

# A Space Information-Enhanced Dense Video Caption for Indoor Human Action Recognition

Bin CHEN  
ISEE, Kyushu University  
Kyushu University  
Fukuoka, Japan  
bin.367@s.kyushu-u.ac.jp  
\*Corresponding author

Yugo NAKAMURA  
ISEE, Kyushu University  
Kyushu University  
Fukuoka, Japan  
y-nakamura@ait.kyushu-u.ac.jp

Shogo FUKUSHIMA  
ISEE, Kyushu University  
Kyushu University  
Fukuoka, Japan  
shogo@ait.kyushu-u.ac.jp

Yutaka ARAKAWA  
ISEE, Kyushu University  
Kyushu University  
Fukuoka, Japan  
arakawa@ait.kyushu-u.ac.jp

**Abstract**—Dense video captioning tasks are used to detect interesting events and provide descriptive text for these events from untrimmed videos. This technology has the potential to be used in security surveillance and human care applications. However, current methods often overlook the relationships between objects in the video, which limits their applicability and makes it challenging to adapt them to specific domains, such as video summarization for indoor human activities. In these scenarios, human activities are closely intertwined with the objects in the scene. In this paper, we propose a plug-and-play module designed to enhance existing dense video captioning methods with spatial information. Specifically, we extract spatial information about the interesting objects using Red-Green-Blue-depth (RGB-D) images and the results of image segmentation. We then integrate this information into the captions generated by the Dense Video Captioning (DVC) method using a fine-tuned Large Language Model (LLM).

We evaluate the performance of our model on a custom dataset and demonstrate that our system provides a convenient and effective approach for obtaining space-enhanced captions.

**Index Terms**—Dense video caption, Image segmentation, Depth camera, Human action recognition

## I. INTRODUCTION

Video is an important information medium in security monitoring and smart home applications. The large volume of video content demands the need for automated methods to summarize and compactly represent the essential content [1]. One promising approach to creating content summaries is to use dense video captions, a technique that generates descriptive text for each frame of video [3].

However, existing dense video captioning methods and related datasets have often overlooked spatial information and object relationships within the visual contents of videos. This limitation hinders the adaptability of these methods to other application scenarios, such as generating activity log files for indoor individuals or indoor security monitoring, where the same activity can have varying interpretations with different indoor positions. Intuitively, to clearly and unambiguously describe indoor human behavior, captions should include information about the action class, location, and associated objects.

In this paper, we introduce a plug-and-play module to enhance spatial information in captions, building upon the current popular approach. As illustrated in Fig. 1, we use

Red-Green-Blue-Depth (RGB-D) data as input and directly extract accurate spatial information from the RGB-D images. Subsequently, we combine this spatial information with object relationships using a fine-tuned sequence-to-sequence (Seq2Seq) text generation model to generate the unambiguous and detailed captions. In our scheme, the Region of Interest (RoI) to which we pay particular attention include: 1. The human’s position in relation to the main object. 2. The position of the object interacting with the human. 3. The ambiguous object in relation to other objects of the same class. We collected a dense video caption dataset enhanced with spatial information and object relationships of interesting objects and then evaluated our system on the dataset to show its effectiveness. Our results demonstrate that the system is capable of effectively generating dense video captions with space information.

Our contribution is here:

- 1) We collected a dataset for dense video captioning, enriched with spatial information and the relationships among interesting objects.
- 2) We designed a plug-and-play module for existing methods to generate space-enhanced unambiguous captions.

