mathematics

MDPI

Article

Parallel Dense Video Caption Generation with Multi-Modal Features

Xuefei Huang [1], Ka-Hou Chan [1,2], Wei Ke [1,2,*] and Hao Sheng [1,3,4]

[1] Faculty of Applied Sciences, Macao Polytechnic University, Macau 999078, China; xuefei.huang@mpu.edu.mo (X.H.); chankahou@mpu.edu.mo (K.-H.C.); shenghao@buaa.edu.cn (H.S.)
[2] Engineering Research Centre of Applied Technology on Machine Translation and Artificial Intelligence of Ministry of Education, Macao Polytechnic University, Macau 999078, China
[3] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[4] Zhongfa Aviation Institute of Beihang University, 166 Shuanghongqiao Street, Pingyao Town, Yuhang District, Hangzhou 311115, China
* Correspondence: wke@mpu.edu.mo

Abstract: The task of dense video captioning is to generate detailed natural-language descriptions for an original video, which requires deep analysis and mining of semantic captions to identify events in the video. Existing methods typically follow a localisation-then-captioning sequence within given frame sequences, resulting in caption generation that is highly dependent on which objects have been detected. This work proposes a parallel-based dense video captioning method that can simultaneously address the mutual constraint between event proposals and captions. Additionally, a deformable Transformer framework is introduced to reduce or free manual threshold of hyperparameters in such methods. An information transfer station is also added as a representation organisation, which receives the hidden features extracted from a frame and implicitly generates multiple event proposals. The proposed method also adopts LSTM (Long short-term memory) with deformable attention as the main layer for caption generation. Experimental results show that the proposed method outperforms other methods in this area to a certain degree on the ActivityNet Caption dataset, providing competitive results.

Keywords: dense video caption; video captioning; multimodal feature fusion; feature extraction; neural network

MSC: 68T45

Citation: Huang, X.; Chan, K.-H.; Ke, W.; Sheng, H. Parallel Dense Video Caption Generation with Multi-Modal Features. *Mathematics* **2023**, *11*, 3685. https://doi.org/10.3390/math11173685

Academic Editor: Zhiming Cai, Wencai Du, Zhihai Wang, Zuobin Ying

Received: 27 July 2023
Revised: 22 August 2023
Accepted: 23 August 2023
Published: 26 August 2023

1. Introduction

With the widespread use of video as an information transmission medium, recordings for playback and live broadcasting have become increasingly popular today. Video processing has gradually become a hot research topic in computer vision [1,2]. Video caption generation is an important task that provides understanding and representation of videos between two media: frame-to-text. This task has also involved critical artificial intelligence (AI) technologies [3,4]. Such technologies have potential applications in the development of smart glasses to assist the visually impaired, intelligent commentary on sports events, early childhood education, and the generation of video surveillance reports [5–7].

Dense video captioning tasks mostly use datasets directly crawled from online sources such as YouTube, whose videos typically consist of long content without pruning [8]. Unlike traditional video captioning, which uses concise sentences to explain the video content, dense video captioning requires not only dividing long videos into various events, but also describing the behaviours in a series of events as accurate as possible. The objective of dense video captioning is to generate as detailed and general description for videos clearly.

A cross domain method is required to perform such video analysis and comprehension, so as to represent the events as sentences. In practice, the video is always based on a sequence of frames of fast-playing images that also contain important audio information usable in the captioning. Therefore, the computer needs to perform a high-level understanding of the video content, aiming to localise/categorise the interesting objects, then represent their motion and behaviour in detail.

Reviewing previous achievements, the process of generating dense video captions can be summarised into two main procedures: dividing event regions and generating descriptive sentences, as shown in Figure 1a. There are different orders to arrange the video localisation and description, mostly following a sequential top-down or bottom-up structure. Inevitably, this makes the generation of captions more dependent on the quality of the previous steps [9–12]. In other words, the performance of the generated descriptions can decrease if the former module does not perform well, and the complexity of the module design is less relevant. Moreover, these methods are not trained end-to-end in the traditional sense and require additional steps for extensive and complicated training, which also affects the results to a certain extent. The parallel method shown in Figure 1b defines dense video caption generation as a set of prediction tasks and decodes the divided events and sentences simultaneously, which solves the problem of dependence on previous results [13]. Although this type of method has produced good results, there are still bottlenecks in the decoding branch that limit the fine-grained description of the video. Therefore, further improvement is necessary.

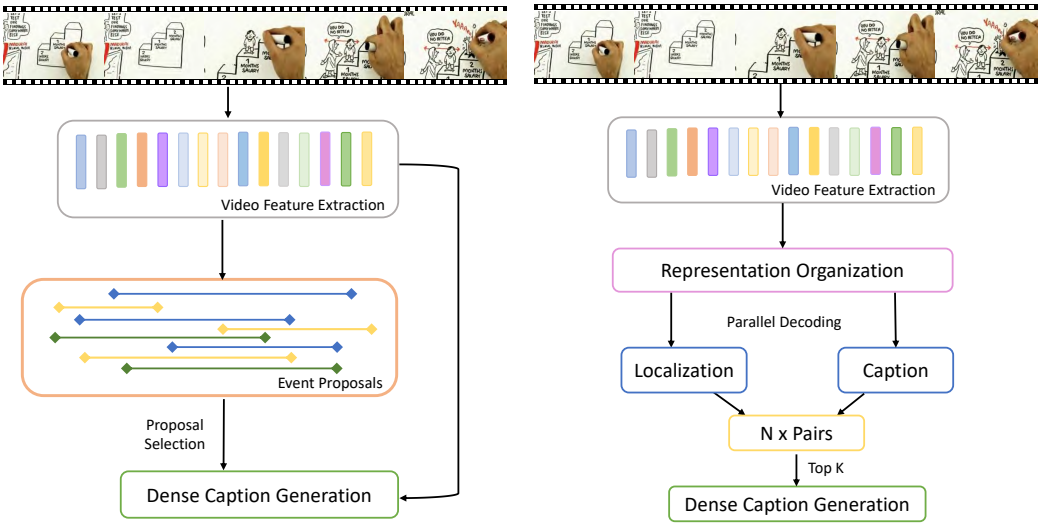

Figure 1. Comparison between existing methods and the parallel method. (**a**) Localise-then-describe method. (**b**) Parallel method.

This work builds on previous research and further explores the techniques to alleviate the fine-grained bottleneck that arises when generating dense video captions in parallel. Additionally, this approach fully exploits the multi-modal features of videos. The main contributions are summarised below:

- A novel model for video caption generation is proposed, which effectively utilises the visual–audio features. Unlike the common conventional sequential mechanism of localise-then-describe approach, the proposed model reasonably associates the proposal and caption modules through parallel paths, which enhances the comprehensiveness of the textual expression.
- In addition, a simplified method is proposed to eliminate the redundancy generated by the anchor mechanism on which the maximum suppression algorithm relies and to reduce the steps of manually setting hyperparameters for end-to-end training of the model.

