captioning.

ABSTRACT

Applying deformable transformer for dense video captioning has achieved great success recently. However, deformable transformer only explores local-perspective perception by attending to a small set of key sampling points, which will make the decoder short-sighted and generate semantically incoherent and contradictory dense captions for a long video. In this paper, we propose a novel Multi-Perspective Perception Network to improve this problem. We first introduce a hierarchical temporal-spatial summary method to generate global-perspective summary context for each decoder layer and avoid redundant information. Then our new designed multi-perspective attention encourages the model to selectively incorporate the multi-perspective perception feature. Finally, we propose a novel multi-perspective generator to perform both multi-perspective feature fusion and caption generation. Experiments show that our proposed model outperforms previously published methods and achieves a competitive performance on ActivityNet Captions and YouCook2. The design of our model also shows the universality of other visual tasks that we obtain comparable results by applying our model for Object Detection and Paragraph Video Captioning.

1. Introduction

Video captioning [1] is an important branch of visual captioning, which needs to produce a single description by giving a short video. Nevertheless, generating only one sentence may be insufficient to describe a long and untrimmed video in the real world. To tackle this issue, dense video captioning (DVC) [2–6] is proposed to automatically localize and describe multiple events in the video. Thus, DVC can be divided into two subtasks. The first subtask aims at localizing all possible events in a video, while the second focuses on representing these proposals to generate language descriptions.

Most previous approaches [7,2,4,8] model temporal localization and caption generation as two fully separate tasks. However, they rigidly follow the rule to first localize events in videos by employing temporal action proposal models [9] and then describe each event with a sentence, where the caption generation performance will be affected by event localization. To overcome this problem, PDVC [10] jointly models these two tasks with an end-to-end framework to further improve dense video captioning performance. It leverages deformable transformer [11] to model long sequential video, for vanilla transformer suffers from slow convergence and high memory usage.

Despite showing advantages, PDVC has limitations. For deformable transformer [11], the backbone of PDVC, adopts local-perspective attention [12] to accelerate the inference time. This local perspective perception may bring semantic information decay in deeper decoder layers and make the decoder short-sighted, which will restrain the capability of the model. For another, it is essential to capture global perspective perception to generate semantically coherent dense captions for a long video. In Fig. 1 (a), the deficiency of global perspective perception leads to the generated captions showing worse coherence and video-semantic consistency. Previous work [13] tries to use self-attention to enhance global-perspective perception for videos. Nevertheless, directly applying self-attention to a long video suffers from redundant connections and over-smoothing [14], which makes the model even harder to optimize. Therefore, the key point is how to provide global perspective perception with lower redundancy.

Based on above analyses, we propose a novel architecture Multi-Perspective Perception Network (MPP-Net) for dense video captioning,

* Corresponding author.
E-mail addresses: wywy@tju.edu.cn (Y. Wei), yuanshaozu@jd.com (S. Yuan), chenmeng20@jd.com (M. Chen), u6498962@anu.edu.au (X. Shen), longbiao_wang@tju.edu.cn (L. Wang), shenlei20@jd.com (L. Shen), yanzhling@mail2.sysu.edu.cn (Z. Yan).
[1] Equal contribution.

https://doi.org/10.1016/j.neucom.2023.126523
Received 2 January 2023; Received in revised form 20 June 2023; Accepted 30 June 2023
Available online 7 July 2023

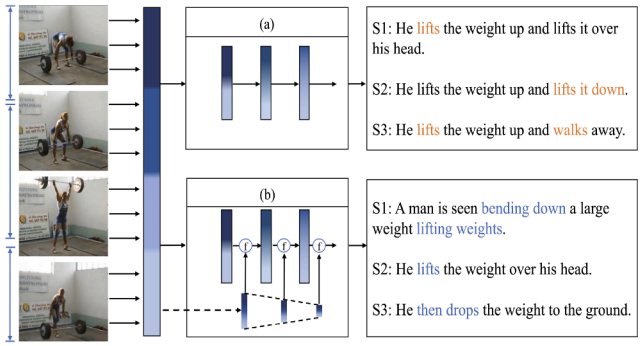


Fig. 1. In example (a), it provides an incoherent and inconsistant caption result by merely utilizing local-perspective perception. Example (b) shows hierarchical summary contexts can further improve generated captions in terms of coherence and video semantic consistency.

which is equipped with three modules including a Hierarchical Temporal Spatial Summary method (HTSS), a new designed multi-perspective attention (MPA), and a multi-perspective captioning generator (MSG). 1) To avoid irrelevant information and generate video summary context for global perception, we first propose hierarchical temporal-spatial summary method (HTSS) to extract the most distinguishing pixel in a temporal-spatial representation space. Inspired by previous work [15,16] that shows deeper layers of transformer trend to focus more on long distance context, HTSS will gradually magnify the temporal duration size in the deeper decoding layer. 2) A multi-perspective perception is introduced to integrate a multi-perspective attention (MPA) layer. MPA takes advantage of both self-attention and deformable attention and weighs multi-perspective contributions via a learned gate. 3) We also design a multi-perspective captioning generator (MPCG) with a two-layer LSTM structure, with the first LSTM guiding visual attention combined with multi-perspective information and the second LSTM generating captions. We evaluate our proposed model on ActivityNet Captions and YouCook2, and the experimental results show that our model outperforms all previous models significantly in all metrics and achieves new state-of-the-art performance. Moreover, we conduct both quantitative and qualitative analyses to demonstrate that MPP-Net can significantly enhance the quality and coherence of the dense video descriptions. The architecture of Multi-Perspective Perception also shows universality in other visual tasks, and we extend our model on another two tasks, Object Detection and Paragraph Video Captioning, improving the performance on both two tasks.

1. We propose a novel architecture, MPP-Net, for dense video captioning, which includes three novel designed modules: a hierarchical temporal spatial summary method, a multi-perspective attention, and a multi-perspective captioning generator.
2. Experimental results show our model establishes a new competitive performance on ActivityNet Captions and YouCook2 in terms of all metrics. Extensive experiments are conducted to verify the contribution of different components.
3. The architecture of MPP-Net shows universality in other visual tasks, where it also improves the performance on Object Detection and Paragraph Video Captioning.

2. Related Work

2.1. Video Captioning

Inspired by the success of neural models in translation systems [17], numerous effective models [18–21], have been developed for video captioning. The underlying concept of this approach is to train two Recurrent Neural Networks (RNNs) in an encoder-decoder structure. Specifically, an encoder inputs a set of video features, aggregates its hidden state, and then transfers it to a decoder for generating a caption. In order to enhance the performance of the captioning model, several methods have been proposed, such as shared memory between visual and textual domains [22,23], spatial and temporal attention [24], reinforcement learning [25], semantic tags [26], and other modalities [27,28]. Later efforts such as CLIP4Caption [29] and SWINBERT [30] boost the performance by employing pretrained modules to enhance multimodal representation. However, these approaches aim at generating only one sentence for an input video, which is insufficient for describing long and untrimmed videos. To tackle this issue, dense video captioning is proposed to automatically describe multiple events in the video.

