This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

PWS-DVC: Enhancing Weakly Supervised Dense Video Captioning with Pretraining Approach

WANGYU CHOI[1], JIASI CHEN[2], (Member, IEEE), and JONGWON YOON[1], (Member, IEEE)
[1]Department of Computer Science and Engineering (Major in Bio Artificial Intelligence), Hanyang University, Ansan 15588, South Korea
[2]Department of Electrical Engineering and Computer Science, University of Michigan, MI 48104, USA

Corresponding author: Jongwon Yoon (e-mail: jongwon@hanyang.ac.kr).

This work was supported by Basic Science Research Program through the National Research Foundation of South Korea (NRF) funded by the Ministry of Education NRF-2022R1A2C1008743 and the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2023-2020-0-01741) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

ABSTRACT In recent times, there has been a notable increase in efforts to simultaneously comprehend vision and language, driven by the availability of video-related datasets and advancements in language models within the domain of natural language processing. Dense video captioning poses a significant challenge in understanding untrimmed video and generating several event-based sentences to describe the video. Numerous endeavors have been undertaken to enhance the efficacy of the dense video captioning task by the utilization of various approaches, such as bottom-up, top-down, parallel pipeline, pretraining, etc. In contrast, the weakly supervised dense video captioning method presents a highly promising strategy for generating dense video captions solely based on captions, without relying on any knowledge of ground-truth events, which distinguishes it from widely employed approaches. Nevertheless, this approach has a drawback that inadequate captions might hurt both event localization and captioning. This paper introduces PWS-DVC, a novel approach aimed at enhancing the performance of weakly supervised dense video captioning. PWS-DVC's event captioning module is initially trained on video-clip datasets, which are extensively accessible video datasets by leveraging the absence of ground-truth data during training. Subsequently, it undergoes fine-tuning specifically for dense video captioning. In order to demonstrate the efficacy of PWS-DVC, we conduct comparative experiments with state-of-the-art methods using the ActivityNet Captions dataset. The findings indicate that PWS-DVC exhibits improved performance in comparison to current approaches in weakly supervised dense video captioning.

INDEX TERMS Weakly supervised, Pretraining, Dense video captioning.

I. INTRODUCTION

Recently, there has been a notable increase in efforts [1]–[6] to understand the correlation between video and language within the domain of computer vision. Attention towards tasks like video captioning [7]–[11] and temporal action localization, which aim to gain an in-depth comprehension of videos, has significantly increased due to the availability of extensive video-related datasets [12]–[16]. Dense video captioning [17] is an advanced task that integrates video captioning and temporal action localization. Its objective is to identify multiple temporal bins that encompass significant scenes within a video, followed by providing informative descriptions for each bin in the form of a single sentence. This task holds high potential for application in a variety of video analytics tasks.

Dense video captioning presents significant challenges, especially when encoding lengthy videos without trimming them. While several effective approaches have been developed for video understanding, they mainly focus on short videos lasting between 10 to 30 seconds, typically centered around a single topic. However, dense video captioning is specifically designed for videos that exceed a duration of 120 seconds. These longer videos contain a multitude of activities and objects, requiring a comprehensive and detailed representation for encoding. The outputs of dense video captioning, which include events and captions, are complex. In untrimmed videos, multiple events can occur simultaneously and temporally overlap with each other. Therefore, it is crucial to consider both the quality and coherence of the captions.

In addition to challenges from an input and output perspective, dense video captioning task has another non-trivial challenge of a nondeterministic nature, wherein individuals may

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

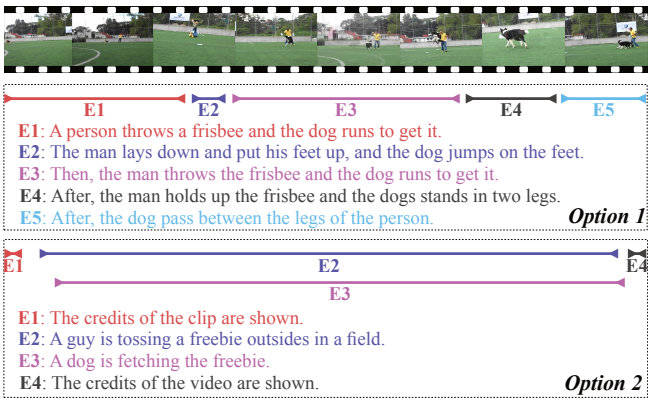


FIGURE 1. Example of dense video captioning for a video extracted from the ActivityNet Captions dataset. Note the different definitions of events and captions for the same video.

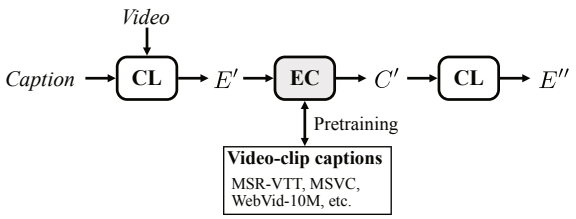


FIGURE 2. The approach of the proposed method. Our technique primarily focuses on leveraging the absence of ground-truth data in weakly supervised dense video captioning. This is achieved by pretraining the event captioning module using video clip-level datasets. CL and EC stand for caption localization and event captioning, respectively.

define events and captions for videos in varying ways. The provided visual representation, depicted in Figure 1, serves as an illustrative example of a sample extracted from the ActivityNet Captions [17] dataset. The video exhibits two distinct instances of ground-truth events along with corresponding captions. The presence of non-determinism in the model contributes to an elevated level of uncertainty [18], thereby impeding the generation of descriptive output. Increasing the number of training samples would effectively address the existing uncertainty. However, it is important to note that the availability of dense video captioning-specific datasets such as ActivityNet Captions [17] and YouCook2 [15], which encompass events and corresponding captions for videos, is currently limited. There exists a conspicuous dearth of temporal action region datasets in comparison to text corpora that are pertinent to captioning. The process of creating supplementary datasets entails substantial investments in terms of both time and financial resources.

Numerous attempts [17], [19]–[34] have been made to tackle the issue of dense video captioning, but it is important to note that, to the best of our knowledge, all existing approaches take a deterministic approach and mainly focus on restricted datasets. For example, some methods [17], [29]–[31], [34] divide the problem of event detection into two subtasks: event localization and event captioning. They use box prediction and regression techniques to address event localization. Some approaches [24], [32], [33] have improved their performance by developing parallel pipelines that consider the interdependence of the two subtasks. However, these models are still trained deterministically using datasets specifically designed for restricted dense video captioning. Recently, pretraining methods [23] have been proposed that use narrated video datasets based on Auto Speech Recognition (ASR) technology. These methods aim to overcome the constraints of dense video captioning datasets. However, these datasets lack faithfully defined events and captions, and their effectiveness is limited for videos with little or no speech.

On the other hand, in addition to conventional approaches, some weakly supervised dense video captioning approaches [35]–[37] have been proposed. For instance, Duan et al. [35] introduced a weakly supervised dense video captioning method. Rahman et al. [36] proposed a multi-modal approach to improve the performance of weakly supervised dense video captioning. Chen et al. [37] recently introduced a bridging network to enhance the connection between the event captioning module and sentence localization. One advantage of these methods is that they do not require explicit event supervision. However, since they directly localize captioning to the video, high-quality descriptive captions are required. Inadequate captions result in a lack of localization, leading to non-descriptive outcomes with limited occurrences. To enhance model performance, it is widely acknowledged that language models benefit from extensive datasets and large-scale models [38]. For the captioning module, a substantial corpus of textual data is necessary. However, existing methods are limited by their reliance on the designated dataset specifically designed for dense video captioning.

