

Retrieval-augmented-generation-enhanced Dense Video Caption for Human Indoor Activities: Disambiguating Caption Using Spatial Information Beyond Field of View Constraints

Bin Chen,^{*} Yugo Nakamura, Shogo Fukushima, and Yutaka Arakawa

Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University,
744 Motoooka, Nishi-ku, Fukuoka 819-0395, Japan

(Received July 16, 2024; accepted September 9, 2024)

Keywords: depth camera, 3D reconstruction, 3D detection, human activity recognition, dense video caption, retrieval-augmented generation, large language model

Dense video captioning aims to extract every goal event and its corresponding period, and has gathered significant attention owing to its potential and valuable applications in smart homes, human care, security monitoring, and more. However, current methods do not sufficiently reduce ambiguity in the generated captions and have a limited field of view, making it difficult to integrate the relationship between people and their surrounding environment into the captions. The limitations restrict their applicability in smart home or indoor security systems. These systems require clear distinctions between normal and abnormal human actions, as similar actions can have different interpretations depending on the surrounding environment. For instance, in indoor security systems, ambiguous captions might fail to distinguish between harmless activities such as a person walking and potentially concerning behaviors such as unauthorized access attempts. In this article, we propose a retrieval-augmented generation (RAG)-based system to enhance existing methods, making them suitable for recording human activity indoors. Our key ideas are as follows. First, we collect information about the house environment using a 3D reconstruction and 3D detection process to build a knowledge base. Second, we design a RAG procedure to extract the relevant environmental context. Third, we develop a spatial information enhancement query based on human detection results from RGB and depth image pairs. We utilize the summarization and reasoning capabilities of a large language model to fuse all information, thereby obtaining spatially enhanced dense captions of human indoor activities. We evaluate our method by comparing it with PDVC (end-to-end Dense Video Captioning with Parallel Decoding), GVL (Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos), and SG-PDVC (Scene Graph-enhanced PDVC) on a custom video dataset collected from two houses with eight different camera positions, using Recall, Precision, Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and Consensus-based Image Description Evaluation (CIDEr) as metrics. Our method outperforms the three compared methods in

^{*}Corresponding author: e-mail: bin.367@s.kyushu-u.ac.jp
<https://doi.org/10.18494/SAM5238>

BLEU-3, METEOR, CIDEr, and Precision metrics, but shows a small decline in ROUGE-L and Recall metrics compared with GVL. These results demonstrate that our method effectively incorporates spatial information and reduces ambiguity for indoor human activity caption applications.

1. Introduction

The video medium serves as important input in many vision-based security monitoring and smart home applications. However, the video stream is very difficult to search and store, so the large volume of video content requires automated methods to summarize and compactly represent essential contents.⁽¹⁾ Fortunately, with the development of deep learning, the dense video caption, which is a technique that generates descriptive text for each event of the video,⁽²⁾ provides a convenient way to create content summaries and record the content in the video.

However, current dense video captioning methods are not designed to effectively capture the spatial relationship between individuals and their surrounding environment in indoor human activity. In Fig. 1, the targeted problem is illustrated where a person in a video performs a single action across three different positions, each with various environmental information. Current existing methods such as end-to-end Dense Video Captioning with Parallel Decoding (PDVC) and Learning Grounded Vision-Language Representation for Versatile Understanding in Untrimmed Videos (GVL), popular in the dense video field and commonly trained on the ActivityNet dataset, focus solely on human action categories. Consequently, they often fail to distinguish between the three events, resulting in ambiguous captions. This limitation often has a significant negative effect on application scenarios where spatial context is important, such as indoor security monitoring. In these cases, detecting an abnormal behavior is the primary focus, and understanding the position of a person relative to the environment is crucial for interpreting activities correctly. For example, in indoor security systems, ambiguous captions might fail to distinguish between harmless activities, such as a man standing by a chopping board and cutting something, and potentially suspicious behaviors, like a man cutting something on the left side behind a bed. In video-based remote care systems for patients, ambiguous captions might fail to distinguish between normal activities, such as a man sleeping on the bed, and potentially abnormal behaviors, like a man lying in the area in front of the door.

There is also much research aimed at solving the problem of low-quality, ambiguous captions mentioned above. In some research, large datasets are utilized to train and construct large models, such as Vid2Seq,⁽³⁾ which is pretrained on YT-Temporal-1B,⁽⁴⁾ a collection of 20 million YouTube videos with English subtitles covering diverse topics. Training and adapting such end-to-end models to specific scenarios are challenging as they require retraining the model to incorporate contextual knowledge. Tu *et al.* devised a textual-temporal attention (TTA) model that extends the conventional temporal attention model by adding another textual attention model to reduce the ambiguity in the video caption task.⁽⁵⁾ Specifically, they extracted object information from video frames and used the labels of these objects to enhance the quality of the captions. Following their work, the Scene Graph-enhanced PDVC (SG-PDVC)⁽⁶⁾ that we proposed also utilizes object information detected in the frames to reduce ambiguity, and more

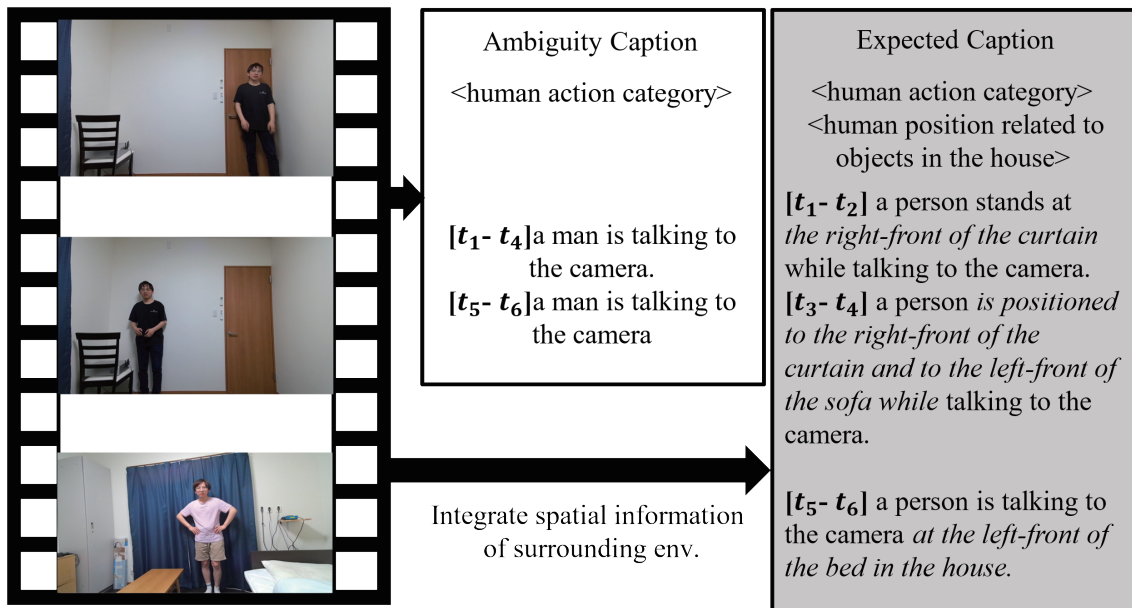


Fig. 1. (Color online) Target problem and expected result. The video includes three events of the same class but with different spatial information.