2.2. Two-stage Dense Video Captioning

Dense video captioning aims at combining both event localization and event captioning. Due to the lack of spatial annotation, earlier approaches show inferior performance, especially in localizing the events. [2] introduced the dense captioning task and a benchmark dataset: ActivityNet Captions. They also proposed a two-stage pipeline that first detects events with a variant method of action detection proposal, then describes the event with an attention-based LSTM. After that, Bi-SST [31] applied a bidirectional event proposal module which exploits both past and future context for proposal prediction. [32] directly dealt with the dense video captioning task by designing different levels of hierarchical LSTMs to retrieve intra-event and inter-event descriptor. [7] proposed an event-centric hierarichal network to employ scene-level, event-level, and frame-level contents simultaneously. In addition, some researches [8,6] have also been proposed to detect other

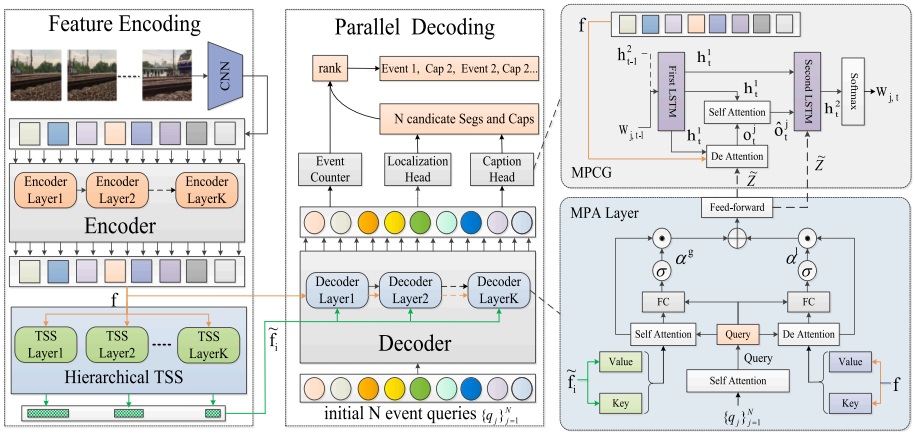


Fig. 3. The overall architecture of our model. In the encoder network, we propose the hierarchical temporal spatial summary(HTSS) to guide the model to produce distinguishing global-perspective visual features. In the encoder network, the multi-perspective attention (MPA) layer is proposed to explore multi-perspective perception in the video. We further designed a multi-perspective captioning generator (MPCG) to generate captions incorporating multi-perspective information dynamically.

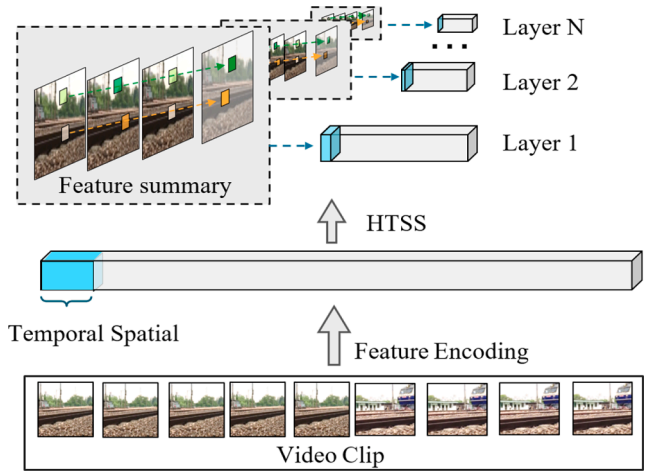


Fig. 2. HTSS extracts the most distinguishing regions in a temporal spatial representation to obtain video summary context.

modalities (e.g., audio and speech) to improve dense video captioning. Generally, all these methods take the two-stage "localize-then-describe" pipeline to boost dense video captioning.

2.3. End-to-End Dense Video Captioning

[4] proposed a novel framework that jointly trained the event localization and event captioning in an end-to-end manner. Moreover, massive approaches have investigated the application of the Transformer [33] model to the dense video captioning task. In particular, [13] proposed an end-to-end transformer model to exploit a differentiable masking network to ensure the consistency between proposal and captioning module during training. However, all these methods focus on generating a large number of proposal-caption pairs, which directly leads to redundancy or inconsistency and ultimately reduces the readability of generated captions. To alleviate this problem, [34] introduced an event selection network to adaptively select a sequence of event proposals, thus improving the readability and coherence of the generated descriptions. Though promising results were achieved, this method is not an end-to-end model, and the subtasks need to be trained separately. Most recent work [10] combines the advantages of [13,34] and boosts the performance by formulating the dense caption generation as an end-to-end task. It accelerates the inference time via the deformable transformer and models the DVC tasks with three parallel subtasks. However, the deformable attention only explores local-perspective perception, and that leads to incoherent and inconsistent captions.

3. Methodology

The overview of our MPP-Net is shown in Fig. 3, which can be conceptually divided into a feature encoding network and a decoding network. The feature encoding network encodes the input video into visual representations, where the frame-level features are extracted by pretrained action recognition network [35,36]. Then the encoding feature and summary context obtained from HTSS will be fed into multi-perspective attention module to explore multi-perspective perception for the video. On the top of the network, three parallel subtasks are applied to generate the dense captions. We will introduce each component in detail in the following subsections. We use the interpolation operation to re-scale the feature temporal dimension to a fixed number T. Then we add M temporal convolutional layers to get feature sequences across multiple resolutions. Finally, the multi-scale features with their positional embeddings are fed into the deformable transformer encoder. The output features are denoted as $f \in \mathbb{R}^{L\times C}, L = \{T + T/2^1 + \ldots + T/2^M\}$.