To briefly summarize the challenges faced by dense video captioning as mentioned above, they are as follows (i) The datasets available for training dense video captioning are quite limited. To effectively handle the task, a substantial dataset comprising extensive events and captions is necessary. Creating a dataset manually annotated by humans from untrimmed video requires substantial effort. (ii) The weakly supervised learning method is appealing as it does not rely on ground truth events, but it is limited by its dependence on the performance of the captioning module. Therefore, meticulous attention must be given to the design and training of the captioning module.

We present several important observations to address these challenges. Previous works have demonstrated that pretraining techniques result in significant improvements in the performance of dense video captioning methods. These techniques leverage imperfect transcribed text datasets to overcome the limited availability of event-caption datasets. We have access to a substantial amount of high-quality datasets for video-clip captioning. However, their drawback is the absence of event information, which renders them unsuitable for direct use in dense video captioning. Fortunately, weakly supervised dense video captioning does not require event supervision, enabling extensive utilization of video-clip datasets.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

This work introduces PWS-DVC, a method for dense video captioning that utilizes weak supervision and can be trained on video-clip-level description datasets. The PWS-DVC system comprises two modules: event captioning and caption localization. The event captioning module is pretrained using readily-available datasets [4], [39]–[42]. We then fine-tune the event captioning module in tandem with the caption localization module through an iterative process of fine-tuning, utilizing weakly supervised learning in a downstream task. Our approach enables a robust event captioning module that is based on pretraining, resulting in the generation of detailed and descriptive captions.

As shown in Figure 2, utilizing the pretraining approach in our event captioning module enhances both the quality of captions and event localization. Our proposed weakly supervised dense video captioning strategy differs from the conventional approach by replacing the two modules of event localization and event captioning with the modules of event captioning and caption localization. In the conventional technique, the improvement of the captioning module is not affected by the localization process. However, in the weakly supervised method, the effectiveness of the captioning module has a direct impact on the localization outcome. Pretraining the captioning module is advantageous for both conventional and weakly supervised approaches, but it has a stronger synergy with the weakly supervised method.

To demonstrate the effectiveness of PWS-DVC, we first train the event captioning module on MSR-VTT [42], a commonly used dataset for generating captions for video clips. By utilizing the event captioning module, PWS-DVC outperforms other video-clip captioning methods. We then fine-tune PWS-DVC on the ActivityNet Captions [17] dataset, training the event captioning and caption localization modules in an alternating manner. This approach results in superior performance in both captioning and localization tasks compared to other weakly supervised learning methods.

The remaining sections of this paper are structured as follows. Section II provides an overview of the existing literature on video-clip captioning and dense video captioning, highlighting the distinctions between our proposed model, PWS-DVC, and existing approaches. In Section III, we provide a comprehensive explanation of the technique employed in PWS-DVC. Section IV demonstrates the efficiency of PWS-DVC through a series of comprehensive tests. Finally, we conclude the paper in Section V.

The main contributions of PWS-DVC can be summarized as follows:

- We introduce a novel pretraining method for weakly dense video captioning, which is the first of its kind to the best of our knowledge.
- Our approach directly utilizes multiple accessible video-clip datasets for dense video captioning, demonstrating its effectiveness.
- The PWS-DVC model outperforms existing weakly supervised dense video captioning methods.

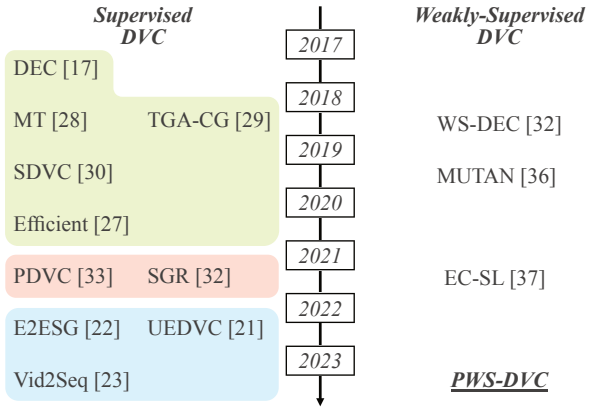


FIGURE 3. History of dense video captioning. There are two main categories of approaches in the field of dense video captioning: supervised approaches and poorly supervised approaches.

II. RELATED WORK

A. VIDEO CLIP CAPTIONING

The objective of video clip captioning is to describe a video using a single sentence. Recent approaches mainly follow the encoder-decoder structure [8], [43]–[47]. The encoder usually consists of a Convolutional Neural Network (CNN) followed by a Recurrent Neural Network (RNN). On the other hand, the decoder is an RNN that generates predictions for individual tokens throughout each decoding step. However, videos often contain informative content and multiple events that cannot be captured in a short sentence. Therefore, recent methods [48]–[50] have employed the approach of generating a comprehensive and elaborate paragraph to provide a detailed description of the video. In this approach, each sentence within the paragraph describes a specific event or occurrence depicted in the video.

B. DENSE VIDEO CAPTIONING

Dense video captioning involves describing untrimmed videos through the use of events and captions. Numerous efforts [17], [19]–[37] have been made to improve the efficacy of dense video captioning using various approaches. Based on the training approach, existing dense video captioning methods can be categorized into two groups: supervised and weakly supervised dense video captioning, as shown in Figure 3. Supervised dense video captioning approaches require strict supervision, which includes the availability of ground truth data for both events and captions. Specifically, these methods rely on datasets that have undergone complex human annotation processes for videos, such as ActivityNet Captions. However, supervised methods have emerged as the mainstream approach due to their comprehensive supervision. In contrast, weakly supervised dense video captioning techniques present a promising avenue as they do not require precise knowledge of events. Similarly, weakly supervised dense video captioning is a promising approach as it does not require ground-truth information regarding events. However, this approach has limitations due to the demand for high-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

quality language models and the complexity of the training process.

Supervised approach. Supervised dense video captioning can be categorized into three distinct classifications based on the architecture employed, each represented by different color boxes. The methods contained within the light green box address the two distinct subtasks of event localization and captioning as separate entities. The methods within the light red box take into account the interdependence between these two subtasks and simultaneously generate events and captions. The techniques employed within the light blue box utilize a pretraining approach on large-scale extra datasets. The authors additionally take into consideration the event and caption set as a sequential entity in an auto-regressive fashion.

Certain approaches [17], [27], [29]–[31] aim to address the subtasks of localization and captioning as distinct subtasks. The DEC [31] method is a pioneering approach in the field of dense video captioning, as it employs Long Short-Term Memory (LSTM) networks for both the tasks of localization and caption generation. The TGA-CG [29] model utilizes bidirectional Long Short-Term Memory (LSTM) networks in order to effectively leverage both preceding and subsequent contextual information. MT [28] employs a transformer architecture and incorporates masked memory in order to generate captions. The proposed approach by SDVC [30] involves the utilization of a network for event sequence creation that is composed of two LSTM layers. This design choice is made in order to effectively account for the temporal dependency that exists between events. The proposed strategy by Efficient [27] aims to dynamically choose frames that contain relevant information in order to minimize computing complexity. All of the aforementioned methodologies yield substantial enhancements in performance, although with limits stemming from the neglect of the interplay between events and captions.

To explore the correlation between the two subtasks, several approaches [32], [33] have been developed to execute event localization and captioning concurrently. PDVC [33]'s proposed approach involves using a deformable masked transformer model to create contextual representations for both events and captions. The model then performs separate tasks of localization and captioning based on each context. SGR [32]'s approach, on the other hand, is top-down. It first generates a paragraph for the video content and then links sentences to specific temporal segments. Ultimately, the last step refines the quality and precision of each sentence.