details were added for improved caption results. However, this type of model lacks knowledge of the surrounding environment owing to its limited field of view (FOV), making it difficult to extract the relationship between the person and their surroundings. As a result, it is not suitable for certain human-serving applications where understanding the spatial relationship between the person and their surroundings is crucial.

To create a more convenient system that can be easily adapted to various application scenarios and introduce the relationship between human actions and the surrounding environment beyond the camera's FOV, we designed a system based on the RAG, which can combine retrieval and in-context learning (ICL) techniques to enhance the results of large language models (LLMs). The system overcomes FOV constraints by using spatial context beyond the camera's view, leveraging a knowledge base to clarify actions and positions outside the visible area. It integrates information about surrounding objects to generate high-quality, natural descriptions of indoor human activity. The RAG-based architecture of the system is shown in Fig. 2. The architecture of our system can be simply divided into three parts: (1) 3D reconstruction and 3D object detection are used to build a knowledge base. Building the knowledge base only needs to be performed once when setting up the system (initial setup); (2) a spatial-information-enhanced query that integrates the spatial information of humans and the original captions is obtained from RGB-D stream, allowing RAG to extract related information from the knowledge base; (3) a designed RAG-enhanced LLM module that fuses and summarizes all related information in the query and spatial context to output the spatial-information-enhanced caption for indoor human action. It is convenient to adapt our system to different rooms, as the initial setup only needs to be run once. This effectively reduces the computational burden.

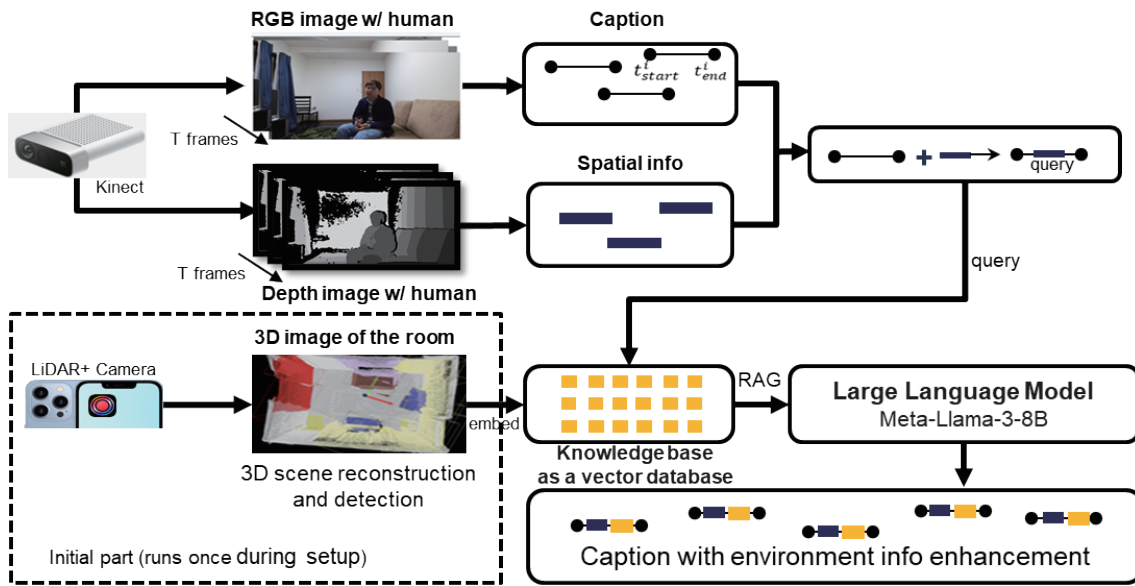


Fig. 2. (Color online) Architecture of the whole system. The initial part only needs to be run once when configuring the system in a new room, reducing the computational cost.

We evaluated our method on a custom video dataset collected from two houses with eight different camera positions. We compared our method with PDVC and GVL, two popular methods in the dense video captioning field, as well as a similar system, SG-PDVC,⁽⁶⁾ using the same class and temporally sensitive pretraining (TSP) features. The evaluation metrics include Recall, Precision, BLEU, METEOR, ROUGE-L, and CIDEr. Our method outperforms the three compared methods in BLEU-3, METEOR, and CIDEr metrics, but shows a small decline in ROUGE-L metrics compared with GVL. For localization performance, our method achieves the highest Precision score among the compared methods but shows only a slight improvement in Recall, which is still lower than that of GVL. These results demonstrate that our method effectively incorporates spatial information and reduces ambiguity for indoor human activity caption applications. Our contributions are as follows.

1. **Dataset Collection:** We collected a dataset for dense video captioning focused on the spatial information of indoor human activity. This dataset emphasizes the spatial relationships between people and objects in indoor environments.
2. **Dense Video Captioning System with Improved Spatial Awareness:** We developed a dense video captioning system specifically designed to address the limitations of current methods in capturing the spatial relationships between an individual and the surrounding environment for recording indoor human activity. This enhancement reduces ambiguity in the generated captions, particularly in scenarios where spatial context is critical, such as indoor security monitoring.
3. **Integrating Spatial Knowledge Base and Overcoming FOV Constraints:** Our system uniquely integrates a spatial knowledge base that encodes information about objects within the environment, producing captions enhanced with spatial context. Additionally, it also addresses traditional FOV constraints, enabling the generation of contextually relevant captions.

2. Related Work

2.1 Dense video caption

Dense video caption tasks are used to detect interesting events and provide descriptive text for these events from untrimmed videos.⁽²⁾ Compared with common video captioning methods, dense video captioning has the following two characteristics. (1) It captures all detected events. In the dense video captioning procedure, it aims to output all detected actions, meaning that it can include multiple action sentences for the same period. (2) It is applicable to untrimmed video. Unlike common video captioning methods, dense video captioning techniques, such as PDVC and GVL, do not require human-annotated boundaries of the video. These methods, which directly use untrimmed video as input, can output the time period along with the captions. As a result, it has great potential to be used in security surveillance and human care systems.