3.1. Hierarchical Temporal Spatial Summary Method

In order to solve the issue that self-attention suffers from redundant connections, over-smoothing, and relation ambiguity, we present an intuitive but effective method named Hierarchical Temporal Spatial Summary (HTSS) to reduce redundant information. As shown in Fig. 2, we hypothesize that the visual features are very similar in a temporal-spatial representation space. The encoding feature can be denoted as $X \in R^{T\times A}$, temporal duration size as $t \in \{1, 2, \ldots T\}$, where A represents the spatial size and T is total length of the video. Thus, the encoding feature can be formulated as $X_1, X_2, X_t \ldots X_T$, where $X_t \in R^{t\times A}$. To avoid bring extra computing burden, HTSS simply retains the most distinguishing with max representing value to obtain the temporal summary s_t. For one layer, the output of Hierarchical Temporal Spatial Summary (TSS) is $[s_1, s_2, s_t \ldots s_T]$. It summaries the video from a more macroscopic perspective compared to deformable attention.

$$\left\{\widetilde{f}_i\right\}_{i=1}^{K} = \left\{TSS_1\left(f\right), TSS_2\left(f\right), TSS_j\left(f\right), \ldots . TSS_K\left(f\right)\right\} \tag{1}$$

where TSS_j denotes Temporal Spatial Summary (TSS) with different temporal duration size for j^{th} decoding layer and $\{\widetilde{f}_i\}_{i=1}^{K}$ denotes final video summary context of HTSS. Note that the number of the pathways K is equal to the number of the decoder layers.

3.2. Multi-Perspective Perception Decoder

This decoder aims to obtain multi-perspective perception according to encoding feature, video summary context and event queries.

Multi-perspective Attention (MPA). MPA aims to integrate the output of the encoder $f \in \mathbb{R}^{L\times C}$ and the global-perspective summary context $\{\widetilde{f}_i\}_{i=1}^{K}$ conditioned on the event queries. It takes advantage of both deformable attention and self-attention. Concretely, self-attention is exploited to explore global-perspective perception with summary context generated by HTSS, while deformable attention is exploited to learn local-perspective perception. Besides, we apply a learned gate to weigh the multi-perspective contributions at each decoding layer.

In order to show this more clearly, we illustrate MPA in formula. Given initial event query vectors $\{q_j\}_{j=1}^{N} \in \mathbb{R}^{N\times C}$ and encoding feature f, we first employ a multi-scale deformable attention(MSDAtt) to explore local-perspective perception $O^l \in R^{N\times C}$ denoted by:

$$O^l = \mathrm{MSDAtt}\left(\{q_j\}_{j=1}^{N}, \{p_j\}_{j=1}^{N}, f\right) \tag{2}$$

where $\{p_j\}_{j=1}^{N}$ denotes the reference points of $\{q_j\}_{j=1}^{N}$. To avoid bringing semantic decay for video summary context $\widetilde{f}_i$, we then apply parallel multi-head self attention(MHAtt) layer to model $\{q_j\}_{j=1}^{N}$ and $\widetilde{f}_i$ to capture global-perspective perception $O^g \in R^{N\times C}$, denoted by:

$$O^g = \mathrm{MHAtt}\left(\{q_j\}_{j=1}^{N}, \widetilde{f}_i, \widetilde{f}_i\right) \tag{3}$$

These multi-perspective perception vectors are then selectively integrated through a learning gate, which can be defined as:

$$M = \alpha^g \otimes O^g + \alpha^l \otimes O^l \tag{4}$$

where $\alpha^g \in \mathbb{R}^{N\times C}$ and $\alpha^l \in \mathbb{R}^{N\times C}$ are the weight matrices, which modulate the contribution of O^g and O^l, respectively. These two weight matrices are calculated as:

$$\alpha^g = \sigma(W^g([Y, O^g]) + b^g) \tag{5}$$

$$\alpha^l = \sigma\left(W^l\left(\left[Y, O^l\right] + b^l\right)\right) \tag{6}$$

where $[.,.]$ represents concatenation, σ is the sigmoid activation. $\{W^g,$

$W^l\}$ are $2C \times C$ weight matrices, and $\{b^g, b^l\}$ are learnable bias vectors.

Architecture of decoding layers. At each decoding layer, a self attention layer is first utilized to perform interaction between different event queries $\{q_j\}_{j=1}^N$. Then, we apply our multi-perspective attention (MPA) to capture multi-perspective perception. Besides, the decoder layer contains a position-wise feed-forward layer, and all components are normalized with AddNorm(AN) operations. The output Z of the current decoder layer can be written as:

$$\{Z_j\}_{j=1}^N = \mathrm{AN}\left(\mathrm{MPA}\left(\mathrm{AN}\left(\mathrm{MHAtt}\left(\{q_j\}_{j=1}^N\right)\right)\right)\right) \tag{7}$$

$$\left\{\widetilde{Z}_j\right\}_{j=1}^N = \mathrm{AN}\left(\mathscr{F}\left(\{Z_j\}_{j=1}^N\right)\right) \tag{8}$$

where $\{q_j\}_{j=1}^N$ are the vectors of input sequence, and $\mathscr{F}$ denotes the position-wise feed-forward layer. Finally, we stack multiple decoding layers together and generate the event queries for the following parallel subtasks.

3.3. Parallel Subtasks

To generate the final dense captions, there are three parallel subtasks on the top of the network, including a designed multi-perspective captioning generator(MPCG), a localization detector, and an event counter.

Multi-perspective captioning generator (MPCG) To further encourage the model to incorporate multi-perspective information and generate captions dynamically, we devise a two-layer LSTM structure for both multi-perspective feature fusion and caption generation. Let denote h_t^1 and h_t^2 as the hidden states for the first LSTM and the second LSTM respectively, where t is current time step. The input for the first LSTM at each time step consists of the previous output h_{t-1}^2 of the second LSTM and previously generated word $w_{j,t-1}$, given by:

$$h_t^1 = \mathrm{LSTM}\left(\left[h_{t-1}^2, w_{j,t-1}\right], h_{t-1}^1\right) \tag{9}$$

where $w_{j,t-1}$ is an embedded matrix cross the vocabulary Σ.

Given the output of the first LSTM h_t^1 and the output of the MPA decoder $\{\widetilde{Z}_j\}_{j=1}^N$, at each time step t, we use a multi-scale deformable attention layer to generate sampling points from each f^l, thus generating the local-perspective relations at each time step. To model the global-perspective relations, we employ a multi-head self-attention layer between the output of MSDAtt and h_t^1, which can be given by:

$$\widehat{O}_t^j = \mathrm{MHAtt}\left(h_t^1, \mathrm{MSDAtt}\left(\left[h_t^1, \left\{\widetilde{Z}_j\right\}_{j=1}^N\right], p_j^h, f^l\right)\right) \tag{10}$$

where p_j^h represents the reference points and $\widehat{O}_t^j$ denotes the multi-perspective attended features.

The input to the second LSTM consists of the attended features $\widehat{O}_t^j$, concatenated with the output of the attention LSTM h_t^1 and the event query $\widetilde{q}_j$, given by:

$$h_t^2 = \mathrm{LSTM}\left(\left[h_t^1, \widehat{O}_t^j, \left\{\widetilde{Z}_j\right\}_{j=1}^N\right], h_{t-1}^2\right) \tag{11}$$

At each time step t, the output of the language LSTM h_t^2 is leveraged for the prediction of next word $w_{j,t}$ via an FC layer with softmax activation.