Recently, there have been suggestions for pretraining techniques to improve performance in event and caption generation, as the restricted availability of datasets with ground-truth information hinders further advancements. The E2ESG [22] model utilizes a pretrained T5 [51] model on the WikiHow [52] dataset, which consists of instructional articles, as a captioning module. The UEDVC [21] model pretrains on the ActivityNet Captions dataset to learn about the interplay between video and language. Vid2Seq [23] employs narrated video datasets [14] with transcribed text from an Automatic Speech Recognition (ASR) system to capture the relationship between video content and language. These pretraining approaches use an architectural framework that treats generating events and captions as a sequence generation task while accounting for their inherent relationship.

Weakly supervised approach. In addition to supervised dense video captioning, weakly supervised dense video captioning methods [35], [37] have emerged, which adopt event captioning and caption localization modules instead of event localization and captioning modules. These approaches use event captioning and caption localization modules instead of event localization and captioning modules. The two modules are then trained in an alternating manner, offering the advantage of not requiring prior knowledge of occurrences. WS-DEC [35] focuses on weakly supervised dense video captioning, achieving comparable performance to supervised learning approaches despite the absence of event-specific supervision. The subsequent approach, EC-SL [15], introduces an induced set attention block module to enhance the consideration of the interplay between captioning and localization.

C. POSITION OF PWS-DVC

The proposed model, PWS-DVC, aims to improve dense video captioning by enhancing the event captioning and language modules. Although weakly supervised dense video captioning has the promising feature of not requiring event ground truth, it has the limitation that insufficient sentences generated by the captioning module are difficult to accurately localize. To address this, we generate a robust language model by pretraining the captioning module and fine-tuning it on the ActivityNet Captions dataset. The approach of pretraining the captioning module leads to more accurate localization by ensuring that it produces clear sentences about the video. Fortunately, the event captioning module can be trained on annotations of video clips and sentences. Therefore, we take advantage of this and pretrain it on MSR-VTT [42], a dataset of video captioning at the video-clip level. We present the training method of PWS-DVC and its position in terms of architecture in Figure 3.

III. METHOD

A. OVERVIEW

Our proposed method, PWS-DVC, addresses the constraint of existing dense video captioning methods, which rely on specialized datasets for dense video captioning, such as ActivityNet Captions [17]. Instead, we leverage weakly supervised dense video captioning to its maximum capacity by pretraining on readily accessible datasets for video clip-level captioning. As depicted in Figure 4, PWS-DVC comprises two distinct modules, namely the caption localization module and the event captioning module. The module for captioning localization is responsible for determining the specific temporal interval in a video where a particular caption is positioned. On the other hand, the event captioning module is designed to construct a phrase that describes a video clip based on the provided video and event. To address the difficulty of dense video captioning, we employ a two-step approach. Firstly,

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

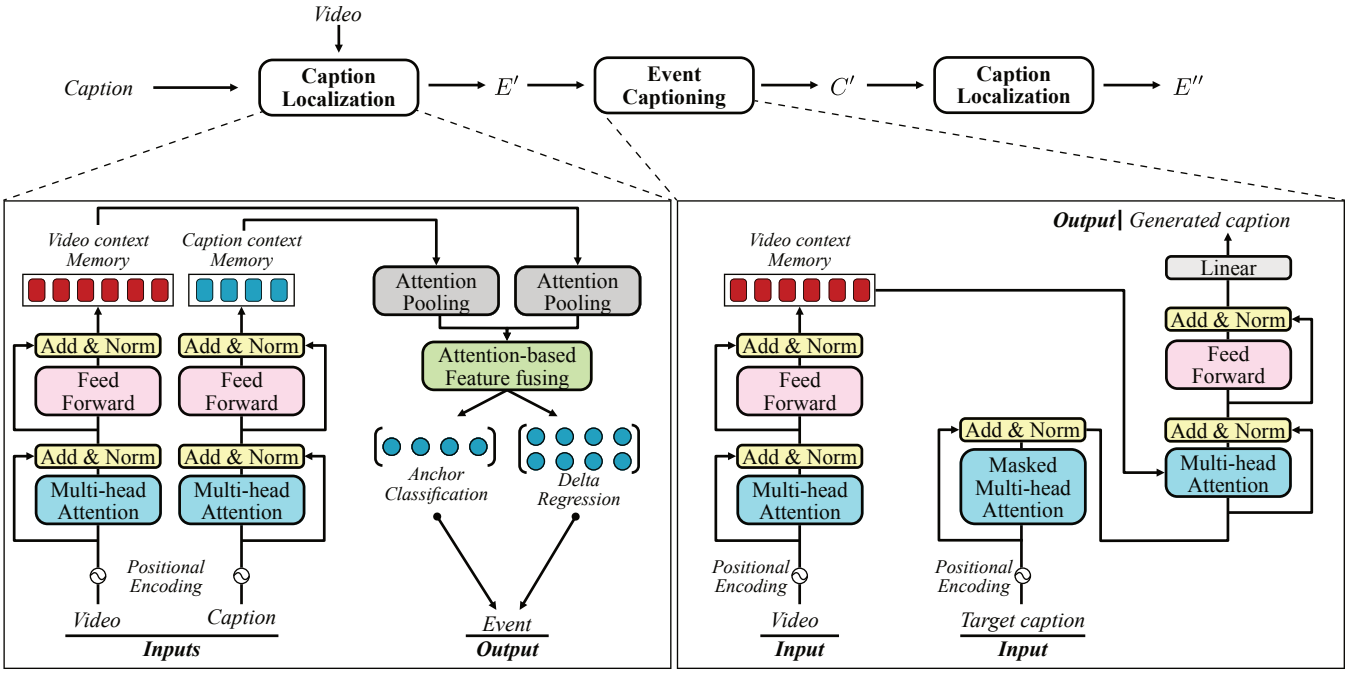


FIGURE 4. Network architecture of the proposed model, PWS-DVC. PWS-DVC consists of two modules: caption localization and event captioning. The caption localization module takes a video and a caption as input and aims to localize which temporal region the caption corresponds to in the video. The event captioning module aims to generate a sentence describing the event given a video and its event. Both modules enable dense video captioning by alternatively training with video and ground-truth captions without supervision on events.

we pretrain the event captioning module using video clip-level datasets. Subsequently, the two modules are trained in an alternating manner.

B. CAPTION LOCALIZATION MODULE

The goal of the sentence localization module is to accurately determine the temporal position of a sentence generated by the event captioning module, which is based on a specific segment of a video. To achieve this objective, the event localization module analyzes the correlation between video frames and captions. This task is accomplished by leveraging the temporal correlation of representations within the videos.

We adopt transformer encoder [53] to encode both video and caption, enabling exploration of the temporal association between these two modalities through their respective representations, which is proven highly effective in capturing sequential data. The process involves sequentially encoding video and caption using transformer encoders, resulting in the generation of video context memory and caption context memory, respectively. The input is then subjected to an attention-pooling layer to rectify its temporal structure. Finally, video and caption context memory integration is achieved by employing an attention network, allowing for the examination of video representations that align with the given fixed caption temporally.

We formulate this process as follows. In the context of the transformer model, the initial step involves the application of positional encoding to both a sequence of video frames and a sequence of words inside a sentence, as illustrated in Figure 4.

$$H_v^0 = \mathrm{PE}(\mathrm{Linear}\,(F)) \tag{1}$$

The positional encoding function, denoted as $\mathrm{PE}(\cdot)$, is applied to the frame-level video feature F, which represents the input video v or the embedding vector of the input caption. The frame-level video feature is based on the C3D model proposed by [54].