- The decoding side introduces the representation organisation module as intermediate information for event localisation and description and extracts temporal boundary information from the video to generate all potential events.

2. Related Work

In early years of video captioning, the main approach was to use template matching to generate logical and specific sentences through keyword sorting or selection. However, this approach only supported simple sentence structures and was inflexible for complex event representation, making it poor for understanding multiple scenes in long videos [14]. With the development of powerful neural networks, deep learning approaches using artificial intelligence (AI) have become possible for extracting multimedia information, marking a milestone in the field of video captioning [15]. Inspired by image captioning methods, such technology can be directly extended to video captioning, enabling the discovery of correlations between image sequences [16–18]. For dense video captioning tasks, significant breakthroughs have been made in the generation of detailed and rich descriptive sentences for event representation. Deep learning in image processing has shown that the Sequence-to-Sequence (Seq2Seq) framework can be applied to video captioning [3,19]. This framework consists of two neural network models over an encoder and a decoder. Most of the input videos use CNN-based network models such as VGGNet [20], VGGreNet [21], or ResNet [22] for the encoder. Conversely, the decoder uses RNN-based network models to generate native sentences for the final output. This encoder–decoder design projects visual features into text sentences, extracting important abstract information and discarding noise in the application [23].

In recent years, attention mechanisms have shown outstanding performance when integrated into various neural models. They have the potential to play a prominent role in image captioning and are increasingly used to address the problem of video captioning [24,25]. Dense video captioning involves visual understanding processes that locate different events in a video and generate descriptive captions for each interesting object. This approach represents video content in detail by transforming frame sequences into multiple descriptive sentences among multiple clips in a long video [26–28]. In the research of Shen et al. [29], dense image captioning is migrated to the video field by combining the multi-scale suggestion module and a visual context perception mechanism. Additionally, Huang et al. in their work [30] divide the long video into several different regional sequences and comprehensively express the video content. However, the feature extraction of regions within the frame sequence is a complicated process in the Seq2Seq framework. It is not an end-to-end method in the traditional sense, and extensive hyperparameters may be required for the input of a non-fixed length video. Furthermore, if too many small regions are split within a shot, it becomes difficult to represent the entire video and to categorise objects within specific regions, making it difficult to discover their correlations.

Since a video can be considered as a sequence of images with an additional time dimension, multiple scene events can occur, and objects can appear and perform actions within a range of frames. To address the neglected time series problem in video more comprehensively, Tran et al. [5] introduced a 3D CNN approach for extracting video features. An advanced model, C3D, was developed based on the 3D CNN, which can handle more complex cases in terms of various scenes [1,31–33]. In addition, Carreira et al. [34] added optical flow features in the encoder part, combined with C3D to form a new Inflated 3D ConvNet (I3D) model, which enhanced the quality of the extracted video features to a certain extent. Qiu et al. [35] and others took advantages of the residual connection's ability to deepen the convolutional network, decomposed the 3D video features into a 2D spatial convolution and a 1D temporal convolution, and constructed a Pseudo-3D Residual Network (P3D) that greatly reduced the need for labelled video data, increased the network's depth, and reduced the amount of convolutional computation.

To enhance the capability of feature extraction, it is important to improve the caption generation module. Caption generation methods use advanced NLP technology such as

LSTM [36], BERT [37], Transformer [38], and other variants [39], which have been applied to video captioning with competitive results [40]. In the approach of Pasunuru et al. [41], a multi-task learning method is proposed that uses LSTM to share parameters between different tasks and improve model performance with more data. Additionally, Shetty et al. [42] used an improved deep LSTM in the decoding section, trained on two different video features, and used an evaluation network to judge video features and generated sentence keywords to improve sentence quality. The EEDVC model, the first to encode video features using Transformer, solves the problem of long-term dependence in LSTM [43]. It converts each extracted proposal into a mask, combines it with video features, and completes the end-to-end training of the model.

In addition, Yao et al. [44] applied the attention mechanism in NLP to the video captioning task. They introduced the attention weight α based on the codec structure. It is different from the attention to region in image captioning but is used to compute different features of a video along the time sequence. This approach allows the decoder to automatically select a more relevant time period when generating words, helping the model filter out irrelevant information and reduce the workload, ultimately improving the evaluation index.

The captioning and event generation modules in the method described above can only be trained independently. However, the results of the captioning module can theoretically be used to train the proposal process. In order to improve the localise-then-describe scheme and fully exploit the two subtasks of event localisation and caption generation, Li et al. [11] proposed a bridging idea using desperation regression to link the two subtasks. This approach allows the prediction of description complexity in the proposal module. The caption module captures video features and achieves the goal of jointly training the two subtasks. However, since many generated sentences are redundant and produce inconsistent results, it is necessary to use Non-Maximum Suppression (NMS) [45] or an Event Sequence Generation Network (ESGN) [9] to select the proposal. These modules introduce many hyperparameters and are highly dependent on manual thresholding strategies, which can affect the model results. The PDVC model [46] proposed a parallel decoding method to address these issues. By designing two parallel prediction heads (localisation head and caption head), both the scope and text description of the event query are predicted. This approach allows the PDVC to directly use the video features to match the split target events, thus providing more unique features. Experiments have shown that the parallel design can make the loss of the caption module improve the performance of event localisation. Furthermore, the absence of thresholding and NMS mechanisms makes model training more efficient. Li et al. proposed a transfer learning method that can simultaneously utilise knowledge from two types of source domains, spatial appearance and temporal motion, and transfer them to the target domain [47]. The core of the CMG-AAL [48] model is a cross-modal foundation module that is composed of two complementary attention mechanisms, which can effectively establish correspondence between text and vision, thereby improving the model's understanding and generation capabilities.

Furthermore, existing methods for analysing and understanding video content mostly rely on visual features, without taking into account clues provided by other modalities such as sound or subtitles [49]. However, incorporating other modalities can help computers understand video content and produce more detailed text descriptions. For instance, in a video of a female announcer broadcasting, the content of the broadcast may not be clear without sound. To address this issue, Jin et al. [50] developed a model that combined multiple types of features by extracting them separately, weighting the average, and, finally, fusing them together as the input for LSTM. This model aims to make full use of more comprehensive feature information to represent videos and proposes a new approach to using multi-modal features to improve the quality of video captions. Other models, such as EMVC [51] and BMT [52], have also proposed methods for incorporating audio features. The EMVC model integrates audio features to support visual cues in event generation, while the BMT model extracts feature vectors for both video and audio using I3D and

VGGish, respectively, and uses a Transformer framework to improve the quality of the generated text.