Localization Detector. It aims to predict the segment location and the foreground confidence conditioned on each event query induced by the proposed decoder. After processed by multi-layer perceptron, a set of detected events $\left\{t_j^s, t_j^e, c_j^{loc}\right\}_{j=1}^N$ are obtained, where t_j^s and t_j^e are the start time and the end time of j-th event. c_j^{loc} is the localization confidence of the event query $\widetilde{q}_j$.

Event counter. It is a module that transforms the event queries $\{\widetilde{q}_j\}_{j=1}^N$ into a fix-size vector r_{len} through a max-pooling layer and a FC layer with softmax activation, where each value refers to the possibility of a specific number. During the inference stage, the outputs are obtained by selecting the top N_{set} events with accurate boundaries and good captions from N event queries, where N_{set} is obtained by argmax operation. The confidence of each event query is calculated by:

$$c_j = c_j^{loc} + \left(\frac{\mu}{M_j^{\gamma}} \sum_{t=1}^{M_j} \log\left(c_{jt}^{cap}\right)\right) \tag{12}$$

where c_{jt}^{cap} is the probability of the generated word, γ is a modulation factor for rectifying the influence of caption length and μ is the balanced vector.

3.4. Training loss

Following previous works [37,10], we also use the Hungarian algorithm to find the bipartite matching results based on N trained events with their locations and captions. Specially, it contains two parts: the matching cost and the prediction loss. For the matching cost, we define it as follows:

$$C = \alpha_{giou} L_{giou} + \alpha_{cls} L_{cls} \tag{13}$$

where L_{giou} represents the generalized IOU [38] between predicted temporal segments and ground-truth segments. L_{cls} represents the focal loss [39] between the predicted classification score and the ground-truth label. For the prediction loss, we calculate the weighted sum of generalized IOU loss, classification loss, countering loss, and caption loss:

$$L = \beta_{giou} L_{giou} + \beta_{cls} L_{cls} + \beta_{ec} L_{ec} + \beta_{cap} L_{cap} \tag{14}$$

where caption loss L_{cap} and countering loss L_{ec} are both the cross-entropy loss between the predicted result and the ground truth.

4. Experiments

In this section, we first introduce the experimental setup. Then, we present the comparative results and conduct ablation study with more model variants. Finally, we show the advantage of our model by a case study and discuss the limitation of our approach.

4.1. Datasets and Evaluation

All our experiments were conducted on two popular dense video captioning benchmarks ActivityNet Captions [2] and YouCooks [40]. ActivityNet Captions contains 20k long untrimmed videos of various human activities, which follows a standard split including 10009 training videos, 4925 validation videos, and 5044 testing videos. Each video, on average, lasts 120 s and contains 3.65 temporally localized captions. YouCook2 consists of 2000 untrimmed videos of cooking procedures, each has 7.7 annotated segments with associated sentences. We use the official split with 1333/457/210 videos for training, validation, and testing.

We evaluate our model with two kinds of evaluation metrics. To evaluate the captioning performance, we take the most widely used automatic metrics like BLEU [41], METEOR [42], ROUGE-L [43], CIDEr [44], and SODA_c [45] that is recently proposed for evaluating the coherence of caption. And the performance of event proposal is verified with precision, recall and F1-score (i.e., the harmonic mean of precision and recall), by calculating average IOU score with threshold of 0.3, 0.5, 0.7, 0.9. And we report more detailed BLEU scores in ablation study to fully measure the performance of different modules.

Table 1
Event localization results on the validation set of ActivityNet Captions.

	Recall	Precision	F1
MFT [48]	24.31	51.41	33.01
SDVC [34]	55.58	57.57	56.56
PDVC [10]	55.42	58.07	56.71
MPP-Net	**55.58**	**58.37**	**56.86**

4.2. Implementation Details

We take the same feature extraction method with previous works [13,10] for fair comparison. We employ the widely used C3D [35] action recognition model as the backbone to obtain features on ActivityNet Captions and adapt TSN [46] model to extract features for YouCook2. The MPP-Net encoder is composed of two stacked encoder layers with 4 levels multi-scale deformable attention, and the decoder consists of three stacked decoder layers with the proposed multi-perspective attention. Note that the number of event queries is 10 for ActivityNet Captions and 100 for YouCook2. Besides, the hidden size is set to 512 in MSDAtt/MHAtt layers and 2048 in feed-forward layers. For the proposed MPCG, hidden dimension is 512 for both the attention LSTM and the language LSTM. During training stage, we use an initial learning rate of $5e^{-5}$. All models are trained with Adam optimizer [47] and mini-batch size of 1. To search the optimal parameters for HTSS module, we test different scales of HTSS with the temporal duration size of $4, 8$ and 16 respectively on ActivityNet Captions. For the event counter, we choose the maximum count as $10/20$ for ActivityNet Captions/YouCook2. In Eqn. 12, γ and μ are set to 2 and 1.0 respectively. The cost ratios in bipartite matching are $\alpha_{\text{giou}} : \alpha_{\text{cls}} = 2 : 1$ and the loss ratios are $\beta_{\text{giou}} : \beta_{\text{cls}} : \beta_{\text{ec}} : \beta_{\text{cap}} = 2 : 1 : 1 : 1$.

4.3. Main Results

In this section, we report main results of the proposed model MPP-Net and the SOTA methods on both localization task and dense captioning task. And we evaluate the caption performance on above two public datasets.

Localization performance. Table 1 shows the results on event localization task. Here, the comparison methods can be divided into two categories. The first follows a two-stage scheme (i.e. MFT [48] and SDVC [34]), which generates event proposals by a "localize-then-describe" paradigm. The second follows an end-to-end scheme (i.e. PDVC [10] and MPP-Net), which outputs the event proposals in a parallel manner. It can be seen that MPP-Net not only exhibits better performances than the two-stage methods MFT and SDVC, but are also slightly better than the end-to-end method PDVC in terms of precision, recall and F1-score. Similar to the findings in [10], we conjecture that the parallel decoding can implicitly capture the location-aware features from caption supervision, helping the optimization of the event localization.

Dense captioning performance. Table 2 summarizes the experimental results of various models on ActivityNet Captions. For fair comparison, we respectively show the performance by utilizing ground-truth proposals and predicted proposals. Under the challenging setting that the model needs to predict proposals by itself, MPP-Net outperforms all the baselines by a large margin especially in terms of CIDEr and SODA_c. Specially, MPP-Net obtains CIDEr score of 29.76, which is the most important metric to measure the performance of video caption. This is to-date the best performance and outperforms previous SOTA model PDVC by **4**. Besides, the SODA_c score of MPP-Net is 5.61, which is 3.5 higher than the previous best performance. Under the easier setting with GT proposals, the trend is similar to above observations that MPP-Net also surpasses all the previous best methods. The improvements generally demonstrate the advantage of exploiting multi-perspective interactions via multi-perspective attention, which facilitates both the quality and coherence of generating descriptions. To further verify generalization of our model, we report the dense captioning performance on YouCook2 in Table 3. The substantial improvements of MPP-Net on the YouCook2 continues to verify the importance to integrate both local and global dependences.