The video and caption encoding components are based on the transformer encoder, which comprises a multiple of N layers. The composition of each layer consists of self-attention and feed-forward modules, with a residual connection and layer normalization appended after each module. The output H^{l+1} of the transformer encoder is obtained in a single cycle, using the input H^l. The final output of the last cycle, which corresponds to the last layer, is denoted as H. The subsequent pair of equations can be employed to delineate encoding based on transformers.

$$H_{self}^l = \Psi\left(H^l + \mathrm{MA}\left(H^l, H^l, H^l\right)\right) \tag{2}$$

$$H^{l+1} = \Psi\left(\mathrm{FFN}\left(H_{self}^l\right) + H_{self}^l\right) \tag{3}$$

$$\mathrm{FFN}(x) = \max\left(0, xW_1 + b_1\right)W_2 + b_2 \tag{4}$$

$$\mathrm{MA}(Q, K, V) = \mathrm{Cat}(\{\mathrm{ATT}_i(Q, K, V)\}_{i=1}^h)W_O \tag{5}$$

$$\mathrm{ATT}(Q, K, V) = \mathrm{softmax}(\frac{(QW_Q)(KW_K)^T}{\sqrt{d_m}})(VW_V) \tag{6}$$

where, $\Psi(\cdot)$ represents the layer normalization function. $\mathrm{Cat}(\cdot)$ denotes vector concatenation and W_* are trainable

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

Algorithm 1 Training procedure of PWS-DVC.

Input: $\mathcal{D}_{\text{MSR-VTT}}, \mathcal{D}_{\text{ANetC}}$ // datasets used for training
Input: $\mathcal{C}, \mathcal{G}$ // event captioning and caption localization modules
Input: $\left\{E_j^{(a)}\right\}_{j=0}^{N_a}$ // anchors for training $\mathcal{G}$
Input: $E^{(0.5,1)}$ // event that covers the entire video.
Output: $\theta_{\mathcal{C}}, \theta_{\mathcal{G}}$ // trained parameters

Pretrain procedure
for $(V, \{C_i\}_{i=0}^{N_v}) \in \mathcal{D}_{\text{MSR-VTT}}$ **do**
  $C = ChooseOne(\{C_i\}_{i=0}^{N_v})$
  $C' = \mathcal{C}(V, E^{(0.5,1)})$
  $\mathcal{L}_c = CrossEntropy(C, C')$
  $\theta_{\mathcal{C}} = \theta_{\mathcal{C}} - \eta_{\mathcal{C}} \frac{\partial \mathcal{L}_c}{\partial \theta_{\mathcal{C}}}$
end for
Fine-tuning procedure
for $(V, \{C_i\}_{i=0}^{N_v}) \in \mathcal{D}_{\text{ANetC}}$ **do**
  $C = ChooseOne(\{C_i\}_{i=0}^{N_v})$
  $E', \text{confidence} = \mathcal{G}(V, C + \delta)$
  $C' = \mathcal{C}(V, E')$
  $E'' = \mathcal{G}(V, C')$
  $\{C_j\}_{j=0}^{N_a} = \{\mathcal{G}(V, E_j)\}_{j=0}^{N_a}$
  $\{\text{score}_j\}_{j=0}^{N_a} = \{\text{METEOR}(C, C_j)\}_{j=0}^{N_a}$
  $\text{best} = \text{argmax}_j\{\text{score}_j\}_{j=0}^{N_a}$
  $\mathcal{L}_c = CE(C, C')$
  $\theta_{\mathcal{C}} = \theta_{\mathcal{C}} - \eta_{\mathcal{C}} \frac{\partial \mathcal{L}_c}{\partial \theta_{\mathcal{C}}}$
  $\mathcal{L}_l = MSE(E', E'') + CE(\text{confidence, best})$
  $\theta_{\mathcal{G}} = \theta_{\mathcal{G}} - \eta_{\mathcal{G}} \frac{\partial \mathcal{L}_l}{\partial \theta_{\mathcal{G}}}$
end for
return $\theta_{\mathcal{C}}, \theta_{\mathcal{G}}$

parameters. We repeat the above process as many times as the number of encoder layers. The output of the video encoder is the output of the last layer H_v and H_c, for the video and the caption, respectively.

The attention-based feature fusing process is a mechanism used to align and fuse video context memory and caption context memory in order to determine the temporal region in a video that corresponds to the caption. This process involves several steps. To start, we have two sets of context memories: the video context memory represented as H_v and the caption context memory represented as H_c. These context memories contain information about the video and caption, respectively.

The first step in attention-based feature fusing is to achieve temporal dimension matching between these two sets of memories. This means that we want to make sure that both memories have the same dimensions and can be effectively compared. To achieve this matching, an attention pooling layer is employed. This layer computes attention weights for each element in the memories, effectively giving importance scores to each element.

The formula used to calculate the attention weights is given as follows:

$$a_i = \frac{e^{w \cdot h_i}}{\sum_{j=1}^{K} e^{w \cdot h_j}} \tag{7}$$

where, a_i represents the attention weight assigned to each element of memory H, denoted as h_i. w represents the attention weights, which are learnable parameters in the model. K represents the number of dimensions in the output vector. This value is determined by the hidden dimension of the transformer hyperparameter (K corresponds to the hidden dimension). Once we have computed these attention weights, we proceed to fuse the memories. The pooled memory, denoted as H_{pooled}, is calculated by taking a weighted sum of the elements in the memory, where the weights are determined by the attention scores:

$$H_{pooled} = \sum_{i=1}^{N} a_i \cdot h_i \tag{8}$$

where, H_{pooled} represents the fused memory that combines information from both the video and caption context memories. Each element of this pooled memory is a weighted combination of the corresponding elements in the original memories, with the weights determined by the attention scores.

Now, with the pooled video context memory $H_{v,\text{pooled}}$ and caption context memory $H_{c,\text{pooled}}$, we have effectively aligned and fused the information from both modalities. The final step in attention-based feature fusing is to compute the importance of the video context memory to the caption context memory. This importance score helps identify the temporal region in the video that corresponds to the caption, allowing the model to focus on relevant video features for localization.

The last step in caption localization involves localizing the temporal region in the video that corresponds to the caption. This is done using anchor classification and delta regression techniques. The use of simple regression, as employed in the existing methods, may aid in achieving this task. However, it should be noted that this approach is prone to encountering local minima. To address this issue, the localization process is divided into two distinct steps: classification and refinement. The classification step aims to select the most appropriate one among several predefined anchors, and the refinement step aims at regression to compensate for the selected anchor.

To do this, we adopt two fully connected layers, FC_{cls} and FC_{reg}. We first identify the best anchor from a set of predefined anchors $S = \{e_1, e_2, \cdots, e_{N_a - 1}\}$.

$$\{o_i\}_{i=0}^{N_a} = \{\text{FC}_{\text{cls}}(e_i)\}_{i=0}^{N_r} \tag{9}$$

where o_i means the classification logit for the i-th anchor. Then we formulate the index of best anchor $i^{(\text{best})}$ as follows.

$$i^{(\text{best})} = \arg\max_i \{\text{MeteorScore}(o_i)\}_{i=0}^{N_a} \tag{10}$$

where, the $\text{MeteorScore}(\cdot)$ function calculates the METEOR [72] score of the caption generated for the anchor. We regress Δw and Δm to refine the final best anchor.

$$\Delta w, \Delta m = \text{FC}_{\text{reg}}(e_{i^{(\text{best})}}) \tag{11}$$

We formulate the anchor that is finally selected, the anchor with the best score is $e^{(\text{best})} = (w_{i^{(\text{best})}} + \Delta w, m_{i^{(\text{best})}} + \Delta m)$. $w_{i^{(\text{best})}}$ and $m_{i^{(\text{best})}}$ are the width and center for the best anchor, respectively.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

C. EVENT CAPTIONING MODULE

The objective of the event captioning module is to provide a concise and informative sentence that describes an event, based on a video and its corresponding temporal region. To achieve this, we follow the established framework of the vanilla transformer [53]. Specifically, we use a two-step process: first, we take a whole video and an associated event as inputs; second, we apply a mask to the temporal region of the video feature that corresponds to the event, and feed it into the transformer encoder. The last step involves allowing the visual context memory output, derived from the transformer encoder, to undergo cross-attention with the transformer's decoder. This process facilitates the generation of the relevant caption.