Multimodality-based methods. These methods tend to use multiple source modalities to work together, making the output more accurate. For example, MDVC, a popular dense video captioning method proposed by Iashin and Rahtu in 2020, utilizes audio, speech, and visual information as modalities to design a multimodal generator that outputs a distribution over the vocabulary.⁽¹⁾ Vid2Seq, proposed by Yang *et al.* in 2023, uses video and transcribed speech as inputs to achieve state-of-the-art results in the dense video caption task.⁽⁷⁾ However, various modalities may contain overlapping or contradictory information, and negative interactions between these modalities can reduce the effectiveness of caption generation.

Vision-modality-based methods. For example, PDVC, an end-to-end framework with parallel decoding proposed by Wang *et al.*, uses video data as the input modality.⁽²⁾ In their method, the dense video captioning problem is modeled as a set prediction task. They first used a CNN and transformer encoder to extract and encode image features. Then, they applied a parallel decoding phase with a transformer decoder, which features three heads (event counter, localization header, and caption header) to generate the localization and captions independently.⁽²⁾ Wang *et al.* proposed GVL,⁽⁸⁾ which is a grounded vision-language learning framework for untrimmed video aimed at the automatic detection of informative events and the alignment of sentence descriptions with event segments. In their work, they introduced the Event-to-Text generation module and Text-to-Event grounding module to learn event-level representation to increase the accuracy of the dense video caption task. However, similar to PDVC, GVL also does not pay enough attention to the relationship between humans and their surrounding environment. This limitation hinders their use in security surveillance and human care applications. In our previous work, we proposed SG-PDVC, a plug-and-play system designed to enhance existing dense video captioning methods with spatial information, improving the recording of indoor human actions.⁽⁶⁾ In our work, we used the panoptic scene graph generation method to extract the relationship between people and their surrounding environment and utilize the Falcon-7B model to integrate all information to obtain spatially enhanced captions. However, owing to the limitation of FOV, the relationships they extract are incomplete, resulting in the generation of unnatural captions.

2.2 RAG for LLMs

LLMs such as the GPT series,⁽⁹⁾ Llama series,⁽¹⁰⁾ and Gemini,⁽¹¹⁾ have attracted increasing interest in the field of natural language processing because they show impressive knowledge mastery abilities and exceed the human behavior in many tasks.⁽¹²⁾ However, LLMs always have two shortcomings: (1) lack of specific knowledge in certain domains and (2) possible fabrication of answers. These two shortcomings are mainly due to the fact that LLMs are trained on existing datasets. Therefore, if you ask an LLM a question that is beyond its training data domain or inquire about the latest information, it may provide incorrect answers.

RAG is an AI framework that combines retrieval and ICL techniques to improve the performance of an LLM.⁽¹²⁾ It is designed to enhance the performance of LLMs by providing external knowledge, enabling them to answer domain-specific questions or generate results on the basis of the latest information. RAG is also characterized as a cost-effective approach owing to its convenience in introducing new knowledge to LLMs without requiring retraining. The RAG framework typically operates with the following steps. (1) A user poses a query to the LLM. (2) Search algorithms extract relevant information from the knowledge base to provide context. (3) The extracted information is integrated into the prompts for the LLM, which will generate results on the basis of the enriched context.

3. Methods

We designed a system based on the RAG framework to enhance the captions generated by the existing dense video caption system, aiming to create a specific system for better human indoor action recognition and recording. It can be divided into two parts: (1) the initial part for building a knowledge database for the LLM and (2) the caption generation part that will use the spatial-information-enhanced query and enriched context extracted by an RAG process to generate the final caption result. The whole system is shown in Fig. 2.

3.1 3D reconstruction and 3D object detection to build a knowledge base (initial part)

In this step, our goal is to extract spatial information about the objects in the house, and this section can be divided into four subtasks: (1) 3D scene reconstruction for the house, (2) 3D object detection to extract the object information in the house, (3) the calculation of the position of the camera and alignment of the scene and 3D object detection results to the camera coordinates, and (4) the building of the knowledge base for the RAG procedure.

3D scene reconstruction. We use camera and LiDAR sensors from the iPhone 14 Pro as the device to generate the cloud points of the room using the Scaniverse application. Theoretically, the cloud point representation of the room can be generated by any method.

3D object detection. We implemented the method VoteNet,⁽¹³⁾ which is trained on the ScanNet 3D dataset⁽¹⁴⁾ to generate the 3D object detection results. VoteNet is an end-to-end 3D object detection network that operates directly on point cloud data and has achieved state-of-the-art 3D detection on two large real indoor 3D datasets, namely, ScanNet and SUN RGB-D.⁽¹⁵⁾ We

take the 3D point cloud of the house as the representation of the 3D space and use it as the input for VoteNet. The effectiveness of VoteNet is evaluated on ScanNet with the Average Precision (*AP*) metric, which is calculated as the area under the Precision–Recall curve and gives a single score to summarize the precision–recall trade-off of a detector. $AP@\alpha$ stands for the *AP* calculated for an intersection over union (*IoU*) in the threshold of α .

Coordinate transformation. In this step, we aim to determine the position of the camera and transform all contextual information to the camera coordinates. To increase the accuracy of the registration results, we use global registration and a fine-tuned registration process, the iterative closest point (ICP) algorithm. The whole procedure is shown in Fig. 3.