4.4. Ablation Study

In this section, we conduct extensive ablation studies to verify the contribution of each component in our MPP-Net, including HTSS, MPCG, and MPA.

Table 3
Dense captioning on YouCook2. We report BLEU(B4), METEOR(M) and CIDEr (C) to align with previous method.

Methods	Feature	Predicted proposals		
		B4	M	C
MT [13]	TSN	0.30	3.18	6.10
ECHR [7]	TSN	-	3.82	-
PDVC [10]	TSN	0.80	4.74	22.71
MPP-Net	TSN	**0.99**	**4.81**	**24.11**

Table 4
The performance with varying temporal duration size of HTSS. "no summary" means removing the HTSS module. Notice that we replace the MPCG module with ordinary LSTM to improve comparability and experiment confidence.

#d	temporal duration size	Predicted proposals					
		B1	B2	B3	B4	M	C
2	(no summary)	15.13	7.57	3.74	1.77	7.37	27.53
2	(2/4)	15.21	7.65	3.72	1.83	7.42	27.39
2	(2/8)	15.20	7.63	3.75	1.81	7.47	27.51
2	(2/16)	15.21	7.70	3.71	1.83	7.45	27.73
2	(4/8)	15.21	7.73	3.83	**1.90**	7.39	27.67
2	(4/16)	15.30	7.74	**3.84**	1.86	7.29	28.48
2	(8/16)	15.08	7.60	3.77	1.88	7.34	27.17
3	(no summary)	15.04	7.55	3.70	1.79	7.51	26.62
3	(2/4/8)	15.24	7.69	3.80	1.88	7.48	28.34
3	(2/4/16)	15.26	7.66	3.74	1.83	7.49	27.82
3	(2/8/16)	15.24	7.54	3.60	1.76	7.41	27.02
3	(4/8/16)	**15.50**	**7.78**	3.81	1.89	**7.56**	**28.55**

Table 2
The results of the dense video captioning task on the ActivityNet Captions validation sets in terms of BLEU4(B4), METEOR(M), CIDEr(C) and SODA_c.

Method	Feature	GT proposals			Predicted proposals			
		B4	M	C	B4	M	C	SODA_c
DCE [2]	C3D	1.60	8.88	25.12	0.17	5.69	12.43	-
TDA-CG [31]	C3D	-	9.69	-	1.31	5.86	7.99	-
DVC [4]	C3D	1.62	10.33	25.24	0.73	6.93	12.61	-
SDVC [34]	C3D	-	-	-	-	6.92	-	-
Efficient [49]	C3D	-	-	-	1.35	6.21	13.82	-
ECHR [7]	C3D	1.96	10.58	39.73	1.29	7.19	14.71	3.22
PDVC [10]	C3D	2.64	10.54	47.26	1.65	7.50	25.87	5.26
MPP-Net	C3D	**2.71**	10.59	50.07	**2.04**	**7.61**	**29.76**	**5.61**

Table 5
Ablation study. Here, '- MPCG' '- L', '- G' and '- gate' represent removing MPCG, local-perspective perception, global-perspective perception and the learned gate respectively.

	Predicted proposals					
	B1	B2	B3	B4	M	C
MPP-Net	**15.61**	**8.01**	**4.03**	**2.04**	**7.61**	**29.76**
- MPCG	15.50	7.78	3.81	1.89	7.56	28.55
- MPCG - gate	15.31	7.72	3.82	1.90	7.49	27.95
- MPCG - gate - L	15.32	7.69	3.70	1.75	7.39	28.47
- MPCG - gate - G	15.04	7.55	3.70	1.79	7.51	26.62

Table 6
The performance of MPP-Net when changing the number of event queries.

#q	Predicted proposals							
	Recall	Precision	B1	B2	B3	B4	M	C
5	55.86	56.70	14.72	7.10	3.23	6.98	7.07	25.81
10	55.65	58.33	15.50	7.78	3.81	1.89	7.56	28.55
20	53.21	59.03	15.02	7.59	3.78	1.91	7.62	25.67
30	50.42	57.94	15.10	7.76	3.81	1.92	7.70	26.11
50	52.82	57.91	15.02	7.62	3.80	1.94	7.53	26.07
100	51.31	58.40	15.12	7.61	3.73	1.85	7.66	26.17

Performance of HTSS. In order to study the effectiveness of HTSS and explore the optimal setting of HTSS for following experiments, we verify the performance of HTSS with varying temporal duration size in Table 4. Overall, HTSS shows its superiority compared to no summary model and the best result is achieved with temporal duration size of $(4/8/16)$. It reveals increasing a larger initial temporal duration size is more conducive to the captioning performance, which is consistent with previous finding in [16] that higher decoder layer tends to focus more on longer distance of semantic contexts relevant to the task. According to above results, we choose the relatively optimal setting with 3 decoding layers and the temporal duration size of $(4/8/16)$ in our experiments.

Performance of MPCG. To evaluate the effectiveness of our multi-perspective captioning generator (MPCG), two model variants are compared: the first is MPP-Net without MPCG and the second is MPP-Net equipped with MPCG. Notably, "without MPCG" represents the model only using one LSTM layer to implement both visual attention and caption generation. From Table 5, we observe that MPCG can exhibit better captioning performance across all the compared metrics, demonstrating the advantage of enhancing the interaction between visual content and natural sentence via our two-layer LSTM structure for dense video captioning.

Performance of MPA. We also implement several ablated models by removing the learned gate, local-perspective perception, and global-perspective perception separately to examine the contribution of each components in Table 5. It's observed that the performance of all model variants degrades compared to MPP-Net, which proves the effectiveness of integrating multi-perspective perception. And it also indicates global-perspective perception play a critical role in this task because when we remove the global-perspective attention and video summary context (global-perspective perception), the caption performance degrades drastically. Besides, the result also shows superiority of the learned gate on incorporating the advantages of both multi-perspective features.

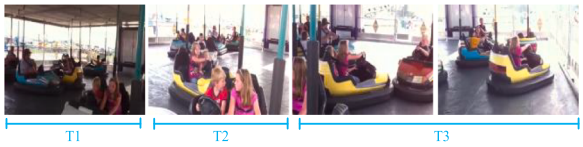
Impact of number of events. Furthermore, we verify the influence of using different initial event queries in multi-perspective perception decoder. From Table 6, we observe that the number of initial event queries has a significant impact on the results. To balance the locating and captioning performance, we choose the number of initial event queries to 10 in our experiments. This setting is also consistent with the event distribution of the ActivityNet Captions.