We apply an event mask to the video and let H_v be the feature of the video after applying positional encoding, and the above process can be expressed as the following equations.

$$H^l_{self} = \Psi\left(H^l_v + \mathrm{MA}\left(H^l_v, H^l_v, H^l_v\right)\right) \tag{12}$$

We denote the output of the last encoder layer, which is the video context memory, as H_v, the decoding process of the transformer be represented as follows.

$$H^l_{cross} = \Psi\left(H_v + \mathrm{MA}\left(H^l_{self}, H_v, H_v\right)\right) \tag{13}$$

$$H^{l+1}_{cross} = \Psi\left(H^l_{cross} + \mathrm{FFN}\left(H^l_{cross}\right)\right) \tag{14}$$

We repeat the aforementioned process, where H^{l+1}_{cross} is set as H_c for the final layer. H_c is the output of the transformer decoder, which classifies a word from one of the vocabulary as the last to pass through the linear transformation as shown in Figure 4.

D. TRAINING PROCESS

As described in Algorithm 1, the PWS-DVC training process involves two steps: pretraining and fine-tuning. During pretraining, the focus is on the event captioning module. Developing a solid language model can be a challenging task, as it requires a substantial text corpus. Therefore, we use a pretraining approach to generate descriptive sentences that describe the corresponding videos. This approach helps to achieve more accurate event localization. For pretraining, we use widely available video clip-level datasets such as MSR-VTT [42] and MSVD [41]. These datasets are ideal for pretraining the event captioning module because they are already trimmed to a short event level. To obtain video features, we pass a video through a C3D [54] backbone, resulting in C3D features in 16 non-overlapped frames of the video. We then pass these video features through the event captioning module to train with the following loss function.

$$\mathcal{L}_c = -\sum_{t=1}^{T_c} c_t \cdot \log\left(\hat{c}_t \mid c_0 : c_{t-1}\right) \tag{15}$$

where, $\hat{C} = \{\hat{c}_i\}_{i=0}^{T_c}$ means the ground-truth caption, and $C = \{c_i\}_{i=0}^{T_c}$ means the prediction output by the event captioning module.

Subsequently, we proceed with the process of fine-tuning using datasets such as the ActivityNet Captions dataset, which is specifically designed for dense video captioning tasks. The event captioning and caption localization modules are trained in an alternating manner. Although our method for weakly supervised dense video captioning has the potential to improve over multiple rounds of training, we focus on a single-round iteration, as depicted in Figure 4. The equation denoted as Equation (15) is employed for the event captioning module, and the following loss function is taken into account for caption localization.

$$\mathcal{L}_l = \left[(m' - m'')^2 + (w' - w'')^2\right] + \left[-\sum_{i=0}^{N_a} y_i \log p_i\right] \tag{16}$$

The first term is the L2 norm for regression on the center and width of the event, and the second term is the cross-entropy loss for anchor classification. Specifically, m' and w' refer to the center and length of E', and m'' and w'' refer to the centre and length of E'' in Figure 4. For classification, we define a one-hot label $y = \{y_1, \cdots, y_{N_a}\}$ for the predefined anchors. If y_i is 1, it indicates that the i-th anchor is the best, otherwise it is 0. $p = \{p_1, \cdots, p_{N_a}\}$ is the prediction of the caption localization module.

E. INFERENCE

Since the PWS-DVC model lacks a direct training mechanism for dense video captioning, we explain the inference procedure. To conduct inference on a video V, we begin by selecting a set of random event anchors denoted as $\{E^{(a)}_i\}^{N_r}_{i=0}$, where N_r is set to 15. Specifically, N_r random numbers are selected from the interval $[0, 1]$ to represent the width and center of the anchors. Next, we combine them to obtain anchors consisting of pairs of N_r widths and centers. After performing a single iteration for each event anchor, we obtain a set of refined events denoted as $\{S_i\}^{N_r}_{i=0}$. We filter the events based on their values with high temporal Intersection over Union (tIoU), retaining only those that meet the validity criteria. Finally, we input the filtered events into the event captioning module to provide captions for each event.

IV. EXPERIMENT

A. EXPERIMENTAL SETUP

Datasets. The MSR-VTT dataset [42] is used to pretrain the event captioning module. This comprehensive video benchmark dataset contains 10,000 web video clips, each accompanied by 20 annotations written by humans. A total of 200,000 clip-sentence pairings were used as samples for pretraining. Additionally, we employ temporal augmentation techniques to increase the number of samples from 200K to 1M.

For dense video captioning, we use the ActivityNet Captions [17] dataset, which comprises 20K untrimmed videos capturing various human activities. On average, videos have a duration of 120 seconds and contain an average of 3.7 events. While other datasets exist for dense video captioning, such as YouCook2 [15] and ViTT [16], we focus exclusively on the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