In the global part, we utilize six corresponding points of the scene point cloud and the 3D fragment (point cloud), generated by the RGB-D image pairs taken by the depth camera, to calculate the initial 3D transformation matrix. We use the reconstruction system procedure in the open3D library^(16,17) to build the point cloud to represent the 3D fragment obtained by the camera. Here, to provide a good initial 3D transformation matrix for the fine-tuned registration process, we calculate the rotation and translation matrices through the following mathematical steps: (1) calculate the covariance matrix used to capture the relationship between the centered

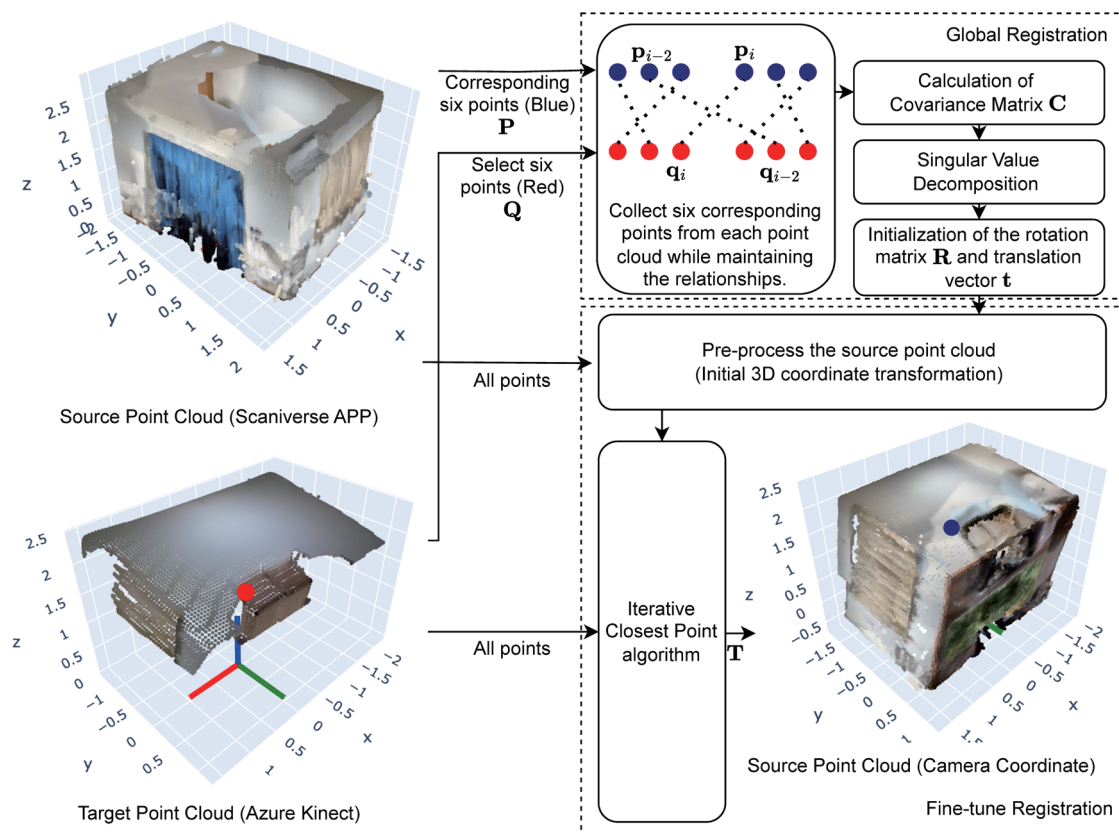


Fig. 3. (Color online) Procedure to align the source point cloud to the target point cloud by calculating the 3D coordinate transformation matrix.

source and target points, (2) perform singular value decomposition (SVD) on the covariance matrix to obtain rotation-related matrices, and (3) calculate the rotation and translation matrices to construct the initial 3D transformation matrix. In the first step, we calculate the covariance matrix C using Eq. (1), where p and q represent the selected source and target points, respectively. The columns of matrix P represent the source points, while the columns of matrix Q represent the target points. Q' is the matrix whose columns are the centered target points and P' is a matrix whose columns are the centered source points. In the second step, we perform SVD on the covariance matrix C to calculate rotation-related matrices using Eq. (2), where U and V are the orthogonal matrices to represent rotations, and S is a diagonal matrix containing the singular values. In the third step, the initial 3D transformation matrix is constructed and calculated using Eq. (3), where R and t represent the rotation matrix and the translation vector, respectively.

$$C = Q'^T P' = \left(Q - \frac{1}{6} \sum_{i=1}^6 q_i \right)^T \left(P - \frac{1}{6} \sum_{i=1}^6 p_i \right) \quad (1)$$

$$C = USV^T \quad (2)$$

$$M = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} VU^T & \frac{1}{6} \sum_{i=1}^6 p_i - VU^T \frac{1}{6} \sum_{i=1}^6 q_i \\ 0 & 1 \end{bmatrix} \quad (3)$$

In the fine-tuned registration process, we use the ICP algorithm to calculate the alignment between two point clouds. The algorithm is modeled below.

$$\min_T \sum_{i=1}^N \|Tp_i - q_i\|^2 \quad (4)$$

Here, T is the final fine-tuned 3D transformation matrix and N is the number of points. All points in the two point clouds are utilized to determine the final 3D transformation matrix.

Building the knowledge base. In this step, we build the knowledge base with the context information of the environment. The context information is represented as a latent vector that forms the basis of a searchable knowledge base for the RAG process.

To construct this knowledge base, two subtasks are involved: (1) embedding the knowledge and (2) providing suitable distance criteria for the knowledge vectors. To balance accuracy and model size, we utilize the GTE-small model,⁽¹⁸⁾ which is based on the BERT framework. This model has been pretrained on a large-scale corpus of relevant text pairs, covering a wide range of domains and scenarios, and serves as our embedding model. The embedding model converts the raw data from the CSV file, which includes 2D object positions, class labels, and object sizes from various camera setups obtained by processing 3D detection results using the corresponding

transformation matrices, into high-dimensional latent vectors. Then, we utilize Facebook AI Similarity Search (FAISS) and Euclidean distance metrics to organize and search within this vector database. The Euclidean distance measures the difference between vectors, enabling the efficient retrieval of relevant context information during the caption generation process.

3.2 Design of spatial-information-enhanced query and RAG-enhanced LLM module (caption part)

In this step, our goal is to design a module to use the spatial-information-enhanced query and extract related information from the knowledge base to output the spatial-information-enhanced caption for indoor human action.

Spatial-information-enhanced query. In this subtask, we use the caption generated by PDVC as the original caption and extract the position of the person in the camera coordinates. The PDVC is a popular and efficient vision-based dense video caption method. In our work, the PDVC we reimplemented is trained with ActivityNet⁽¹⁹⁾ and the TSP⁽²⁰⁾ feature, which improves the temporal sensitivity and is more suitable for the dense video caption task than traditional features such as C3D.⁽²¹⁾ The original caption is represented as $\{s_j, t_j^s, t_j^e, c_i^{loc}\}_{j=1}^N$, where t_j^s and t_j^e represent the start and end times of the j event, respectively. s_j and c_i^{loc} represent the caption and confidence of the j event, where $s_j = \{w_{j1}, w_{j2}, \dots, w_{jM_j}\}$; w_{jt} means the t th proposal and M_j the sentence length.