Having reviewed the existing work, it suggests that there is still room for improvement in the relationship between video event localisation and text description. Currently, there are bottlenecks in the branch of parallel decoding methods, and the auxiliary function of multi-modal features is equally important. However, due to the challenge of unifying different video lengths, the calculation of the number of contained events remains difficult. Therefore, developing a method for the computer to use the characteristics of multi-modal data in the video and to fully consider the connection between the two subtasks of event localisation and caption generation is still a challenging task.

3. Methodology

For an unedited video, the task of dense video captioning is to divide multiple events in the video and generate corresponding description sentences. In order to fully exploit the correlation between caption and event proposals, as well as the multi-modal features of the video to improve the efficiency of text description generation, we design a parallel multi-modal dense video caption generation model.

3.1. Model Overview

The entire framework and the data flow between the various parts in the schematic diagram are shown in Figure 2.

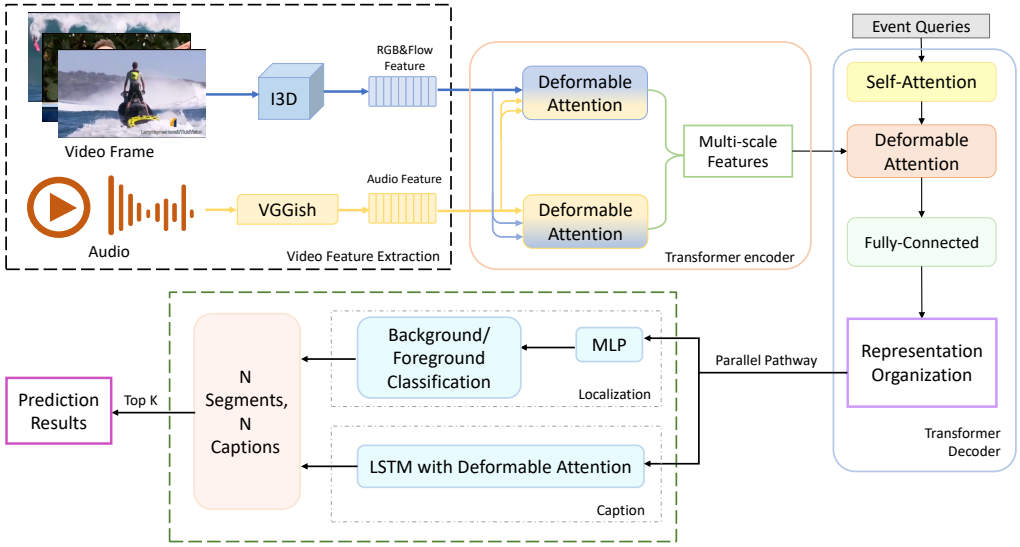

Figure 2. Overall framework of the proposed model.

The proposed model uses pre-trained I3D and VGGish to extract the visual and audio features of the video, respectively. It then merges the encoded multi-scale video features into a more characteristic feature set based on the deformable Transformer encoder and decoder framework [53]. Such a representation organisation allows for a more intuitive understanding of the core context of the video. In addition, the model inputs the video features into the captioning and positioning modules in parallel, rather than directly performing proposal localisation and generating captions from the video features in sequence. Finally, the model selects multiple sets of proposal-caption pairs with higher confidence to ensure content integrity and produces more logical and detailed video captions.

3.2. Video Encoder

The video encoding process consists of two parts: a multi-modal feature extraction component and a position encoder. The convolutional network is responsible for extracting

feature information from the video, while a sequence data encoder based on the Transformer framework is used to understand the information association between contexts.

3.2.1. Feature Extraction

In order to enhance the use of multi-modal information in videos, the work decided to incorporate audio features based on the findings of visual modality research. To extract the features of each modality separately, the work used the proven pre-trained combination of I3D and VGGish.

For the visual modality feature extraction, the work chose the I3D network, which can solve the problem of 2D CNN not being able to extract spatial features in videos and adapt to video inputs of different lengths and resolutions by adjusting the network structure and output characteristics accordingly. I3D is constructed by expanding 2D CNN into 3D CNN, which can inherit the knowledge and parameters learned by 2D cellular neural networks in image classification and recognition tasks, without the need for training from scratch. Compared with some other models (such as C3D with only 8 layers), the 20 layers of I3D have a deeper and more complex network structure, including a multi-branch structure composed of multiple convolutional kernels of different sizes, which can capture features at different scales and reduce the number of parameters and computational costs. The I3D network can not only process RGB features, but also optical flow features and average the outputs of the two networks during testing, thus integrating colour information and motion information. Therefore, it can extract spatial features present in videos better than other options.

For audio modality feature selection, we use VGGish to extract the features, which has a strong generalisation ability with pre-trained parameters and can effectively transform audio features into feature vectors that conform to natural language logic. In our work, VGGish converts audio into 128-dimensional semantic feature vectors, which have stronger expressiveness with high-level feature vectors.

3.2.2. Feature Encoding

Previous methods have attempted to concatenate features with common weights, but this has proven to be insufficient. It is not possible to fuse them together because the visual and audio features have different dimensions extracted form a video. To address this issue, the work introduced the deformable Transformer as a novel component in the proposed model. The deformable Transformer can distinguish different attention heads in the framework, thereby improving the model expressive and generalisation abilities. Among them, deformable sampling locations are introduced into the pre-filtering mechanism to reduce computational complexity and memory consumption, while maintaining efficient information transmission. The deformable Transformer uses deformable attention to replace the self-attention module in the encoding part of traditional Transformers, as well as the cross-attention class module in the decoding part. This allows the model to better capture long-distance dependencies and local details in the sequence, thereby improving performance. This process can be thought of as converting video features from a video sequence to a set sequence, which is essentially a learnable positional coding. The deformable Transformer encodes the position of the extracted multi-modal video features, unfolds pixels into a one-dimensional sequence, and computes the correlation between pixels; thus, the global information of the video is fully learned. To better exploit the multi-scale features in event prediction, the work added L timescale convolutional layers to obtain feature sequences spanning multiple resolutions. The multi-scale deformable attention module helps alleviate the convergence problem of self-attention by focusing on the sparse space near the reference point.

Let $\boldsymbol{X}$ be the set of feature maps, given by

$$\boldsymbol{X} = \left\{ \boldsymbol{x}^l \right\}_{l=1}^{L}, \tag{1}$$

where each multi-scale feature map $\boldsymbol{x}^l$, with size $\boldsymbol{x}^l$ is $C \times H^l \times W^l$, is extracted from the feature map output in the previous stage for $1 \leq l \leq L$. A projection matrix $\boldsymbol{H}_{milt}$ is used to project the sample offset into the features. The $\boldsymbol{H}_{milt}$ matrix is associated with a linear operator, expressed as the offset of the t-th sampling point of the i-th query element on the l-th scale in the m-th attention head,

$$\boldsymbol{H}_{milt} = \phi_l(\hat{\boldsymbol{p}}_i) + \Delta \boldsymbol{p}_{milt}, \tag{2}$$

where $\hat{\boldsymbol{p}}_i$ is the coordinate of the reference point of each query element $\boldsymbol{q}_i$ in the $[0, 1]^2$ space, ϕ_l is a function that converts the normalised reference point to the input feature map at the l-th layer, and $\Delta \boldsymbol{p}_{milt}$ is a sampling offset that is derived from a linear transformation on the query element.