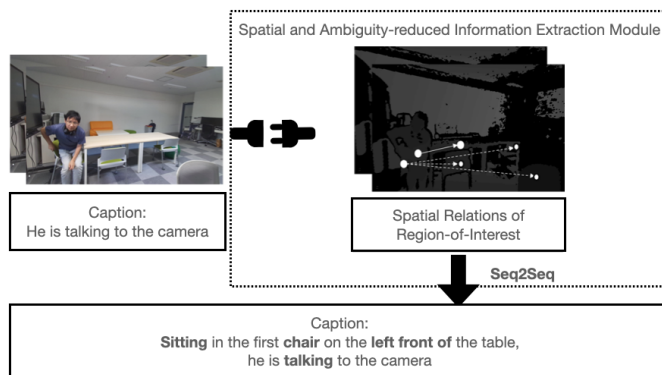


Fig. 1. Main scheme of the system

## II. RELATED WORK

### A. Dense video captioning and Dataset

Dense video captioning is an emerging sub-field in video understanding, with a focus on event localization and event captioning within long, untrimmed videos that often contain a mix of events and irrelevant background content. Compared to traditional video captioning methods, dense video captioning aims to generate descriptions for all events within a video, making it suitable for application scenarios that require detailed textual descriptions.

Based on the input, there are two main categories of methods in this field: 1. Multi-modality-based methods and 2. Vision-modality-based methods. For example, Iashin et al. utilize audio, speech, and visual information as input and merge the extracted features with a multi-modal generator to output a distribution over the vocabulary [1]. Yang et al. propose vid2seq [2], which uses video and transcribed speech as input to achieve state-of-the-art results in the dense video captioning task. However, in the real world, audio or speech inputs often come with significant background noise, which can potentially impact the results of an application. Fortunately, there are also models based solely on vision that have achieved state-of-the-art results.

For instance, Wang et al. introduced a framework for end-to-end dense video captioning with parallel decoding (PDVC) using only visual input. They formulated dense caption generation as a set prediction task. In their pipeline, they employed CNN and Transformer encoders to extract features. In the parallel decoding phase, they implement a Transformer decoder with three heads (event counter, localization header, and caption header) to generate the localization and caption independently.

However, current vision-based end-to-end methods often neglect the spatial information and fail to address ambiguity. Furthermore, the existing datasets used in dense video captioning do not adequately incorporate this information, despite the presence of some spatial details within the dataset’s ground truth sentences. The current popular datasets used in dense video captioning include YouCook2 and ActivityNet. YouCook2 is one of the largest task-oriented instructional video datasets, containing 2000 long untrimmed videos from 89 cooking recipes. On average, each distinct recipe has 22 videos [5]. The ActivityNet Captions dataset is based on ActivityNet v1.3 and includes 20,000 YouTube untrimmed videos with 100,000 caption annotations. These videos have an average length of 120 seconds [6]. We observed that the ground truth captions in ActivityNet often contain sentences that describe the relationships among interesting objects for accurate captions. These datasets are typically used to train end-to-end models and require substantial human effort for collection. Hence, there is a need for a more accurate and efficient method to incorporate spatial information and expand the range of application scenarios.

### B. Scene graph generation and image segmentation

Scene Graph Generation (SGG) is an important technology for vision-based scene understanding. Its primary goal is to analyze and recognize visual content and represent their relationship as a graph structure. In previous methods, objects are detected using bounding boxes, followed by the prediction of their pairwise relationships. In 2022, J. Yang et al. introduced Panoptic Scene Graph Generation (PSG), a novel task that challenges models to create a more comprehensive scene graph representation based on panoptic segmentation, as opposed to rigid bounding boxes. In contrast to SGG, the PSG task consistently involves three sub-tasks: semantic segmentation, instance segmentation, and relationship prediction. In their paper, they present a transformer-based method and define 56 predicate relationships. Following their design, Zijian Zhou et al. introduced a new method, Hilo [9], for the PSG task. In their paper, they utilize a DETR-based panoptic segmentation method, Mask2Former [8], as the preprocessing step to obtain accurate panoptic segmentation results. They subsequently generate more precise outcomes.

As a result, the existing methods are typically required to predict all the relationships they detect. However, in application scenarios where only spatial information about interesting objects is necessary, existing methods encounter 3 issues: 1. An excessive number of relations. 2. Predictions are not always accurate. 3. Less emphasis on spatial information.