The position of the person is obtained in two parts: (1) train You Only Look Once (YOLO) v8 for human detection and (2) align the RGB result to the depth image and transform the coordinates to the camera coordinates.

YOLOv8 is the latest version of the YOLO⁽²²⁾ series and is popularly used for real-time object detection and classification in the computer version. In this work, we use YOLOv8 as the backbone to detect the position of the person in pixel coordinates. YOLOv8 is trained with the COCO8 dataset, which includes the “person” class. In our work, we aim to disambiguate the original captions with spatial information. Specifically, we compare the position of the person at the start and end times to decide whether the event should be divided. The schema can be represented by

$$Event = \begin{cases} \{(s_j, t_j^s, t_j^c), (s_j, t_j^c, t_j^e)\}, & \text{if } d_h > thr, \text{ where } t_j^c = \frac{1}{2}(t_j^s + t_j^e) \\ \{(s_j, t_j^s, t_j^e)\}, & \text{if } d_h \leq thr \end{cases} \quad (5)$$

where t_j^c is the center time stamp, d_h is the distance between the two positions of the person, and thr represents the threshold.

We utilize the trained YOLOv8 to output the detection results of the person in pixel coordinates, $[p_x, p_y]$. Then, we align the result of the RGB image to the depth image $[p_x, p_y, d]$,

where d is the value of the pixel in the depth image. We use Eq. (6) to map $[p_x, p_y, d]$ to the camera coordination.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d \cdot \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \quad (6)$$

Here, $[x, y, z]^T$ represents the detection result of the person in the camera coordinates. f_x and f_y represent the focal lengths in the x - and y -directions, respectively, and $[c_x, c_y]$ denotes the principal points (the pixel coordinates of the camera center).

LLM to output the final spatial-information-enhanced caption. In this subtask, we use the designed prompt, which includes the information of the human spatial enhanced query and context of the objects, for the LLM to generate the spatial- information-enhanced caption.

The context of the objects is obtained by the RAG procedure. The input of the RAG model uses a combination of variables that include the setups of the camera and the name of the test room. The designed template of the input is shown as the following text. The output of the process is a dictionary in data format, including the names of the objects and their positions within the house.

“”””
 collect the information of the two chosen object in <the name of the test room>. with the camera setting up on <the setups of the camera>.
 The first chosen object is : the largest object in the house.
 The second chosen object is :the second largest object in the house.
 ATTENTION: the chosen objects all come from <the name of the test room> and the camera setting up on the <the setups of the camera>.
 Pay more attention at the <the name of the test room> and the camera setting up on the <the setups of the camera>.
 “”””

The enhanced prompt, which is the input of the LLM, combines two key components, namely, human spatial enhanced query and the context of the objects.

1. Human spatial enhanced query: This initial query consists of the person’s 3D position in the camera coordinates and the original caption, which is extracted by the PDVC procedure, but does not include the spatial relationships of the human and objects.
2. Context of the objects: This is the retrieved position information of the selected objects from the RAG module.

To make the enhanced prompt, the numeric values of the positions of the objects and person are compared. The output sentences have the following fixed sentence pattern:

< he is at {relative position} of the {class of object} >.

If the x value of the person is greater than that of the objects, the position will include “right”. Similarly, if the z value of the person is greater than that of the objects, the position will include “front”. As a result, the enhanced prompt of the LLM can be constructed with the following template, which is from the “Langchain” library.

```

"""
Please answer the following question.
Give me the first result, NOT the code.

here is the format instructions:
{format_instructions}

here is the question:
{question}
    Attention: I use the two different objects to show the position of human respect to them and the description
describe the same position.
    Attention: do not change left-front,left-back,right-front,right-back to left,right,front or back because they are
different.
"""

```

The output sentences with the fixed sentence pattern, which include the information of the human spatial enhanced query and context of the objects, will be inserted into the {question} part to generate the spatial-information-enhanced caption. The following is the template of the {question} part.

```

"""
Rewrite and proofread three short sentences into ONE long sentence to make it CONCISE and suitable for a
conference paper.
the first sentence is :{original caption}
the second sentence is :{ Sentence 1 with fixed sentence pattern }
the third sentence is :{ Sentence 2 with fixed sentence pattern }
"""

```

The LLM we used is Llama 3, which is the latest version of the open-access LLM released by Meta, consisting of two architectures, namely, the 8 billion (8B) and 70 billion (70B) parameter models. Among them, considering the summary and reasoning capabilities of the model, as well as its size, we implement the Llama 3 8B model. The LLM is used to merge all the information and output the final spatial-information-enhanced natural caption.


4. Experiment

4.1 Device and dataset

In our work, the 3D scene represented as a point cloud of the house can theoretically be captured by any device. Thus, to show the convenience of the initial part of our system, we utilize the LiDAR and camera sensors of iPhone 14 Pro and the Scaniverse application from the APP store to generate the 3D scene point cloud of the house.

We utilize Azure Kinect to capture the input video. Azure Kinect is a device that integrates an RGB camera, a spatial microphone array, an orientation sensor, and a depth sensor operating on time-of-flight technology. The device, the sensors used, and the modes of the sensors are shown in Table 1. In the table, “MJPG_720P” means that we use Motion-JPEG as the video compression format with a 720p resolution for the color image. “WFOV_2x2BINNED” means

Table 1
Azure Kinect device and its setting.

Azure Kinect Device	Related Sensor	Mode and Parameter
	RGB camera	MJPG_720P
	Depth sensor	WFOV_2x2BINNED

that the device is used in the Wide FOV mode, where the camera combines adjacent sensor pixels into a bin, effectively reducing the resolution but increasing the FOV. We use the device to capture two types of data. The first type of data consists of RGB-D image pairs without any people. This type of data is used to generate the partial 3D scene fragment that is used to calculate the transformation matrix between the 3D scene point cloud of the house and the 3D scene fragment in order to determine the relationship between the camera coordinates and the world coordinates. The second type of data consists of RGB-D image pairs that include a person and is used to design the spatial query.

In the evaluation dataset, we use two indoor rooms as our test rooms. In each room, we set up four camera positions to capture the video data. In each video, the participant is instructed to stand in different positions in front of the camera. We present the layout of the two houses, generated by the 3D object detection results of the houses along with their corresponding 3D transformation matrices to the camera coordinates, and the videos captured by Azure Kinect for two human participants to obtain the ground truth caption. We ask them to describe the positions and actions of the participant in the video. An example of the video and the generated 2D layout are shown in Fig. 4 where the dotted line represents the position and FOV of the camera.