Then, the deformable Transformer is used to understand the long-distance associations of different segments in long videos and output multi-scale video features. The Multi-Scale Deformable Attention (MSDAttn) dynamically adjusts sampling positions and attention weights by using learnable offsets, allowing for adaptive allocation of attention resources based on the characteristics and needs of the data. MSDAttn can also reduce computational complexity and memory consumption by sampling sparsity, improving the running speed and performance of the model. The traditional attention mechanism requires fully connected operations on all input features, which leads to an exponential increase in computational and storage capacity as the input features increase. This uses deformable convolutional kernels to sparsely sample input features, selecting only a portion of important features for attention calculation, greatly reducing computational and storage costs.

$$\text{MSDAttn}(\boldsymbol{q}_i, \hat{\boldsymbol{p}}_i, \boldsymbol{X}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left(\sum_{l=1}^{L} \sum_{t=1}^{T} A_{milt} \cdot \boldsymbol{W}'_m \boldsymbol{X} \boldsymbol{H}_{milt} \right), \tag{3}$$

where, the MSDAttn module samples L points from multi-scale feature maps, instead of sampling T points from single-scale feature maps. The m denotes attention heads, M is the number of heads, t denotes sampling keys, and T is the total number of sampling keys $(T \leq HW)$, and A_{milt} represents the attention weight of the t-th sampling point in the m-th attention head, which is calculated by using $softmax$ on the query feature $\boldsymbol{q}_i$.

3.3. Video Decoder

In the decoder section, the work used the deformable Transformer to decode video features at multiple scales. These features are then fed in parallel to the localisation and captioning modules. As an intermediate information hub, the work also constructed a representation organisation module that receives video temporal features and implicitly generates multiple event proposals.

3.3.1. Feature Decoding

The work incorporated the query mechanism and set prediction loss from the DETR framework for object detection into the video captioning domain during the decoding phase. The input query is a learnable vector or parameter that represents an initial estimate of the event location at the input layer, which is then updated and optimised by attention at each decoding layer. The output queries are considered as the representations of N events, with each output token from the decoder corresponding to a potential event. Therefore, all tokens predict a set of events without changing the meaning of the original text. Moreover, the work also included a module for organising representations prior to the parallel method. This module serves as intermediate information for event localisation and sentence generation. Representation organisation uses a feed-forward neural network to identify the most important features in a spatio-temporal context and generate all the possible proposal representations. Each representation becomes the central information of the event, including the timestamps and sentences.

It is important to note that the input event query is used to replace the anchor. This method is learnable and allows the model to automatically identify regions in the image that may contain objects, with a capacity of up to 100 such regions. Then, using a bipartite graph matching method, valid prediction boxes are filtered from the 100 prediction boxes, and the loss function is calculated. Therefore, the event query is equivalent to replacing the anchor with a learnable method that avoids generating a large number of invalid boxes that result from using the anchor.

The number of input event queries can be controlled, which limits the maximum number of proposals the model can generate. This may result in a low recall rate if the number of queries is set too low, which may affect the accuracy of the subsequent positioning module and result in the loss of important information. In turn, although it may improve the recall rate to some extent if the number of queries is set too high, it reduces the accuracy rate of the title generation module, resulting in repeated words or unnatural language logic. Therefore, the number of event queries must be carefully considered and counted to ensure the best possible results.

$$\boldsymbol{N}_{set} = \boldsymbol{argmax}(\boldsymbol{v}_{len}), \tag{4}$$

where the $\boldsymbol{v}_{len}$ represents a fixed-size feature vector for prediction. By using the deformable Transformer in combination with the event query during decoding, this work can more accurately divide the extreme point regions of object boundaries, resulting in more precise event edge detection and can accelerate the network training by focusing on sparse spatial locations and combining multi-scale feature representations.

3.3.2. Parallel Pathway

The query features and reference enhanced by the representation organisation are sent to the localisation and captioning proposal modules in parallel. This is followed by a one-to-one matching process and filtering to select the combination that best matches the actual video content to achieve dense video captioning.

Localisation Module

The primary objective of this module is to match the output of the representation organisation module to the video on a one-to-one basis and to determine the centre and timestamp of the event proposal. The work used a multi-layer perceptron for box prediction, which involves regression of the event boundary and binary classification of foreground and background. In box prediction, the work calculates the relative offset of the reference point from the actual ground and determines the centre and duration of the event. The purpose of binary classification is to generate a confidence score for the foreground of each event proposal. The final output time proposal consists of the start time $\boldsymbol{t}_i^{star}$, the end time $\boldsymbol{t}_i^{end}$, and the confidence of the location proposal $\boldsymbol{c}_i^{loc}$.

Caption Module

The captioning module allows the model to focus on video frames that are highly correlated with the output words. This helps mine fine-grained interactions between words and video frames. Most traditional methods rely on LSTM with soft attention to generate captions. This allows the importance of each element to be dynamically determined by restricting the attentional field to event proposals and ensures that the generated words are all contained in the same event for sentence and video matching. However, the work wants to use a parallel method for this task and cannot rely directly on the information from the positioning module. If the work used the above method, the association between the reference text and the video would be lost. Therefore, the work proposed to use deformable soft attention combined with LSTM to generate captions, as shown in Figure 3.

By using deformable soft attention, a reference point can be predicted for each input query as the reference position of the centre point of the event proposal, and the weight of

the sampling point can be calculated to limit the event to a more precise region. Specifically, the hidden state $\boldsymbol{h}_{it}$ of LSTM at the t-th moment is given by

$$\boldsymbol{h}_{it} = \mathrm{LSTM}\left(\boldsymbol{w}_{i,t-1}, \boldsymbol{h}_{i,t-1}, \widetilde{\boldsymbol{q}}_{i,t-1}\right), \tag{5}$$

where $\boldsymbol{w}_{i,t-1}$ is the word generated at the previous moment, and $\widetilde{\boldsymbol{q}}_{i,t-1}$ is the event query output by the representation organisation. Then, take $[\boldsymbol{h}_{it}, \widetilde{\boldsymbol{q}}_i]$ as the query in deformable soft attention to obtain the context feature $\boldsymbol{z}_{it}$,

$$\boldsymbol{z}_{it} = \mathrm{DSAttn}([\boldsymbol{h}_{it}, \widetilde{\boldsymbol{q}}_i]), \tag{6}$$

and the output features $\boldsymbol{z}_{it}$ are restricted to a relatively small region to narrow down the scope of event proposals. Next, LSTM takes $\boldsymbol{z}_{it}$, $\boldsymbol{w}_{i,t-1}$, and $\widetilde{\boldsymbol{q}}_i$ in series as input to obtain the generated t-th word $\boldsymbol{w}_{it}$ and calculates the probability distribution p of the word $\boldsymbol{w}_{it}$ in the entire vocabulary,

$$p(\boldsymbol{w}_{it} \mid \boldsymbol{w}_{i,t-1}) = \mathit{Softmax}(\boldsymbol{w}_{i,t-1}, \boldsymbol{z}_{it}, \widetilde{\boldsymbol{q}}_i). \tag{7}$$

According to the probability distribution p, the embedded word sequence can be continuously sampled until an End-of-Sentence (<EOS>) symbol is encountered, resulting in a complete sentence.