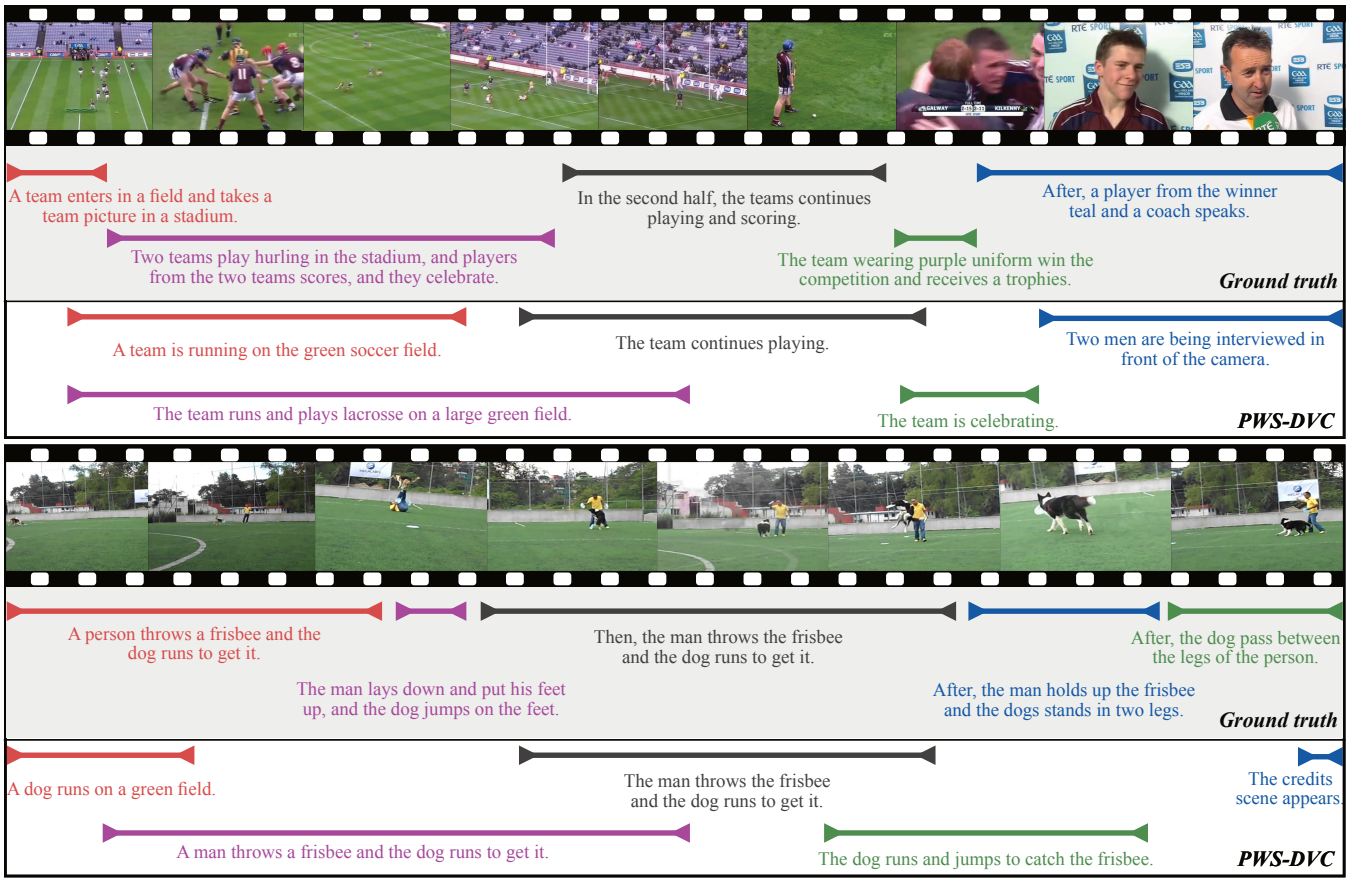


FIGURE 5. **Examples of the output of PWS-DVC for two videos from the ActivityNet Captions validation set.**

Method	WS	Extra dataset	METEOR	CIDEr	ROUGE	BLEU@1	BLEU@2	BLEU@3	BLEU@4	SODA_c
DCE [17]	✗	✗	4.82	17.29	-	17.95	7.69	3.86	2.20	
DVC [31]	✗	✗	6.93	12.61	-	12.22	5.72	2.27	0.73	
Bi-SST [29]	✗	✗	9.60	12.68	19.10	18.99	8.84	4.41	2.30	
MT [28]	✗	✗	9.56	-	-	-	-	4.76	2.23	
SDVC [30]	✗	✗	8.82	30.68	-	17.93	7.99	2.94	0.93	
PDVC [26]	✗	✗	7.50	47.26	2.64	-	-	-	1.65	5.26
E2ESG [21]	✗	✗	7.33	26.92	-	-	-	-	1.45	5.29
Vid2Seq [23]	✗	✗	8.50	30.10	-	-	-	-	-	5.80
MUTAN [36]	✓	✗	4.93	13.79	10.39	10.00	4.20	1.85	0.90	
WS-DEC [35]	✓	✗	6.30	18.77	12.55	12.41	5.50	2.62	1.27	
EC-SL [37]	✓	✗	7.49	21.21	13.02	**13.36**	5.96	2.78	1.33	
PWS-DVC	✓	✓	**8.22**	**28.91**	**13.80**	12.92	**6.38**	**3.11**	**1.87**	**5.45**

TABLE 1. **Dense event captioning performances (%) of the proposed method and state-of-the-art methods on the ActivityNet Captions validation set. "WS" denotes "weakly supervised".**

ActivityNet Captions dataset for the purpose of comparing with current weakly dense video captioning algorithms.

Evaluation metrics. The publicly accessible evaluation code supplied by the ActivityNet Captions Challenge is utilized in our study[1]. The evaluation metrics employed in this paper include recall and precision for event localization, as well as METEOR [72], CIDEr [73], and BLEU [74] for sentence evaluation. The computation of the captioning score for a generated event and sentence pair is contingent upon the presence of an overlap greater than the specified threshold with any ground-truth events. This score is determined by comparing the related sentence from the ground-truth data. Alternatively, if the conditions are not met, the score is assigned a value of zero.

We adopt SODA [75] as an additional evaluation metric. Recently, SODA has been introduced for evaluating multi-sentence captioning tasks, particularly in the context of dense video captioning. The SODA metric considers the coherence and narrative flow between sentences. It is derived from the METEOR metric, but penalizes the use of redundant words,

[1]https://github.com/ranjaykrishna/densevid_eval

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

Method	WS	Pretrain	tIoU=0.1	tIoU=0.3	tIoU=0.5
CTRL [55]	✗	✗	-	47.43	29.01
ABLR [56]	✗	✗	73.30	55.67	36.79
2D-TAN [57]	✗	✗	-	58.75	44.05
FIAN [58]	✓	✗	-	64.10	47.90
WSLLN [59]	✓	✗	75.40	42.80	22.70
LCGB [60]	✓	✗	74.20	44.30	23.60
SCN [35]	✓	✗	71.48	47.23	29.22
WS-DEC [35]	✓	✗	62.71	41.98	23.34
EC-SL [37]	✓	✗	68.48	44.29	24.16
PWS-DVC	✓	✓	70.01	51.94	27.70

TABLE 2. Temporal event localization performances of PWS-DVC and state-of-the-art methods on the ActivityNet Captions dataset. The top-1 recall is quantified by evaluating the overlap between the predicted temporal event and the ground truth across various temporal IoU thresholds. The abbreviation "WS" is used to represent the term "weakly supervised".

Method	BLEU@4	CIDEr	METEOR
FSTA [61]	39.80	41.1	26.50
OA-BTG [62]	41.40	46.9	28.20
MGSA [63]	42.40	47.5	27.60
POS+CG [64]	42.00	48.7	28.20
POS+VCT [65]	42.30	49.1	29.70
SAM-SS [66]	43.80	53.2	28.90
ORG-TRL [67]	43.60	50.9	28.80
STraNet [67]	43.60	50.9	28.80
VNS-GRU [68]	45.30	53.0	29.90
ARB [69] + ACL [70]	42.60	51.3	28.90
MFVCG [71]	44.90	52.2	29.80
PWS-DVC	41.20	49.3	26.99

TABLE 3. Comparison to state-of-the-art methods on MSR-VTT for video clip captioning.

phrases, and sentences. This feature enhances its ability to resist over-evaluation, as it discourages inflated scores through excessive repetition.

Implementation details. The event captioning module and caption localization module of PWS-DVC are based on the transformer architecture. We follow the hyper-parameters of the original transformer [53], including a model dimension of 512, 8 encoders and decoders, and a feed-forward dimension of 2048. We set the dropout rate to 0.1.

To train PWS-DVC, we use the AdamW [76] optimizer in both the pretraining and fine-tuning stages. In the pretraining stage, we gradually decrease the learning rate according to a cosine function with a warm-up period that increases up to 5e-4. In the fine-tuning stage, we train for 30 epochs using a batch size of 8 with a learning rate of 1e-4.

The examination of various hyperparameters' influence on the model is crucial in deep learning frameworks. Our extensive exploration indicates that PWS-DVC is robust to different hyperparameters, meaning that the improvement in performance with better hyperparameter settings is minimal. Instead of aiming for improved outcomes, we adhere to widely adopted settings to ensure fairness when comparing with baseline methods.

B. PERFORMANCE COMPARISON

We conduct extensive comparisons with existing state-of-the-art methods to demonstrate the effectiveness of PWS-DVC. Besides the dense video captioning benchmark dataset, ActivityNet Captions, we also compare the performance with video-clip captioning methods and compare the performance of caption localization with temporal action localization models.

Dense video captioning. Table 1 compares the dense video captioning performance on the ActivityNet Captions validation set. The results show that PWS-DVC achieves the best performance on weakly supervised dense video captioning methods for METEOR, CIDEr, ROUGE, and BLEU, and comparable performance on fully supervised methods. Specifically, PWS-DVC achieves state-of-the-art performance on almost all metrics compared to the weakly supervised approaches MUTAN, WS-DEC, and EC-SL. These results suggest that PWS-DVC's pretraining approach contributes to both localization and captioning by improving the performance of the event captioning module. Table 1 also compares PWS-DVC with several methods trained under weakly supervised as well as fully supervised conditions. PWS-DVC performs comparably to them despite being trained without information about ground-truth events.