4.2 Knowledge base of surrounding environment

We use the Scaniverse app to obtain the point cloud and use VoteNet from MMdetection3D to obtain the 3D object detection result. We input the point cloud to VoteNet, which is pretrained and evaluated on the ScanNet dataset, which includes daily indoor objects. Here, we utilize the checkpoint weights where the AP@0.25 metric is 62.34 and the AP@0.5 metric is 40.08, which are sufficiently accurate for our system. The detection results of the two rooms are shown in Fig. 5.

Next, we aim to find the camera's position and transform all the information into the camera coordinates. The results of the ICP process are shown in Table 2 and Fig. 6. The fitness and correspondence set size are measures of the alignment and matched points between the source and target point clouds, respectively. The inlier root mean square error (inlier *RMSE*) measures the average distance error between the corresponding points. We utilize the volume and position values of the objects and the corresponding transformation matrix to build the knowledge base of the surrounding environment.

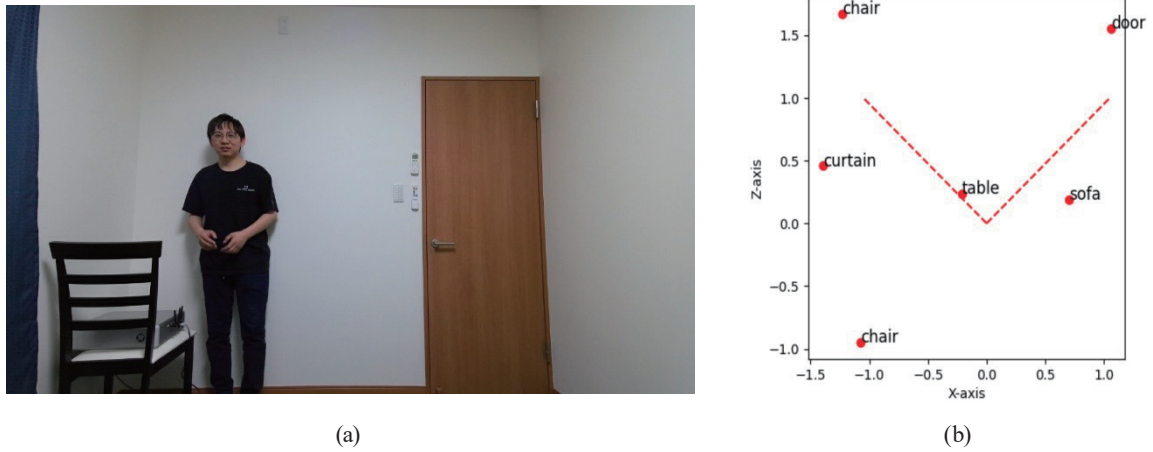


Fig. 4. (Color online) (a) Test video and (b) 2D layout of the house with the corresponding camera position.

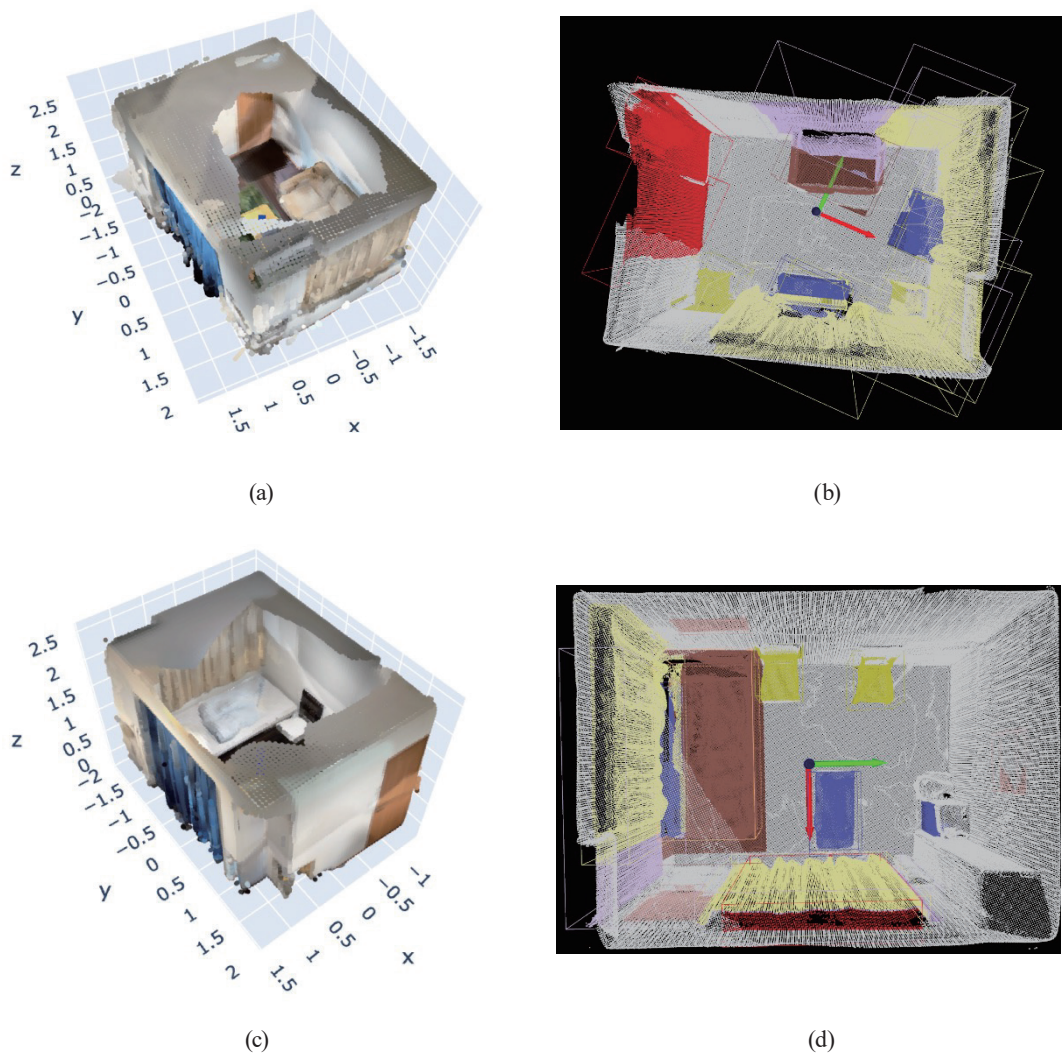


Fig. 5. (Color online) (a–d) 3D scene reconstruction and detection results of the two rooms. The point clouds are all generated by iPhone 14 Pro with the Scaniverse App and the detection result from VoteNet.