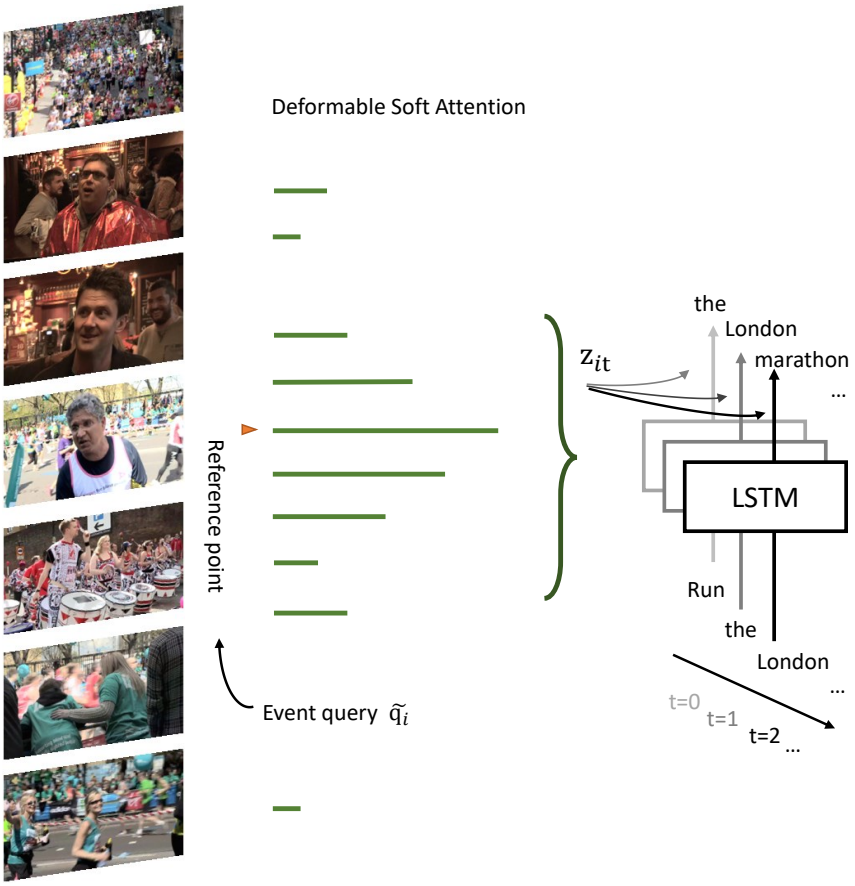

Figure 3. LSTM with deformable attention.

3.3.3. Caption Generation

The localisation information and captions obtained by the parallel method are combined into a proposal set. This set also requires checking for proposal-caption pairs, similar to the number of event queries introduced in Section 3.2.2. However, too many features for output can lead to redundancy and poor readability, while too few features can lead to

missing of important information from the video. To avoid these problems, we perform a bipartite match between N proposals in the proposal set and K captions in the Ground Truth (GT), and measure the difference between the predicted foreground and background regions of the model and the actual annotated regions through the cross-entropy loss function, to obtain proposal-caption pairs. To ensure consistency in semantic information and size, we use the set prediction loss for the calculation, which is the weighted sum of the individual module losses,

$$\boldsymbol{L} = \mu\left(\boldsymbol{L}_{giou} + \boldsymbol{L}_{cls} + \boldsymbol{L}_{cr-e} + \boldsymbol{L}_{cap}\right), \tag{8}$$

where $\boldsymbol{L}_{giou}$ represents the generalised IOU between the generated timestamps and the GT, $\boldsymbol{L}_{cls}$ means the focal loss of binary matching, $\boldsymbol{L}_{cr-e}$ represents the cross-entropy loss between the generated number of event proposals and the GT number, and $\boldsymbol{L}_{cap}$ is the cross-entropy loss between the generated words and the GT words. Based on the calculation result, the work keeps use of the prediction frame with a confidence level higher than the threshold as the final output and obtains a text description that is more suitable for the video content and has an accurate time stamp.

4. Experiment

In order to evaluate our newly proposed method for dense video caption generation, this work verifies the performance on the ActivityNet Caption dataset and compares the results with those from the state-of-the-art methods.

4.1. Dataset and Data Pre-Processing

The ActivityNet Caption dataset is a publicly available dataset that is widely used for dense video captioning tasks. It covers several domains relevant to our method. The dataset consists of 20 K video clips, each with an average duration of 2 min. Each clip is annotated with the events, including the start and end time of each event, along with a human-written textual description of the event content. The dataset is derived from YouTube videos, but some videos have been removed or altered by their original authors and are not available for direct download. The work used an alternative approach provided by the authors to retrieve the missing videos and obtain a complete dataset. Like most researchers, here, we split the dataset into training, validation, and test sets. The training set contains 10 K clips, the validation set contains 4 K clips, and the test set contains 5 K clips. However, the labels for the test set have not yet been released, so this work used the validation set for experimentation and comparison purposes.

Before training the model, the work preprocessed the reference sentences by converting all letters to lowercase, removing non-text characters, and adding special markers `<BOS>`and `<EOS>` at the beginning and end of each sentence. To reduce the impact of large vocabularies and low-frequency words, the work replaced words occurring less than five times with `<UNK>`. However, this replacement resulted in the loss of some semantic information, also known as the out-of-bag error. This work also added a start token to the decoder input, which allowed the caption to be generated word-by-word until the end marker was reached.

4.2. Implementation

Our model was trained on an Ubuntu 20.04 system using two NVIDIA GeForce RTX 3070 GPUs, and the work used PyTorch [54] as the neural network engine. For multi-modal feature extraction, the work followed the approach of BMT [52]. This work used I3D to extract 64 RGB features and 64 optical flow features at 25.0 fps with a size of 224, producing feature vectors with a dimension of 1024. We also used VGGish to extract audio features. The learning rate was set to $1e{-}4$, and the batch size was 32. The work used a multi-scale deformable attention of size 4 and applied a two-layer deformable Transformer to encode and decode the video features. The hidden layer size of the feed-forward network was set to 2048, and we set the number of event queries to 10. In the caption module, the work set the hidden layer dimension of the LSTM to 512 and the word embedding size to 468.