Caption localization. We then evaluate the performance of caption localization on the validation set of ActivityNet Captions. We compare the weakly supervised dense video captioning methods WS-DEC and EC-SL with state-of-the-art models for temporal action localization tasks, such as CTRL and ABLR. Table 2 shows that the PWS-DVC approach achieves superior accuracy in caption localization compared to existing weakly supervised dense video captioning methods WS-DEC and EL-SL. Despite the fact that existing temporal action localization methods are trained on ground-truth events, the caption localization performance of PWS-DVC is comparable. These results demonstrate that PWS-DVC's approach is not only effective in improving the event captioning module, but also in improving caption localization.

Video clip captioning. The event captioning module of PWS-DVC generates a single sentence from a video clip, which is itself the purpose of video-clip captioning. We demonstrate the effectiveness of our event captioning module not only for dense video captioning, but also by comparing it with existing video clip captioning methods on the MSR-VTT video clip dataset. Table 3 shows the performance comparison of PWS-DVC's event captioning module with state-of-the-art video clip captioning methods. While video clip captioning methods have various advantages such as multimodal, pretraining, and object detection, PWS-DVC follows the architecture of a vanilla transformer in terms of model architecture and modality. Nevertheless, the event captioning module of PWS-DVC shows comparable performance to existing state-of-the-art video clip captioning methods.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

#	Method	METEOR	CIDEr	ROUGE	BLEU@4
1	PWS-DVC without pretraining	5.49	13.77	10.02	0.72
2	PWS-DVC pretrained on ActivityNet Captions	7.28	20.59	12.71	1.35
3	PWS-DVC pretrained on MSR-VTT	8.05	27.11	13.17	1.80
4	PWS-DVC pretrained on MSR-VTT with temporal augmentation	8.22	28.91	13.80	1.87

TABLE 4. The results of ablation study.

C. QUALITATIVE RESULTS

Figure 5 presents the qualitative outcomes of PWS-DVC. We select two videos from the validation set of ActivityNet Captions, and show their corresponding events and captions together. Through the examination of these two instances, it is demonstrated that PWS-DVC has the capability to effectively generate events and captions even in the absence of ground-truth events. Furthermore, PWS-DVC produces descriptive sentences by utilizing the event captioning module that has been pretrained on a video clip dataset. Additionally, it accurately identifies the location of each sentence inside the video scene. In the first video, PWS-DVC produces sentences that incorporate certain keywords, such as ''green field,'' ''running,'' and ''lacrosse,'' while also accurately identifying the precise location within the video scene for each sentence. This holds for the second video as well.

Note that the outcomes of PWS-DVC do not align with the ground-truth events and captions. However, this is a reasonable observation considering the inherent non-deterministic character of the dense video captioning task, as depicted in Figure 1.

D. ABLATION STUDY

To demonstrate the efficacy of pretraining the event captioning module, which serves as the core approach of PWS-DVC, we conduct an intensive ablation experiment involving different pretraining pretraining. We train four different variants of PWS-DVC, each with varying levels of pretraining: (i) PWS-DVC without pretraining, (ii) PWS-DVC pretrained using only the ActivityNet Captions dataset without any extra dataset, (iii) PWS-DVC pretrained on MSR-VTT but without temporal augmentation, and (iv) PWS-DVC pretrained on MSR-VTT with temporal augmentation. We present the performance of the four distinct PWS-DVC models on the ActivityNet Captions validation set in Table 4. It is evident that models that have undergone pretraining on ActivityNet Captions or MSR-VTT exhibit superior performance compared to model #1. This observation indicates that pretraining is highly effective in the context of weakly supervised dense video captioning approaches. Furthermore, it is evident that the incorporation of extra datasets, such as MSR-VTT, with the ActivityNet Captions dataset, yields enhanced performance outcomes. This is attributed to the augmented diversity of the training sample. By the same principle, the temporal augmentation serves to enhance the diversity of the training dataset, hence resulting in model #4 attaining the highest level of performance.

V. CONCLUSION

The paper presents a novel pretraining approach to enhance the performance of weakly supervised dense video captioning. Weakly supervised dense video captioning is promising as it does not rely on event supervision. However, inadequate training in the captioning module negatively affects both localization and captioning. To address this, we generate a robust language model by pretraining event captioning modules on publicly available datasets, such as the MSR-VTT. Our robust language model indirectly enhances the performance of event captioning and caption localization. As a result, our proposed weakly supervised dense video captioning approach, PWS-DVC, outperforms existing weakly supervised dense video captioning methods. Our methodology can provide novel insights for other tasks that require solid ground truth.

REFERENCES

[1] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu, ''Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling,'' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 7331–7341.

[2] Y. Sun, H. Xue, R. Song, B. Liu, H. Yang, and J. Fu, ''Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning,'' in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 38032–38045.

[3] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, ''MERLOT: Multimodal Neural Script Knowledge Models,'' in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 23634–23651.

[4] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, ''Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval,'' in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 1728–1738.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, ''Flamingo: a Visual Language Model for Few-Shot Learning,'' in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 23716–23736.

[6] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, ''VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text,'' in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24206–24221.

[7] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, ''Video Captioning With Attention-Based LSTM and Semantic Consistency,'' *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, Sep. 2017.

[8] B. Wang, L. Ma, W. Zhang, and W. Liu, ''Reconstruction Network for Video Captioning,'' in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7622–7631.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

[9] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele, "Translating Video Content to Natural Language Descriptions," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 433–440.

[10] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. W. Microsoft, "SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 17949–17958.

[11] P. H. Seo, A. Nagrani, A. Arnab, and C. Schmid, "End-to-end Generative Pretraining for Multimodal Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17938–17947.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2014, pp. 1725–1732.

[13] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16375–16387.

[14] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2630–2640.

[15] L. Zhou, C. Xu, and J. Corso, "Towards Automatic Learning of Procedures From Web Instructional Videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, Apr. 2018, pp. 7590–7598.

[16] G. Huang, B. Pang, Z. Zhu, C. Rivera, and R. Soricut, "Multimodal Pretraining for Dense Video Captioning," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, ser. AACL'20. Suzhou, China: Association for Computational Linguistics, Nov. 2020, pp. 470–490.

[17] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-Captioning Events in Videos," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017, pp. 706–715.

[18] C. Summers and M. J. Dinneen, "Nondeterminism and Instability in Neural Network Optimization," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 9913–9922.

[19] T. Wang, J. Zhang, F. Zheng, W. Jiang, R. Cheng, and P. Luo, "Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos," Mar. 2023, *arXiv:2303.06378*.

[20] N. Aafaq, A. Mian, N. Akhtar, W. Liu, and M. Shah, "Dense Video Captioning With Early Linguistic Information Fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 2309–2322, 2023.

[21] Q. Zhang, Y. Song, and Q. Jin, "Unifying Event Detection and Captioning as Sequence Generation via Pre-Training," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin, Heidelberg: Springer-Verlag, Jul. 2022, pp. 363–379.

[22] W. Zhu, B. Pang, A. V. Thapliyal, W. Y. Wang, and R. Soricut, "End-to-end Dense Video Captioning as Sequence Generation," in *Proceedings of the International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022, pp. 5651–5665.

[23] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10714–10726.

[24] W. Choi, J. Chen, and J. Yoon, "Parallel Pathway Dense Video Captioning With Deformable Transformer," *IEEE Access*, vol. 10, pp. 129899–129910, 2022.