Table 2
Quality of registration result.

	Room 1	Room 2
Fitness (avg.)	0.122726	0.115741
Inlier <i>RMSE</i> (avg.)	0.0175551	0.0177646
Corresponding Set Size (avg.)	8694.25	7818.75

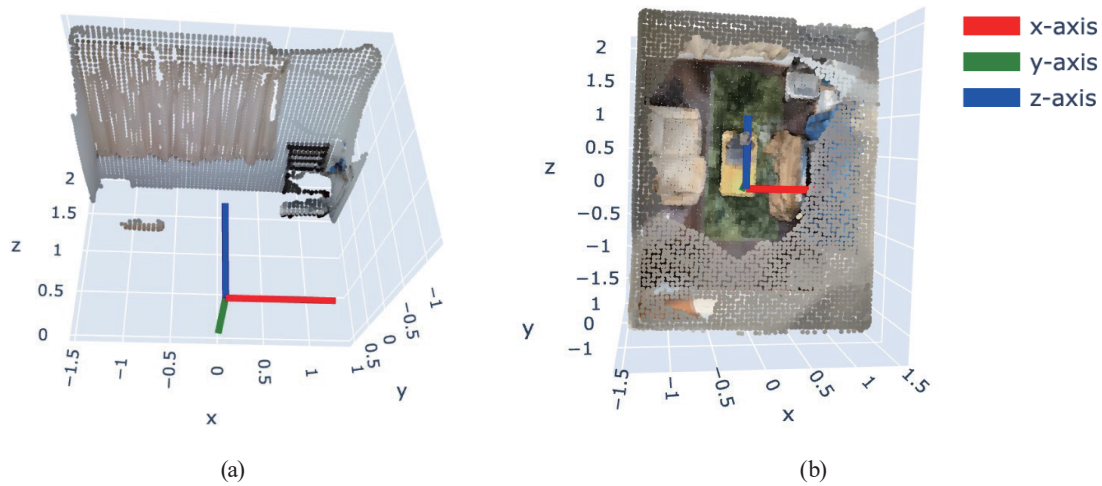


Fig. 6. (Color online) (a) 3D fragment captured with Azure Kinect in the camera coordinates. (b) 3D scene transforms and registers to the corresponding camera coordinates.

4.3 Spatial-information-enhanced query

The spatial-information-enhanced query consists of the original caption generated by the PDVC method and the person 3D position in the camera coordinate system.

We use the pretrained PDVC model, which was trained on the ActivityNet dataset and achieves scores of 31.14 on CIDEr, 8.37 on METEOR2018, 55.79 on Recall, and 57.39 on Precision, to generate the original caption.

To obtain the 3D position of an individual within the house in the 3D camera coordinate system, we first detect the person using YOLOv8 on the RGB image. Next, we align the YOLOv8 human detection results to the depth data and transform the coordinates from the pixel space to the 3D camera coordinate system.

In the first step, we use YOLOv8 to obtain the position of the person. We utilize the pretrained checkpoints of YOLOv8 on the COCO8 dataset, which achieves a mAP@50 metric of 0.872, which is accurate enough for our system.

In the second step, the alignment process adjusts the size of the RGB-D image pair since the FOV of the color image differs from that of the depth image. This adjustment is performed using the point cloud generation procedure of the Open3D library. The transformation process, which converts the person detection results from pixel coordinates to camera coordinates, is carried out using Eq. (6). Since both processes are standard steps in RGB-D data processing, we provide a qualitative result for the alignment of the YOLOv8 human detection results from RGB to depth data, as shown in Fig. 7. The pixel coordinates, along with depth values, are then used to



Fig. 7. (Color online) (a) Human detection result on RGB image. (b) Human detection result on aligned depth image.

calculate the person position in the camera coordinate system through the transformation process.

4.4 RAG-enhanced LLM module

We compare the effect of our system with PDVC, GVL and SG-PDVC⁽⁶⁾ with TSP features in the same human action class in the custom dataset and evaluate the localization and dense video caption performance characteristics by qualitative and quantitative analyses.

Qualitative analysis. The results are shown in Table 3. From the results, we can see that both PDVC and GVL do not pay attention to the surrounding environment and cannot include spatial information in the final caption. The SG-PDVC proposed by Chen *et al.*⁽⁶⁾ can enhance the caption with spatial information, but it is limited by the FOV of the camera. It can only output the person position relative to the chair with words like “first” or “second,” which are unnatural. The captions generated by our system can use the information of the two largest objects, the curtain and sofa, in the house beyond the FOV to provide spatial information. Our system also maintains consistency in direction by using the same camera coordination. Additionally, our system detects changes in position and redetermines the time boundary of the event, updating the caption to eliminate ambiguity.

Quantitative analysis. All metrics are calculated across IoU at [0.3, 0.5, 0.7, 0.9] and averaged, as shown in Table 4. To evaluate the localization performance, Precision and Recall metrics are used. Precision is the fraction of relevant instances among the retrieved instances, and Recall is the fraction of relevant instances that are retrieved. Our method achieves the highest Precision score among the compared methods but shows only a slight improvement in Recall, which is still lower than that of GVL. For evaluating the dense video caption performance, the metrics METEOR, BLEU, CIDEr, and ROUGE-L are used. METEOR is calculated on the basis of the harmonic mean of unigram precision and recall, BLEU is calculated by comparing candidate sentences to reference sentences using the geometric mean of n-gram precision, CIDEr is calculated by measuring the similarity between the generated and reference captions, and ROUGE-L is calculated on the basis of the longest common subsequence between the generated and reference captions.

Table 3
(Color online) Results of qualitative comparison.

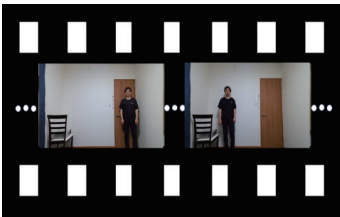
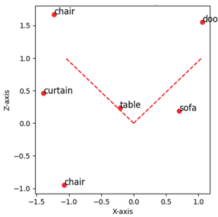
Evaluation Video and Extracted Knowledge		
Ground Truth	[0–20 s] A person stands at the right front of curtain while talking to the camera. [30–40 s] A person is positioned to the right front of the curtain and to the left front of the sofa while talking to the camera.	
PDVC	[3–47 s] A woman is seen talking to the camera.	
GVL	[13–36 s] The man then walks up and down the room and speaking to the camera.	
SG-PDVC	[3–47 s] He is talking to the camera on the right of chair, which is the first chair. [3–25 s] The person is seen speaking to the camera, who is situated at the right-front of both the curtain and the sofa.	
Ours	[25–47 s] The person is seen speaking to the camera, the person is at both the left-front of the curtain and the left-front of the sofa.	