## III. PROPOSED METHOD

### A. Problem Setting

Dense video captioning aims to identify a set of temporally-localized captions of suitable length. In our paper, we model this task with the following equation 1, where  $t_j^s$ ,  $t_j^e$ ,  $S_j$  represent the starting time of the event, ending time of an event, and the caption of an event, respectively.  $f_e$ ,  $f_o$ ,  $f_l$  represent the video caption module, spatial and ambiguity-reduced information extraction module, and sequence-to-sequence text generation module, respectively. In our evaluation section, we will use  $t_j^s$  and  $t_j^e$  to measure localization performance, while  $S_j$  will be used to measure dense video caption performance. We model the constitute of  $S_j$  as equation 2, where  $S_j^s$ ,  $S_j^a$ ,  $S_j^o$  represent space information, action information, and object information.

$$\{(t_j^s, t_j^e, S_j)\}_{j=1}^{N_{set}} = f_l(f_e(RGB), f_o(RGB - D)) \quad (1)$$

$$S_j = f(S_j^s, S_j^a, S_j^o) \quad (2)$$

### B. Framework

The core insight of our system is to design a system that can incorporate spatial information to generate detailed and unambiguous captions. The entire pipeline is illustrated in Fig. 2. The system comprises 3 main components: video caption generation, spatial and ambiguity-reduced information extraction, and sequence-to-sequence(seq2seq) text generation.

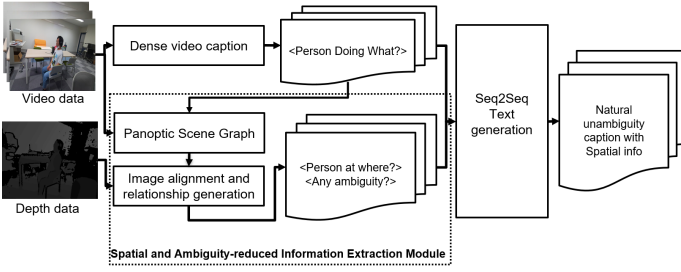


Fig. 2. Architecture of the System

**Feature Extractor.** Humam Alwassel et al. [14], highlight that traditional feature extractors are often trained for trimmed action classification tasks. Consequently, to obtain features more suitable for untrimmed videos, we have implemented the TSP feature extractor to extract the corresponding features, replacing the use of the C3D [10] features.

**Video Captioning Generation.** The Parallel Decoding architecture for dense video captioning is a novel design introduced by Teng Wang et al. in 2022 [3]. In contrast to the traditional two-stage pipeline, they employ a transformer encoder and decoder architecture to simultaneously predict the location and caption, followed by merging the location and caption with the predicted scores.

The localization head is designed to perform box prediction and binary classification tasks using a multi-layer perceptron. The results will be saved as sets of tuples  $\{t_j^s, t_j^e, c_i^{loc}\}_{j=1}^N$ , where  $c_i^{loc}$  stand for the confidence of the event. The caption head is responsible for generating sentences that describe the event. As this head cannot determine the boundaries of the event, the deformable soft attention (DSA) and LSTM are used to generate a sentence  $S_j = \{w_{j1}, \dots, w_{jM_j}\}$ , where the  $w_{jt}$  means the word in the  $j$  proposal and  $M_j$  is the sentence length.

**Spatial and Ambiguity-reduced Information Extraction.** In order to extract spatial and unambiguous information from RGB-D video sequences and generate concise sentences  $S_j^s$  and  $S_j^o$  for descriptions, we utilize the previous output  $(t_j^s, t_j^e)_{j=1}^{N_{set}}$  as the input and consider the RGB and depth images with a time stamp computed as  $(t_j^s + t_j^e)/2$  as the key for the  $j^{th}$  sentence.

We implement a Mask2Former to obtain a more accurate image segmentation result and based on that, we implement a fine-tuned OpenPSG [7] module to generate all the relationships of the visual contents.

To extract spatial and unambiguous information from keyframes, we adhere to the following design principles: 1. The main object is chosen from the three largest objects and centered. 2. Additional clarification is introduced only when ambiguity arises with other objects. We reuse the ‘human action’ class from OpenPSG, employ panoptic segmentation from Mask2Former, and align RGB-D images to directly generate information about:[human position related to the main object],[human interact with what] and [which object he interact with] based on the depth of the interest objects.