[25] V. Iashin and E. Rahtu, "A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer," in *Proceedings of the British Machine Vision Conference (BMVC)*, May 2020.

[26] T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu, "Event-Centric Hierarchical Representation for Dense Video Captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1890–1900, May 2021.

[27] M. Suin and A. N. Rajagopalan, "An Efficient Framework for Dense Video Captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. Association for the Advancement of Artificial Intelligence (AAAI), Apr. 2020, pp. 12039–12046.

[28] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-End Dense Video Captioning with Masked Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018, pp. 8739–8748.

[29] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018, pp. 7190–7198.

[30] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, "Streamlined Dense Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019, pp. 6588–6597.

[31] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly Localizing and Describing Events for Dense Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Jun. 2018, pp. 7492–7500.

[32] C. Deng, S. Chen, D. Chen, Y. He, and Q. Wu, "Sketch, Ground, and Refine: Top-Down Dense Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 234–243.

[33] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and 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.

[34] W. Choi, J. Chen, and J. Yoon, "Step by Step: A Gradual Approach for Dense Video Captioning," *IEEE Access*, vol. 11, pp. 51949–51959, 2023.

[35] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang, "Weakly Supervised Dense Event Captioning in Videos," in *Advances in Neural Information Processing Systems*, ser. NIPS'18, vol. 31. Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 3063–3073.

[36] T. Rahman, B. Xu, and L. Sigal, "Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019, pp. 8908–8917.

[37] S. Chen and Y.-G. Jiang, "Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8425–8435.

[38] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, "Improving language models by retrieving from trillions of tokens," Dec. 2021, *arXiv:2112.04426*.

[39] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "VTX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4581–4591.

[40] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 447–463.

[41] D. Chen and W. Dolan, "Collecting Highly Parallel Data for Paraphrase Evaluation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 190–200.

[42] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016, pp. 5288–5296.

[43] P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang, "Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1029–1038.

[44] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly Modeling Embedding and Translation to Bridge Video and Language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4594–4602.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756

[45] Y. Pan, T. Yao, H. Li, and T. Mei, "Video Captioning with Transferred Semantic Attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6504–6512.

[46] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko, "Sequence to Sequence – Video to Text," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4534–4542.

[47] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "M3: Multimodal Memory Modelling for Video Captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7512–7520.

[48] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4584–4593.

[49] Y. Xiong, B. Dai, and 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.

[50] J. Lei, L. Wang, Y. Shen, D. Yu, T. Berg, and M. Bansal, "MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 2603–2614.

[51] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, Jan. 2020.

[52] M. Koupaee and W. Y. Wang, "WikiHow: A Large Scale Text Summarization Dataset," Oct. 2018, *arXiv:1810.09305*.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NIPS)*, ser. NIPS'17, vol. 30. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.

[54] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE, Dec. 2015, pp. 4489–4497.

[55] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal Activity Localization via Language Query," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5267–5275.

[56] Y. Yuan, T. Mei, and W. Zhu, "To Find Where You Talk: Temporal Sentence Localization in Video with Attention Based Location Regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33. Association for the Advancement of Artificial Intelligence (AAAI), Jul. 2019, pp. 9159–9166.

[57] S. Zhang, H. Peng, J. Fu, and J. Luo, "Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. Association for the Advancement of Artificial Intelligence (AAAI), Apr. 2020, pp. 12870–12877.

[58] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, and Z. Xu, "Fine-grained Iterative Attention Network for Temporal Language Localization in Videos," in *Proceedings of the ACM International Conference on Multimedia (MM)*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 4280–4288.

[59] M. Gao, L. Davis, R. Socher, and C. Xiong, "WSLLN:Weakly Supervised Natural Language Localization Networks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1481–1487.

[60] Z. Chen, L. Ma, W. Luo, P. Tang, and K.-Y. K. Wong, "Look Closer to Ground Better: Weakly-Supervised Temporal Grounding of Sentence in Video," Jan. 2020, *arXiv:2001.09308*.

[61] A.-A. Liu, Y. Qiu, Y. Wong, Y.-T. Su, and M. Kankanhalli, "A Fine-Grained Spatial-Temporal Attention Model for Video Captioning," *IEEE Access*, vol. 6, pp. 68463–68471, 2018.

[62] J. Zhang and Y. Peng, "Object-Aware Aggregation With Bidirectional Temporal Graph for Video Captioning," in *Proceddings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2019, pp. 8327–8336.

[63] S. Chen and Y.-G. Jiang, "Motion Guided Spatial Attention for Video Captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, Jul. 2019, pp. 8191–8198.

[64] B. Wang, L. Ma, W. Zhang, W. Jiang, J. Wang, and W. Liu, "Controllable Video Captioning With POS Sequence Guidance Based on Gated Fusion Network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019, pp. 2641–2650.

[65] J. Hou, X. Wu, W. Zhao, J. Luo, and Y. Jia, "Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 8917–8926.

[66] H. Chen, K. Lin, A. Maye, J. Li, and X. Hu, "A Semantics-Assisted Video Captioning Model Trained With Scheduled Sampling," *Front Robot AI*, vol. 7, p. 475767, Sep. 2020.

[67] Z. Zhang, Y. Shi, C. Yuan, B. Li, P. Wang, W. Hu, and Z.-J. Zha, "Object Relational Graph With Teacher-Recommended Learning for Video Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2020, pp. 13278–13288.

[68] H. Chen, J. Li, and X. Hu, "Delving deeper into the decoder for video captioning," in *Proceddings of the European Conference on Artificial Intelligence (ECAI)*, Jan. 2020.

[69] B. Yang, Y. Zou, F. Liu, and C. Zhang, "Non-Autoregressive Coarse-to-Fine Video Captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35. Association for the Advancement of Artificial Intelligence (AAAI), May 2021, pp. 3119–3127.

[70] S. Li, B. Yang, and Y. Zou, "Adaptive Curriculum Learning for Video Captioning," *IEEE Access*, vol. 10, pp. 31751–31759, 2022.

[71] R. S. Bhooshan and S. K., "A Multimodal Framework for Video Caption Generation," *IEEE Access*, vol. 10, pp. 92166–92176, 2022.

[72] S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 65–72.

[73] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[74] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[75] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, and M. Nagata, "SODA: Story Oriented Dense Video Captioning Evaluation Framework," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin, Heidelberg: Springer-Verlag, 2020, pp. 517–531.

[76] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Nov. 2017.



WANGYU CHOI received the B.S. degree in computer science and engineering from Tech University of Korea in 2017. He is currently pursuing a Ph.D degree in Computer Science and Engineering, at Hanyang University, Korea. His research interests include video streaming, mobile networking, and computer vision.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2023.3331756



JIASI CHEN received the B.S. degree from Columbia University, New York, NY, USA, with internships at AT&T Labs Research, Florham Park, NJ, USA, and NEC Labs America, Princeton, NJ, USA, and the Ph.D. degree from Princeton University, Princeton, NJ, USA. She is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, University of Michigan, MI, USA. Her current research interests include edge computing, wireless and mobile systems, and multimedia networking, with a recent focus on machine learning at the network edge to aid augmented reality (AR)/virtual reality (VR) applications. Dr. Chen was a recipient of the Hellman Fellowship and the UCR Regents Faculty Fellowship.



JONGWON YOON received the B.S. degree in computer science from Korea University in 2007, and the M.S. and Ph.D. degrees from the University of Wisconsin-Madison in 2012 and 2014, respectively. He was a recipient of Lawrence Landweber Fellowship during his Ph.D. study. He is currently an Associate Professor in the Department of Computer Science and Engineering at Hanyang University, Korea. He is the founding director of the Intelligent Machines laboratory which broadly focuses on research in machine learning, artificial intelligence, mobile systems, and HCI.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4