Table 4
Caption and localization performance characteristics.

	GVL	PDVC	SG-PDVC	Ours
BLEU3	7.63	1.81	1.98	7.90
METEOR	6.82	2.32	3.18	7.56
ROUGE_L	11.0	6.18	6.13	8.79
CIDEr	10.8	3.05	3.79	14.9
Recall	82.8	75.0	75.0	77.3
Precision	35.6	33.4	33.4	36.8

Our method achieves significant improvements in CIDEr, slight gains in METEOR and BLEU-3, and a small decline in ROUGE-L compared with GVL. The notable improvement in CIDEr indicates that our system generates captions that are more aligned with human descriptions, which is crucial for accurately interpreting human actions in smart home and security contexts. The slight gains in METEOR and BLEU-3 reflect better word choice and phrasing, which enhance the precision of descriptions. The minor drop in ROUGE-L, which measures sequential overlap, is considered negligible in this context because in smart home and security systems, the emphasis is on accurately conveying key information, which our system achieves despite this small trade-off. Moreover, benefiting from the successful integration of the spatial information of humans and their surrounding environment, our method achieves high scores in all four metrics compared with PDVC, the backbone of our system, and SG-PDVC.

In other words, our system successfully generates spatial-information-enhanced and unambiguous captions of indoor human actions. Compared with previous methods, our approach is more suitable for smart home or indoor security systems where both the position and action of a human are important.

5. Conclusions

In this work, we presented a system RAG- and LLM-based system to enhance the quality of captions generated by the existing methods for smart home and indoor security systems. It incorporates spatial information from the whole-house knowledge base, overcoming traditional FOV constraints in dense video captioning and enabling more accurate, context-aware descriptions. In summary, we utilized 3D reconstruction and detection results to build a knowledge base of the environment. We then designed an RAG procedure to extract relevant information on the basis of spatial cues, enhancing queries for an LLM. Our objective was to ensure that the final caption includes information on the person's location and activity while minimizing ambiguity. We compared our system with PDVC and GVL, two popular dense video caption methods, as well as a similar method proposed in our previous work, using a custom dataset. According to the results, our method outperforms the three compared methods in BLEU-3, METEOR, CIDEr, and Precision metrics, but shows a small decline in ROUGE-L and Recall metrics compared with those of GVL. These results demonstrate that our system successfully generates spatial-information-enhanced and unambiguous captions of indoor human action.

6. Limitation and Future Work

Our method utilizes the PDVC system as the backbone, so the result heavily depends on the performance of the PDVC, which is a complex and large model designed for dense video captioning. However, in our application scenario, we only need to recognize indoor human actions, providing significant leeway to design a more accurate and efficient backbone. In the future, we plan to propose a new model that is smaller and more precise, making our system a more accurate and efficient system for the dense video caption of indoor human actions.

References

- 1 V. Iashin and E. Rahtu: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (IEEE, 2020) 958–959.
- 2 T. Wang, H. Zheng, M. Yu, Q. Tian, and H. Hu: IEEE Trans. Circuits Syst. Video Technol. **31** (2020) 1890.
- 3 A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid: Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (IEEE/CVF, 2023). <https://doi.org/10.1109/CVPR52729.2023.01032>
- 4 R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE/CVF, 2022). <https://doi.org/10.1109/CVPR52688.2022.01589>
- 5 Y. Tu, C. Zhou, J. Guo, S. Gao, and Z. Yu: Pattern Recognit. **111** (2021) 107702. <https://doi.org/10.1016/j.patcog.2020.107702>
- 6 B. Chen, Y. Nakamura, S. Fukushima, and Y. Arakawa: Proc. IEEE Int. Conf. Robotics, Control and Automation (IEEE, 2024)
- 7 A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid: Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2023) 10714–10726. <https://doi.org/10.1109/CVPR52729.2023.01032>
- 8 T. Wang, J. Zhang, F. Zheng, W. Jiang, R. Cheng, and P. Luo: arXiv (2023). <https://arxiv.org/abs/2303.06378>

- 9 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei: Proc. 2020 Advances in Neural Information Processing Systems (Curran Associates, Inc., 2020) 1877–1901.
- 10 H. Touvron, L. Martin, K. R. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. M. Bikel, L. Blecher, C. Cantón Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. S. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. M. Kloumann, A. V. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom: arXiv (2023). <https://arxiv.org/abs/2307.09288>
- 11 G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, and A. Hauth: arXiv (2023). <https://arxiv.org/abs/2312.11805>
- 12 Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang: arXiv (2023). <https://arxiv.org/abs/2312.10997>
- 13 C. R. Qi, O. Litany, K. He, and L. J. Guibas: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE/CVF, 2019) 9277–9286. <https://doi.org/10.1109/ICCV.2019.00937>
- 14 A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2017) 5828–5839. <https://doi.org/10.1109/CVPR.2017.261>
- 15 S. Song, S. P. Lichtenberg, and J. Xiao: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2015) 567–576.
- 16 S. Choi, Q. -Y. Zhou, and V. Koltun: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (IEEE, 2015) 5556–5565. <https://doi.org/10.1109/CVPR.2015.7299195>.
- 17 J. Park, Q. -Y. Zhou, and V. Koltun: Proc. IEEE Int. Conf. Computer Vision (ICCV) (IEEE, 2017) 143–152. <https://doi.org/10.1109/ICCV.2017.25>.
- 18 Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang: arXiv (2023). <https://arxiv.org/abs/2308.03281>
- 19 J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2018) 7190–7198. <https://doi.org/10.1109/CVPR.2018.00751>
- 20 H. Alwassel, S. Giancola, and B. Ghanem: Proc. IEEE/CVF Int. Conf. Computer Vision (IEEE/CVF, 2021) 3173–3183. <https://doi.org/10.1109/ICCVW54120.2021.00356>
- 21 D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri: Proc. IEEE Int. Conf. Computer Vision (IEEE, 2015) 4489–4497.
- 22 J. Redmon, S. Divvala, R. Girshick, and A. Farhadi: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016) 779–788. <https://doi.org/10.1109/CVPR.2016.91>