The work used a dynamic learning rate and set the warm-up steps to 10 epochs, gradually increasing from 0 to $5e-5$. We trained the relation detection with fixed prediction loss and set the learning rate to $5e-4$. This work used the Adam optimiser [55] for the loss function.

4.3. Results and Analysis

The work tested our framework by following the implementation details above and performing experiments on the ActivityNet Caption dataset. We compared the results with those obtained by the state-of-the-art methods. This work also performed an ablation study to investigate how different modules in our framework affected the experimental results. Finally, the work presents the results of the qualitative analysis in a visualisation, which provides a clearer picture of the benefits of our proposed framework.

4.3.1. Comparison to the State-of-the-Art

This work contrasted our proposed model with the state-of-the-art methods for the dense video captioning task, consisting of EEDVC [43], DCE [8], MFT [26], WLT [27], SDVC [9], EHVC [31], MDVC [28], BMT [52], EMVC [51], PPVC [13], and PDVC [46]. The contrast results are displayed in Table 1.

Table 1. Comparison of the performance of our proposed method with the state-of-the-art methods on the ActivityNet Captions dataset.

Models	*B*@1	*B*@2	*B*@3	*B*@4	METEOR	CIDEr
EEDVC [43]	9.96	4.81	2.91	1.44	6.91	9.25
DCE [8]	10.81	4.57	1.90	0.71	5.69	12.43
MFT [26]	13.31	6.13	2.84	1.24	7.08	21.00
WLT [27]	10.00	4.20	1.85	0.90	4.93	13.79
SDVC [9]	**17.92**	7.99	2.94	0.93	8.82	-
EHVC [31]	-	-	-	1.29	7.19	14.71
MDVC [28]	12.59	5.76	2.53	1.01	7.46	7.38
BMT [52]	13.75	7.21	3.84	1.88	8.44	11.35
EMVC [51]	14.65	7.10	3.23	1.39	9.64	13.29
PPVC [13]	14.93	7.40	3.58	1.68	7.91	23.02
PDVC [46]	-	-	-	**1.96**	8.08	28.59
Proposed	15.23	**8.02**	**3.91**	1.75	**9.68**	**29.17**

Bold font indicates the highest result.

As reported in Table 1, *B*@*N* is an evaluation metric known as BLEU [56], which measures the quality of translations by comparing the matching degree of *N*-grams in the candidate and reference translations. BLEU is commonly used for text generation tasks in NLP. METEOR [57] is based on BLEU and uses the F-value as the final evaluation metric, taking into account both recall and precision. CIDEr [58] calculates the similarity between candidate and reference sentences, making it suitable for image and video captioning evaluation tasks. Higher quality text descriptions receive higher scores on these evaluation metrics.

The work thoroughly analysed the comparative results based on the evaluation metrics mentioned above. Our proposed method performed slightly worse in *B*@1 and *B*@4, but outperformed other methods in terms of METEOR and CIDEr. The SDVC uses reinforcement learning to train the model, resulting in a higher *B*@1 score compared to all other methods. Furthermore, all our metrics are higher than those of BMT and EMVC, which also use multi-modal features as inputs. This demonstrates the feasibility of using parallel paths to generate textual descriptions. When compared to the two parallel decoding methods of PPVC and PDVC, it can be seen that most of our metrics are slightly higher. It is worth noting that PDVC's input also includes visual and audio features, indicating that our method is still competitive.

4.3.2. Ablation Study

The work conducted several comparative experiments to analyse the impact of different components of our proposed model on the output results. This includes comparing the performance of the localisation module, investigating the effect of LSTM with deformable attention on the caption module, and assessing the influence of multi-modal features as inputs on model generation.

Table 2 shows the quality of event localisation on the ActivityNet Captions dataset. Our method achieves a higher F1 score compared to MFT and SDVC. It outperforms the traditional localise-then-describe methods and uses event proposal networks to generate event proposals. Meanwhile, our method uses a representation organisation to filter better quality events and overlays a localisation module to accurately locate event proposals. As a result, our method significantly outperforms MFT on various metrics and achieves better overall results compared to other models. These metrics demonstrate the effectiveness of our proposed localisation module. The $@tIoU$ represents the temporal intersection of the unions, with 4 thresholds of $\{0.3, 0.5, 0.7, 0.9\}$.

Table 2. Performance comparison of the localisation module.

Models	Proposal Network	Recall ($@tIoU$)					Precision ($@tIoU$)					F1
		0.3	0.5	0.7	0.9	avg	0.3	0.5	0.7	0.9	avg	
MFT [26]	✓	46.18	29.76	15.54	5.77	24.31	86.34	68.79	38.30	12.19	51.41	33.01
SDVC [9]	✓	**93.41**	76.40	42.42	10.10	55.58	96.71	77.73	**44.84**	10.99	57.57	56.56
PPVC [13]	-	91.71	78.90	56.73	20.60	61.98	96.23	73.80	37.66	12.31	55.07	58.33
PDVC [46]	-	89.47	81.91	44.63	15.67	55.42	**97.16**	78.09	42.68	14.40	58.07	56.71
Ours	-	90.25	**81.97**	48.39	**20.77**	**63.08**	95.51	**79.31**	39.60	**15.72**	**59.10**	**58.43**

Bold font indicates the highest result.

The work compared different methods for the caption module of our model: Vanilla LSTM, LSTM with Soft Attention (SA), and LSTM with Deformable Soft Attention (DSA). The results are shown in Table 3. When generating event proposals and captions in parallel, using Vanilla LSTM results in a lack of interaction between text and features. SA training does not allow all attention weights to be concentrated on a fixed area. Therefore, using DSA as our proposed method effectively addresses the issue of parallel methods not directly accessing event proposals and achieves better results.

Table 3. The effect of LSTM with deformable attention on the caption module.

Method	*B*@1	*B*@2	*B*@3	*B*@4	METEOR	CIDEr
Vanilla LSTM	14.88	7.15	3.84	1.70	8.91	27.39
LSTM with SA	**15.44**	7.61	3.70	1.68	9.23	28.85
LSTM with DSA(Ours)	15.23	**8.02**	**3.91**	**1.75**	**9.68**	**29.17**

Bold font indicates the highest result.

The work also confirmed that combining different types of features can help improve the quality of the text generated by the model. The comparison of the results is shown in Table 4. Experimental results also report that using only audio features is not sufficient to improve the quality of the text and can even have a negative impact on the performance of the model. Using only visual features can produce good results, but is still not as effective as using multi-modal features. By supplementing visual features with audio features, the generated text can be significantly improved.

Table 4. The impact of multi-modal features on the quality of generated captions.