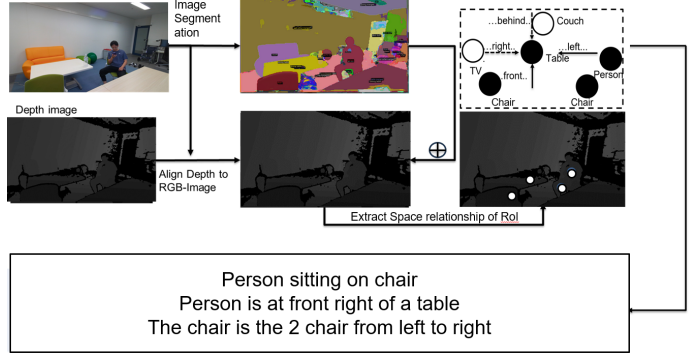


Fig. 3. Spatial and Ambiguity-reduced Information Extraction

**Seq2Seq Text Generation.** To generate natural sentences, we utilize a Large Language Model (LLM) to combine and summarize the short sentences. ChatGPT is a well-known model in the field of natural language processing and has produced excellent results for a variety of tasks. However, ChatGPT is not freely available. To make the system applicable to a wide range of applications, we implemented a fine-tuned ‘falcon-7b<sup>1</sup>’ model that integrates all information into longer sentences with a more natural structure and human comprehension.

## IV. EXPERIMENT

### A. Experiment setting

**Dataset.** There is currently no benchmark dataset that comprises RGB-D images of indoor scenes and annotations in the form of natural sentences describing human behavior. To address this gap, we have gathered a dataset consisting of 30 untrimmed, long videos for training and testing. Within the dataset, we have designed 3 different scenes, 4 position classes, and 2 subjects. Additionally, each video includes both unambiguous and ambiguous action classes.

**Metrics.** The method is evaluated in two aspects: 1. For location performance, the average precision and average recall across IOU(Intersection over Union) at [0.3, 0.5, 0.7, 0.9] are used. 2. For dense captioning performance, we calculate the BLEU4 [10], METEOR [11], and CIDER [12] of the marched pairs between generated captions and ground with IOU thresholds of [0.3, 0.5, 0.7, 0.9].

### B. Implementation details

We collect the dataset with one Microsoft Azure Kinect. The video is 30 FPS with 720P. We implemented a TSP feature-updated PDVC and pre-trained it on ActivityNet dataset as the backbone. In the spatial and ambiguity-reduced information extraction module, we implement a Mask2Former model which is trained on the coco panoptic dataset for detection and segmentation of daily objects. We designe 4 position relations (left, right, front, back) and 4 human-object interaction classes (walking on, standing on, lying on, sitting on) to generate based on OpenPSG.

<sup>1</sup>Falcom LLM <https://falconnllm.tii.ae>

### C. Comparison Results

We provide qualitative comparison results and quantitative comparison results to the existing method on localization performance and dense video caption performance.

**Localization performance.** To create a plug-and-play module aimed at improving the quality of captions generated by existing methods, we reuse the time intervals generated by those methods. As a result, we achieve the same precision and recall scores as the PDVC method, as demonstrated in the table I.



		
Ground Truth	He is sitting on the chair in the front-left of the table. The chair is the first one from the left.	He is standing on the floor behind the right side of the table, playing with a ball.
PDVC	He is seen speaking to the camera lead into he talk	He then put the ball down and down the ball
GVL	A man is talking to the camera	He then begin to play the ball and down the ball
Ours	He is <b>sitting</b> in the <b>front left-hand</b> position of the table, with his chair being the <b>first chair</b> . He is facing the camera and engaging in a conversation.	He is <b>standing on the floor</b> to the <b>right</b> of the table, and he then puts a ball down and kicks it.