Table 7
Paragraph captioning on the ActivityNet Captions dataset [2].

Method	Features	GT proposals		
		B4	M	C
Trans-XL [50]	visual + flow	10.39	15.09	21.67
VTrans [13]	visual + flow	9.75	15.64	22.16
MART [51]	visual + flow	10.33	15.68	23.42
PDVC [10]	visual + flow	11.80	15.93	27.27
MPP-Net	visual + flow	**12.75**	**16.01**	**29.35**

Table 8
Results of object detection. AP denotes the detection accuracy.

Model	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR-DeFormer	62.6	47.7	26.4	47.1	58.0
DETR-MPP-Net	**63.2**	**48.1**	**26.7**	**47.5**	**58.4**



PDVC: ["a group of people are seen riding around a bumper cars", "the people continue to ride around the cars", "the kids continue to get around the car"]

MPP-Net: ["a group of people are riding in bumper cars", "they are driving in bumper cars", "they are driving down and bumping into each other"]

Ground-Truth: ["People drive bumper cars around a bumper car rink in an amusement park", "A yellow bumper car hits another bumper car", "A red bumper car hits the outer wall and stalls out"]

Fig. 5. Examples of dense video captioning results coupled with ground-truth temporal locations. It is noted that all three compared models use the same temporal locations.

4.5. Universality Discussion

In order to evaluate whether the MPP-Net can be applied to other visual tasks, we first replace the subtask of MPP-Net for paragraph video captioning [52,51]. Note that paragraph video captioning is a simplified version of dense video captioning, which needs to generate a coherent paragraph without predicting the temporal location in the video. For fair comparison, we follow the schedule in [10] and remove the localization head to optimize the whole architecture with cross-entropy loss. Table 7 shows the performance comparison between the state-of-the-art methods and our proposed MPP-Net. Similar to the observations in dense video captioning, MPP-Net outperforms several competitive baselines including Trans-XL [50], VTrans [13], MART [51], and PDVC [10] significantly. The improvements demonstrate the advantage of exploiting both multi-perspective interactions via our multi-perspective attention operator for paragraph video captioning.

For computer vision task, we conduct the experiments by applying our model for Objective Detection, where the top caption decoder layers are replaced by detection layer. For fair comparison, all the experimental settings are the same as the original DETR-D [11]. In Table 8, our MPP-Net improves the performance in all metrics compared with Deformable Transformer, proving that the introduction of global-perspective perception helps the model to understand the image content.

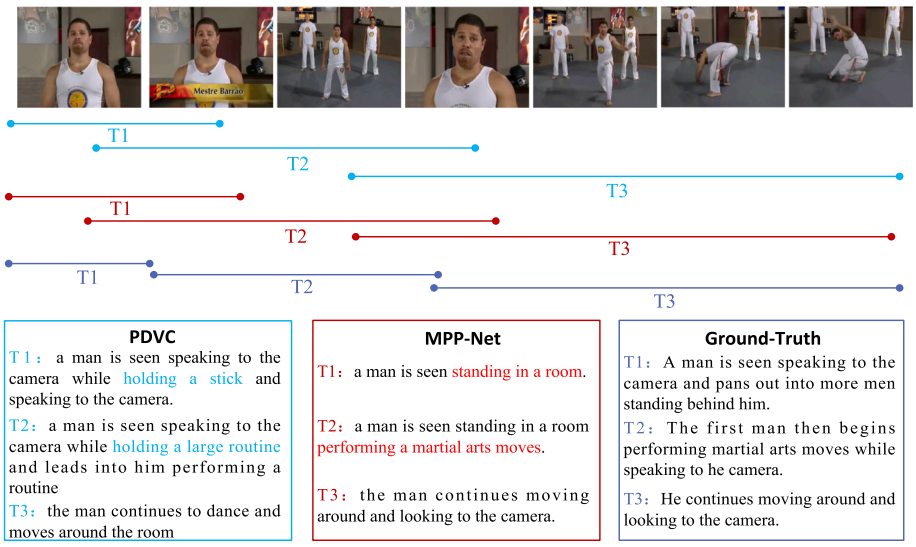


Fig. 4. Case study for an example video from the ActivityNet Captions validation set. The generated captions of our MPP-Net and PVDC [10] are compared with the human-annotated ground truth. Notably, we highlight the inconsistent captions in blue and improving captions in red.

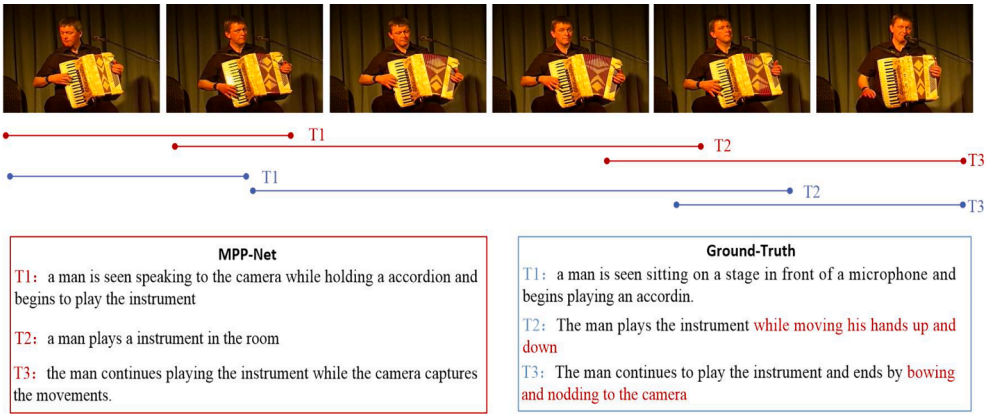


Fig. 6. Another case study for an example video. The generated captions of our MPP-Net are compared with the human-annotated ground truth. The red-marked sentences indicate fine-grained descriptions in ground truth.

4.6. Case Study

In this section, we illustrate some examples to compare the description quality of PDVC [10] and MPP-Net. Two different settings are considered: the first setting is to generate video captions with ground-truth temporal locations, as shown in Fig. 5; the second one is to generate captions with predicted temporal locations, as demonstrated in Fig. 4.