Method	*B*@1	*B*@2	*B*@3	*B*@4	METEOR	CIDEr
Visual-only	13.66	7.43	3.11	1.27	8.35	23.58
Audio-only	13.07	6.69	2.94	1.13	6.81	16.20
Proposed	**15.23**	**8.02**	**3.91**	**1.75**	**9.68**	**29.17**

Bold font indicates the highest result.

4.3.3. Qualitative Analysis

This work demonstrates the proposed method on the act dataset. The generated text descriptions can be seen in Figure 4. The GT is also included for reference.

As shown in Figure 4, the proposed method divides a 2 min video into 4 proposals event and generates logical text descriptions. Compared to GT, our method accurately separates the video based on its content and scenes and avoids the event redundancy. The generated captions fully describe the content of each event. However, GT clearly identifies the name of the male protagonist as "Mr. Bean", which our method does not. This difference may be due to the fact that the video clip is from a popular TV series and GT's captions are manually marked, whereas the protagonist is relatively well known. This comparison shows that our method has not yet successfully identified the protagonist and matched his name in a complex scene.

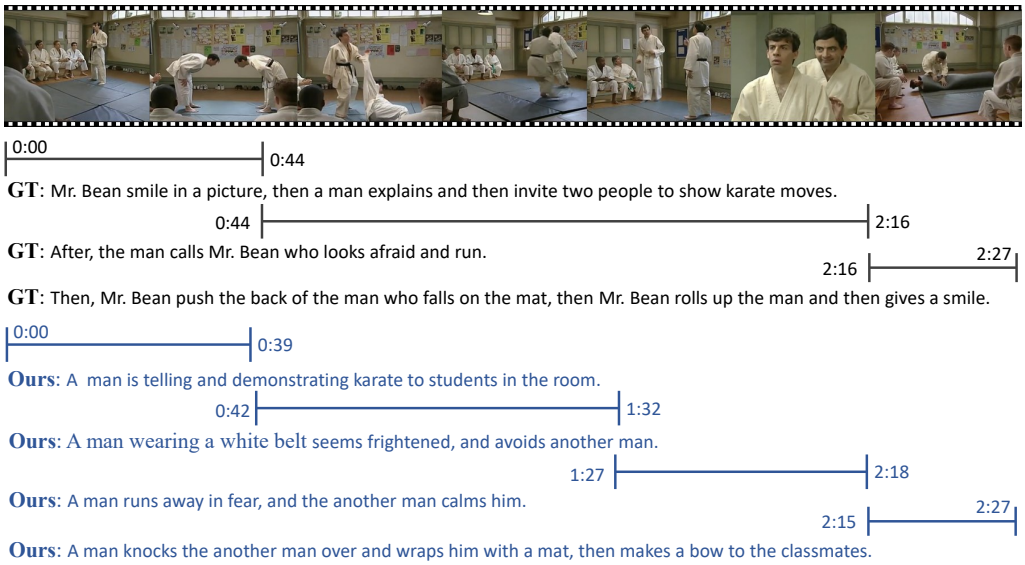

Figure 4. Results of a qualitative analysis of a video from the ActivityNet Caption dataset. The predicted results of the proposed model are compared with the GT reference.

5. Conclusions

The paper has proposed a new model for dense video captioning that achieves competitive results on the ActivityNet Caption dataset, with the following particular improvements.

- The approach was able to effectively exploit the multi-modal features of video and highlights the potential of the audio modality to enhance video details.
- A deformable Transformer was used to encode and decode features, which eliminated the need for complex anchor mechanisms and hyperparameter constraints in non-maxima suppression.
- A representation organisation module was introduced to improve the link between features and context.
- The parallel method for the two subtasks of localisation and captioning was enhanced, allowing fine-grained interaction between the submodules and improving the comprehensiveness and accuracy of the text descriptions generated by the model.

Based on the experimental results, it is evident that our method excels in event localisation and generates text descriptions that follow linguistic logic, while presenting video content from multiple perspectives. Compared to other dense video captioning methods, our proposed method has clear advantages. In the future, we will aim to overcome the branch bottleneck of existing parallel methods, improve the computer's ability to understand and represent video content, and ultimately achieve the beautiful vision of artificial intelligence.

Author Contributions: Conceptualisation, X.H., K.-H.C., W.K. and H.S.; methodology, X.H., K.-H.C., W.K. and H.S.; software, W.K. and H.S.; validation, X.H., K.-H.C., W.K. and H.S.; formal analysis, X.H., W.K. and H.S.; investigation, X.H.; resources, K.-H.C., W.K. and H.S.; data curation, X.H., K.-H.C., W.K. and H.S.; writing—original draft preparation, X.H., K.-H.C., W.K. and H.S.; writing—review and editing, X.H., K.-H.C., W.K. and H.S.; visualisation, X.H., W.K. and H.S.; supervision, W.K. and H.S.; project administration, W.K. and H.S.; funding acquisition, W.K. and H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Key R&D Program of China (No. 2019YFB21 01600), the National Natural Science Foundation of China (No. 61872025), the Macao Polytechnic University (RP/FCA-06/2023, RP/ESCA-03/2020), and the Open Fund of the State Key Laboratory of Software Development Environment (No. SKLSDE-2021ZX-03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Thanks for the support from Macao Polytechnic University and HAWKEYE Group.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018. [CrossRef]
2. Sighencea, B.I.; Stanciu, R.I.; Căleanu, C.D. A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction. *Sensors* **2021**, *21*, 7543. [CrossRef]
3. Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; Saenko, K. Sequence to Sequence—Video to Text. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Piscataway, NJ, USA, 2015. [CrossRef]
4. Tang, M.; Wang, Z.; Liu, Z.; Rao, F.; Li, D.; Li, X. CLIP4Caption: CLIP for Video Caption. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Online, 20–24 October 2021; ACM: New York, NY, USA, 2021. [CrossRef]
5. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Piscataway, NJ, USA, 2015. [CrossRef]
6. Wu, Y.; Sheng, H.; Zhang, Y.; Wang, S.; Xiong, Z.; Ke, W. Hybrid Motion Model for Multiple Object Tracking in Mobile Devices. *IEEE Internet Things J.* **2023**, *10*, 4735–4748. [CrossRef]
7. Wang, S.; Sheng, H.; Yang, D.; Zhang, Y.; Wu, Y.; Wang, S. Extendable Multiple Nodes Recurrent Tracking Framework with RTU++. *IEEE Trans. Image Process.* **2022**, *31*, 5257–5271. [CrossRef]
8. Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; Niebles, J.C. Dense-Captioning Events in Videos. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
9. Mun, J.; Yang, L.; Ren, Z.; Xu, N.; Han, B. Streamlined Dense Video Captioning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
10. Wang, J.; Jiang, W.; Ma, L.; Liu, W.; Xu, Y. Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018. [CrossRef]
11. Li, Y.; Yao, T.; Pan, Y.; Chao, H.; Mei, T. Jointly Localizing and Describing Events for Dense Video Captioning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018. [CrossRef]