Fig. 4. Qualitative Comparison Results

**Dense video caption performance.** In Fig.4, we present the ground truth of our dataset in a video file, along with the corresponding caption results from existing methods and ours. In this qualitative comparison, our methods successfully add spatial information and eliminate ambiguity.

The quantitative comparison results are presented in Table I. All metrics are evaluated across IOU at [0.3, 0.5, 0.7, 0.9] and averaged. Our system shows a small improvement in BLEU4, which measures text similarity based on the precision of 4-grams, and significant improvements in METEOR, which provides a comprehensive evaluation considering word choice and order, as well as CIDEr, which assesses diversity and consensus in captions.

TABLE I  
CAPTIONING AND LOCALIZATION PERFORMANCE

	<i>PDVC</i>	<i>GVL</i>	<i>Ours</i>
Bleu1	9.1	7.6	9.9
Bleu2	5.7	4.8	5.4
Bleu3	3.6	3.1	3.2
Bleu4	1.2	1.3	<b>1.6</b>
METEOR	3.6	2.8	<b>6.1</b>
CIDEr	3.4	2.8	<b>9.9</b>
Recall	49.1	56.24	49.1
Precision	21.0	13.7	21.0

In other words, our system excels in generating space-enhanced and unambiguous captions. Compared to existing methods, it is better suited for application scenarios that

require space information and clarity in human action recognition.

### V. CONCLUSION

In this paper, we have proposed a plug-and-play module to enhance the quality of captions generated by existing methods. The objective is to ensure that captions contain three essential components: [what the person is doing], [where the person is located], and [the actions performed with which objects], while minimizing ambiguity. Our system focuses on the spatial and ambiguity-reduced information extraction module. The module takes depth and color images from the time intervals of the previous step as input and produces spatial information by employing a fine-tuned PSG process based on RoI and depth data. Experimental results obtained from the custom dataset demonstrate the effectiveness of our system as a pipeline for incorporating spatial information and eliminating ambiguity.

### REFERENCES

- [1] V.Iashin, and E.Rahtu, "Multi-modal dense video captioning," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 958–959, 2020.
- [2] A.Yang, A.Nagrani, et al., "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10714–10726, 2023.
- [3] T.Wang, H.Zheng, M.Yu, Q.Tian, and H.Hu, "Event-centric hierarchical representation for dense video captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1890–1900, 2020.
- [4] T.Wang, et al., "Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos," arXiv preprint arXiv:2303.06378, Mar 2023.
- [5] L.Zhou, C.Xu, and J.Corso, "Towards automatic learning of procedures from web instructional videos," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, April 2018.
- [6] J.Wang, W.Jiang, L.Ma, W.Liu, and Y.Xu, "Bidirectional attentive fusion with context gating for dense video captioning," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7190–7198, 2018.
- [7] J.Yang, et al., "Panoptic scene graph generation," European Conference on Computer Vision. Cham: Springer Nature Switzerland, pp. 178–196, Oct 2022.
- [8] B.Cheng, et al., "Masked-attention mask transformer for universal image segmentation," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1290–1299, 2022.
- [9] Z.Zhou, M.Shi, and H.Caesar, "HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation," arXiv preprint arXiv:2303.15994, 2023.
- [10] D.Tran, L.Bourdev, R.Fergus, L.Torresani, and M.Paluri, "Learning spatiotemporal features with 3d convolutional networks," Proceedings of the IEEE international conference on computer vision, pp. 4489–4497, 2015.
- [11] K.Papineni, S.Roukos, T.Ward, and W.B.Zhu, "A method for automatic evaluation of machine translation," the Proceedings of ACL-2002, July 2002.
- [12] S.Banerjee, and A.Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72, June 2005.
- [13] R.Vedantam, C.Lawrence Zitnick, and D.Parikh, "Cider: Consensus-based image description evaluation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566–4575, 2015.
- [14] H.Alwassel, S.Giancola, and B.Ghanem, "Tsp: Temporally-sensitive pretraining of video encoders for localization tasks," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3173–3183, 2021.