The text in different colors (blue vs. red) shows the difference between the two models. It's observed that, our MPP-Net can produce more accurate and descriptive sentences than PDVC. For example, in Fig. 5, MPP-Net understood the video and depicted the clip of T3 as "*bumping into each other*" while PDVC only generated vague expressions such as "*get round the car*". Similarly, in Fig. 4, PDVC made some obvious mistakes during generation, e.g., it mentioned "*holding a stick*", which is incorrect with the video content. Besides, PDVC can only roughly describe in T2 as "*performing a routine*", while MPP-Net can accurately describe "*performing a martial arts moves*". We argue that MPP-Net has the above advantages because it can figure out the connections among different events and also know how they are connected. In MPP-Net, the parallel decoder uses the HTSS method to obtain representative features in each event and uses the self-attention mechanism to measure how well they are related. By this way, the content of other events can effectively correct the mistakes that occurred when generating captions for the current event.

Although the proposed MPP-Net performs well in the reported evaluation metrics, it still has some limitations. As can be seen in Fig. 6, MPP-Net is incapable of MPP-Net is incapable of describing the fine-grained events, such as "moving his hands up and down", "bowing and nodding to the camera". The phenomenon also existed in other dense video captioning methods. The reason for this phenomenon is that when using a temporal convolutional network to extract video features, it can only capture the coarse-grained features of video segments and not the fine-grained ones. As a result, the model is unable to perceive detailed information about events. In future work, we plan to incorporate more prior knowledge, such as fine-grained knowledge of attributes and actions, to assist the model in generating descriptions.

Table 9
Results of human evaluation, with respect to metrics of relevance(Rel), coherence(Coh) and conciseness(Con).

Metrics	Rel	Coh	Con
PDVC [10]	37%	41%	39%
MPP-Net	**63%**	**59%**	**61%**

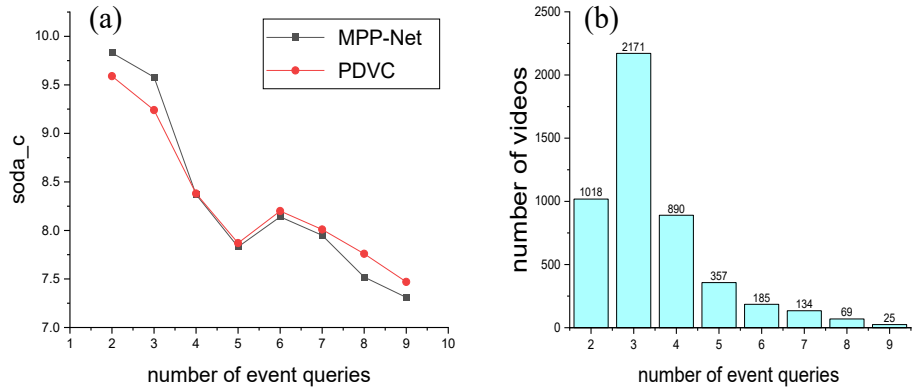


Fig. 7. Figure (a) shows the curve of soda_c score as the video event number increases for PDVC and MPP-Net, respectively. Figure (b) shows the distribution of event numbers in the dataset.

4.7. Human Evaluation.

Considering the most reliable way to evaluate the quality of the description is with human judges, here we conduct further human evaluation in Table 9 to assess the performance of generated captions. Here, we only compare the MPP-Net with the most competitive baseline PDVC [10] for simplicity. Considering that this task is considerably challenging, the human evaluation does not include ground truth. We first apply MPP-Net and PDVC to generate captions for 100 cases randomly sampled from the test set. Then we shuffle the order of predicted captions and hide the model name. A total of 10 college students who are proficient in English were invited to finish the human evaluation. Following previous works [51,53], three metrics are adopted, including **relevance** to verify how related the generated caption is to the content of the video, **coherence** to evaluate the logic, fluency, and readability of the generated caption; and **conciseness** to measure whether the captions are less verbose and repetitive.

4.8. Discussion of Limitation

In our experiments, we found an interesting phenomenon: that the number of events has a great impact on the performance. In Fig. 7 (a), we illustrate the curve of soda_c scores for PDVC and MPP-Net as the event number increases in the dataset. Generally, it shows that the soda_c scores of both models decrease sharply as the number of events increases, for videos containing too many events will make this task more challenging. Compared to PDVC, MPP-Net achieves better performance when the number of events is less than 4, which shows the advantage of global-perspective perception. However, to generate captions for the video with more events, the global-perspective perception shows a negative impact on our model. We guess it is hard to give a summary context for a complicated video, because some cases are usually disordered. Fig. 7(b) shows the distribution of the number of videos conditioned on the different numbers of events. It can be seen that most of the videos in ActivityNet Captions have 2 to 4 events, which explains MPP-Net's better performance than PDVC in our experiments. In the future, we will continue to study how to effectively capture global-perspective perception in complex videos to tackle the above limitations.

5. Conclusion

In this paper, we propose a novel Multi-Perspective Perception Network to explore multi-perspective perception. It includes a hierarchical temporal-spatial summary method that generates global-perspective summary context for each decoder layer while avoiding redundant information; a newly designed multi-perspective attention method that incorporates the summary context and local perspective feature selectively; and a novel multi-perspective caption generator that performs both feature fusion and caption generation. Experiments show that MPP-Net outperforms previously published methods and achieves competitive performance on ActivityNet Captions and YouCook2. It also shows universality and improves the performance in Object Detection and Paragraph Video Captioning. We hope our model can further facilitate related research.

CRediT authorship contribution statement

Yiwei Wei: Conceptualization, Methodology, Software, Investigation, Validation, Writing - original draft, Visualization. **Shaozu Yuan:** Conceptualization, Methodology, Software, Investigation, Validation, Writing - original draft, Visualization. **Meng Chen:** Resources, Writing - review & editing, Formal analysis, Supervision. **Xin Shen:** Resources, Writing - review & editing, Formal analysis, Supervision. **Longbiao Wang:** Writing - review & editing. **Lei Shen:** Writing - review & editing. **Zhiling Yan:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported by National Natural Science Foundation of China(No.62176182) and the Fundamental Research Funds for the Central Universities (22CX04043A).

References

[1] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence-video to text, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4534–4542.
[2] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, J. Carlos Niebles, Dense-captioning events in videos, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 706–715.
[3] Z. Shen, J. Li, Z. Su, M. Li, Y. Chen, Y.-G. Jiang, X. Xue, Weakly supervised dense video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1916–1924.
[4] Y. Li, T. Yao, Y. Pan, H. Chao, T. Mei, Jointly localizing and describing events for dense video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7492–7500.
[5] Z. Zhang, D. Xu, W. Ouyang, C. Tan, Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization, IEEE Trans. Circuits Syst. Video Technol. 30 (9) (2020) 3130–3139.