12. Zhang, W.; Ke, W.; Yang, D.; Sheng, H.; Xiong, Z. Light field super-resolution using complementary-view feature attention. *Comput. Vis. Media* **2023**, *9*, 843–858. [CrossRef]
13. Choi, W.; Chen, J.; Yoon, J. Parallel Pathway Dense Video Captioning With Deformable Transformer. *IEEE Access* **2022**, *10*, 129899–129910. [CrossRef]
14. Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; Saenko, K. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, CO, USA, 31 May–5 June 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015. [CrossRef]
15. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
16. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on attention for image captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4634–4643. [CrossRef]
17. Huang, X.; Ke, W.; Sheng, H. Enhancing Efficiency and Quality of Image Caption Generation with CARU. In *Wireless Algorithms, Systems, and Applications*; Springer Nature: Cham, Switzerland, 2022; pp. 450–459. [CrossRef]
18. Wang, S.; Yang, D.; Wu, Y.; Liu, Y.; Sheng, H. Tracking Game: Self-adaptative Agent based Multi-object Tracking. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; ACM: New York, NY, USA, 2022. [CrossRef]
19. Caspi, Y.; Simakov, D.; Irani, M. Feature-Based Sequence-to-Sequence Matching. *Int. J. Comput. Vis.* **2006**, *68*, 53–64. [CrossRef]
20. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Chan, K.H.; Im, S.K.; Ke, W. VGGreNet: A Light-Weight VGGNet with Reused Convolutional Set. In Proceedings of the 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC), Leicester, UK, 7–10 December 2020; IEEE: Piscataway, NJ, USA, 2020. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016. [CrossRef]
23. Zhao, B.; Li, X.; Lu, X. CAM-RNN: Co-Attention Model Based RNN for Video Captioning. *IEEE Trans. Image Process.* **2019**, *28*, 5552–5565. . [CrossRef]
24. Sawarn, A.; Srivastava, S.; Gupta, M.; Srivastava, S. BeamAtt: Generating Medical Diagnosis from Chest X-Rays Using Sampling-Based Intelligence. In *EAI/Springer Innovations in Communication and Computing*; Springer International Publishing: Cham, Switzerland, 2021; pp. 135–150. [CrossRef]
25. Deng, J.; Li, L.; Zhang, B.; Wang, S.; Zha, Z.; Huang, Q. Syntax-Guided Hierarchical Attention Network for Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 880–892. [CrossRef]
26. Xiong, Y.; Dai, B.; Lin, D. Move Forward and Tell: A Progressive Generator of Video Descriptions. In *Computer Vision—ECCV 2018*; Springer International Publishing: Cham, Switzerland, 2018; pp. 489–505. [CrossRef]
27. Rahman, T.; Xu, B.; Sigal, L. Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019. [CrossRef]
28. Rafiq, G.; Rafiq, M.; Choi, G.S. Video description: A comprehensive survey of deep learning approaches. *Artif. Intell. Rev.* **2023**. [CrossRef]
29. Shen, Z.; Li, J.; Su, Z.; Li, M.; Chen, Y.; Jiang, Y.G.; Xue, X. Weakly Supervised Dense Video Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–16 July 2017; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
30. Huang, X.; Chan, K.H.; Wu, W.; Sheng, H.; Ke, W. Fusion of Multi-Modal Features to Enhance Dense Video Caption. *Sensors* **2023**, *23*, 5565. . [CrossRef]
31. Wang, T.; Zheng, H.; Yu, M.; Tian, Q.; Hu, H. Event-Centric Hierarchical Representation for Dense Video Captioning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1890–1900. [CrossRef]
32. Zeng, R.; Xu, H.; Huang, W.; Chen, P.; Tan, M.; Gan, C. Dense Regression Network for Video Grounding. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; IEEE: Piscataway, NJ, USA, 2020. [CrossRef]
33. Li, K.; Guo, D.; Wang, M. Proposal-Free Video Grounding with Contextual Pyramid Network. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1902–1910. [CrossRef]
34. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–16 July 2017; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
35. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
36. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

37. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
39. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229. [CrossRef]
40. Park, J.S.; Darrell, T.; Rohrbach, A. Identity-Aware Multi-sentence Video Description. In *Computer Vision—ECCV 2020*; Springer International Publishing: Cham, Switzerland, 2020; pp. 360–378. [CrossRef]
41. Pasunuru, R.; Bansal, M. Multi-Task Video Captioning with Video and Entailment Generation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; Volume 1. [CrossRef]
42. Shetty, R.; Laaksonen, J. Frame- and Segment-Level Features and Candidate Pool Evaluation for Video Caption Generation. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; ACM: New York, NY, USA, 2016. [CrossRef]
43. Zhou, L.; Zhou, Y.; Corso, J.J.; Socher, R.; Xiong, C. End-to-End Dense Video Captioning with Masked Transformer. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018. [CrossRef]
44. Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; Courville, A. Describing Videos by Exploiting Temporal Structure. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: Piscataway, NJ, USA, 2015. [CrossRef]
45. Neubeck, A.; Gool, L.V. Efficient Non-Maximum Suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA, 2006. [CrossRef]
46. Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; Luo, P. End-to-End Dense Video Captioning with Parallel Decoding. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; IEEE: Piscataway, NJ, USA, 2021. [CrossRef]
47. Li, B.; Zhang, W.; Tian, M.; Zhai, G.; Wang, X. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5944–5958. [CrossRef]
48. Zhang, W.; Ma, C.; Wu, Q.; Yang, X. Language-guided navigation via cross-modal grounding and alternate adversarial learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 3469–3481. [CrossRef]
49. Hao, W.; Zhang, Z.; Guan, H. Integrating Both Visual and Audio Cues for Enhanced Video Caption. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [CrossRef]
50. Jin, Q.; Chen, J.; Chen, S.; Xiong, Y.; Hauptmann, A. Describing Videos using Multi-modal Fusion. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; ACM: New York, NY, USA, 2016. [CrossRef]
51. Chang, Z.; Zhao, D.; Chen, H.; Li, J.; Liu, P. Event-centric multi-modal fusion method for dense video captioning. *Neural Netw.* **2022**, *146*, 120–129. [CrossRef]
52. Iashin, V.; Rahtu, E. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer. *arXiv* **2020**, arXiv:2005.08271.
53. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
54. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703.
55. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
56. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02, Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002. [CrossRef]
57. Lavie, A.; Denkowski, M.J. The Meteor metric for automatic evaluation of machine translation. *Mach. Transl.* **2009**, *23*, 105–115. [CrossRef]
58. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; IEEE: Piscataway, NJ, USA, 2015. [CrossRef]