[6] V. Iashin, E. Rahtu, Multi-modal dense video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 958–959.
[7] T. Wang, H. Zheng, M. Yu, Q. Tian, H. Hu, Event-centric hierarchical representation for dense video captioning, IEEE Transactions on Circuits and Systems for Video Technology 31 (5) (2020) 1890–1900.
[8] V. Iashin, E. Rahtu, A better use of audio-visual cues: Dense video captioning with bi-modal transformer, in: The 31st British Machine Vision Virtual Conference, British Machine Vision Association, BMVA, 2020.
[9] S. Buch, V. Escorcia, C. Shen, B. Ghanem, J. Carlos Niebles, Sst: Single-stream temporal action proposals, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 2911–2920.
[10] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, P. Luo, End-to-end dense video captioning with parallel decoding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6847–6857.
[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159.
[12] Z. Yao, J. Ai, B. Li, C. Zhang, Efficient detr: improving end-to-end object detector with dense prior, arXiv preprint arXiv:2104.01318.
[13] L. Zhou, Y. Zhou, J.J. Corso, R. Socher, C. Xiong, End-to-end dense video captioning with masked transformer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8739–8748.
[14] J. Tan, J. Tang, L. Wang, G. Wu, Relaxed transformer decoders for direct action proposal generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13526–13535.
[15] B. Chen, P. Li, B. Li, C. Li, L. Bai, C. Lin, M. Sun, J. Yan, W. Ouyang, Psvit: Better vision transformer via token pooling and attention sharing, arXiv preprint arXiv: 2108.03428.
[16] K. Clark, U. Khandelwal, O. Levy, C.D. Manning, What does bert look at? an analysis of bert's attention, arXiv preprint arXiv:1906.04341.
[17] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Advances in neural information processing systems 27.
[18] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4507–4515.
[19] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, Sequence to sequence-video to text, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4534–4542.
[20] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, K. Saenko, Translating videos to natural language using deep recurrent neural networks, arXiv preprint arXiv:1412.4729.
[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.
[22] J. Wang, W. Wang, Y. Huang, L. Wang, T. Tan, M3: Multimodal memory modelling for video captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7512–7520.
[23] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, Y.-W. Tai, Memory-attended recurrent network for video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8347–8356.
[24] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, Stat: Spatial-temporal attention mechanism for video captioning, IEEE transactions on multimedia 22 (1) (2019) 229–241.
[25] X. Wang, W. Chen, J. Wu, Y.-F. Wang, W.Y. Wang, Video captioning via hierarchical reinforcement learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4213–4222.
[26] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, L. Deng, Semantic compositional networks for visual captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5630–5639.
[27] J. Xu, T. Yao, Y. Zhang, T. Mei, Learning multimodal attention lstm networks for video captioning, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 537–545.
[28] C. Hori, T. Hori, G. Wichern, J. Wang, T.-Y. Lee, A. Cherian, T.K. Marks, Multimodal attention for fusion of audio and spatiotemporal features for video description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2528–2531.
[29] M. Tang, Z. Wang, Z. Liu, F. Rao, D. Li, X. Li, Clip4caption: Clip for video caption, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4858–4862.
[30] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, L. Wang, Swinbert: End-to-end transformers with sparse attention for video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17949–17958.
[31] J. Wang, W. Jiang, L. Ma, W. Liu, Y. Xu, Bidirectional attentive fusion with context gating for dense video captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7190–7198.
[32] D. Yang, C. Yuan, Hierarchical context encoding for events captioning in videos, in: IEEE International Conference on Image Processing (ICIP), IEEE, in: 2018 25th, 2018, pp. 1288–1292.
[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30.
[34] J. Mun, L. Yang, Z. Ren, N. Xu, B. Han, Streamlined dense video captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6588–6597.
[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
[36] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks for action recognition in videos, IEEE transactions on pattern analysis and machine intelligence 41 (11) (2018) 2740–2755.
[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.
[38] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: A metric and a loss for bounding box regression, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 658–666.
[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, in: Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[40] L. Zhou, C. Xu, J.J. Corso, Towards automatic learning of procedures from web instructional videos, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7590–7598.
[41] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[42] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the ninth workshop on statistical machine translation, 2014, pp. 376–380.
[43] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, Text summarization branches out (2004) 74–81.
[44] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
[45] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, M. Nagata, Soda: Story oriented dense video captioning evaluation framework, European Conference on Computer Vision, Springer (2020) 517–531.
[46] Q. Wang, A.B. Chan, Describing like humans: on diversity in image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4195–4203.
[47] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.
[48] Y. Xiong, B. Dai, D. Lin, Move forward and tell: A progressive generator of video descriptions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 468–483.
[49] A.R.M. Suin, An efficient framework for dense video captioning, Proceedings of the AAAI Conference on Artificial Intelligence 34 (7) (2020) 12039–12046.
[50] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le, R. Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, arXiv preprint arXiv: 1901.02860.
[51] J. Lei, L. Wang, Y. Shen, D. Yu, T. Berg, M. Bansal, Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2603–2614.
[52] J.S. Park, M. Rohrbach, T. Darrell, A. Rohrbach, Adversarial inference for multi-sentence video description, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6598–6608.
[53] Y. Xiong, B. Dai, D. Lin, Move forward and tell: A progressive generator of video descriptions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 468–483.

Yiwei Wei is a lecturer in the department of computer science, China University of Petroleum (Bei Jing) at karamay. He is currently pursuing his PhD degree now studying in Tianjin University, Tianjin, China. His current research interests include multimodal fusion, image/video caption and multi-modal sentiment analysis.

Shaozu Yuan received the Master degree in the college of computer and communication engineering, China University of Petroleum(East China), Now, he works in JD Conversational AI, Beijing, China. His current research includes multimodal understanding, NLP and computer version.

Meng Chen is Director, JD Conversational AI, Beijing, China. His research includes NLP, speech recognition and multimodal understanding. He currently serving as the program committee member for several top academic conferences. His research interests include Virtual Reality, Computer Vision, Deep Learning, Data Mining, and Pattern Recognition.

Xin Shen received the Bachelor degree in the College of Engineering, Computing and Cybernetics from the Australian National Univerisity, Canberra, Australia, in 2022. He is currently working toward the Ph.D. degree in the School of Information Technology & Electrical Engineering, University of Queensland. His research interests include natural language processing and multimodal machine learning.

Longbiao Wang received his Dr. Eng. degree from Toyohashi University of Technology, Japan, in 2008. He was an assistant professor in the faculty of engineering at Shizuoka University, Japan from April 2008 to September 2012. From October 2012 to August 2016 he was an associate professor at Nagaoka University of Technology, Japan. Currently he is a Professor at the Tianjin University, China. His research interests include robust speech recognition and speaker recognition.

Lei Shen received the Ph.D. degree in the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her research interests include natural language processing, text generation, and dialogue systems.

Zhiling Yan received her Bachelor degree of Science in Statistics from Sun Yat-sen University. Currently, she is a graduate student in Nanyang Technological University. Her research interests include natural language processing and multimodal machine learning.