

# A Comprehensive Survey of 3D Dense Captioning: Localizing and Describing Objects in 3D Scenes

Ting Yu, *Member, IEEE*, Xiaojun Lin, Shuhui Wang, *Member, IEEE*, Weiguo Sheng, *Member, IEEE*,  
Qingming Huang, *Fellow, IEEE*, and Jun Yu, *Senior Member, IEEE*



Fig. 1: Illustration of a 3D dense captioning task: localizing and describing objects in 3D scenes. The task involves the combined process of object localization and captioning to generate natural language descriptions for objects in a 3D scene. It takes a 3D point cloud source as input and produces a diverse range of bounding boxes along with multiple descriptions.

**Abstract**—Three-Dimensional (3D) dense captioning is an emerging vision-language bridging task that aims to generate multiple detailed and accurate descriptions for 3D scenes. It presents significant potential and challenges due to its closer representation of the real world compared to 2D visual captioning, as well as complexities in data collection and processing of 3D point cloud sources. Despite the popularity and success of existing methods, there is a lack of comprehensive surveys summarizing the advancements in this field, which hinders its progress. In this paper, we provide a comprehensive review of 3D dense captioning, covering task definition, architecture classification, dataset analysis, evaluation metrics, and in-depth prosperity discussions. Based on a synthesis of previous literature, we refine a standard pipeline that serves as a common paradigm for existing methods. We also introduce a clear taxonomy of existing models, summarize technologies involved in different modules, and conduct detailed experiment analysis. Instead of a chronological order introduction, we categorize the methods into different classes to facilitate exploration and analysis of the differences and connections among existing techniques. We also provide a reading guideline to assist readers with different backgrounds and purposes in reading efficiently. Furthermore, we propose a series of promising future directions for 3D dense

This work was supported by the National Natural Science Foundation of China under Grant No. 62002314, 62125201, 62020106007, 62022083, 62236008 and Zhejiang Provincial Natural Science Foundation of China under Grant No. LY23F020005. (Corresponding author: Jun Yu.)

T. Yu, X. Lin, and W. Sheng are with the School of Information Science and Technology, Hangzhou Normal University, Hangzhou 311121, China (e-mail: yut@hznu.edu.cn; linxiaojun@stu.hznu.edu.cn; w.sheng@ieee.org).

S. Wang is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangshuhui@ict.ac.cn).

Q. Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: qmhuang@ucas.ac.cn).

J. Yu is with the Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (email: yujun@hdu.edu.cn).

captioning by identifying challenges and aligning them with the development of related tasks, offering valuable insights and inspiring future research in this field. Our aim is to provide a comprehensive understanding of 3D dense captioning, foster further investigations, and contribute to the development of novel applications in multimedia and related domains.

**Index Terms**—3D dense captioning, vision-language bridging, visual captioning, 3D point cloud.

## I. INTRODUCTION

Humans possess a remarkable ability to rapidly recognize and describe various shape details and spatial relationships of objects in unfamiliar scenarios with just a brief glimpse [1]. However, replicating this capability in current computer systems remains challenging. Previous influential research in related fields has predominantly focused on the task of image captioning [2], [3], [4], [5], [6], [7], [8], which involves bridging visual content understanding with natural language description [9], [10] to generate captions that highlight the overall content of the entire image. Subsequently, dense image captioning [11], [12], [13], [14], [15] emerged as a natural extension of image captioning, witnessing a surge of interest in cross-media unified expression facilitated by advances in deep learning technology. Unlike image captioning, dense image captioning places greater emphasis on identifying and expressing local visual details in multiple natural languages.

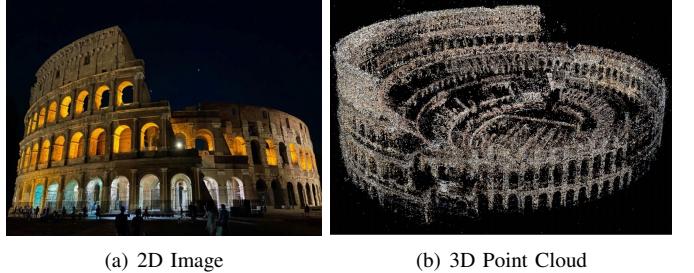
Despite the widespread popularity and significant success of the aforementioned captioning tasks, they are not without limitations. One of the main limitations is that they rely solely on 2D image sources, which inherently have a single viewpoint and can result in misaligned, distorted, and obscured appearances, making it challenging to capture comprehensive

physical details and object location relationships [16]. However, more recently, with the advent of 3D point cloud data collection techniques [17], [18], [19] and pioneering point cloud processing methods [20], [21], [22], there has been a growing interest in 3D cross-modal learning. Unlike 2D images with uniform pixels and grids, 3D point clouds are represented by sparse and disordered points, as illustrated in Fig. 2, and provide rich geometric information, including physical characteristics such as size and shape, as well as spatial relationships between objects from multiple rotatable perspectives. This property of 3D point clouds significantly compensates for the limitations of 2D visual data. Consequently, Chen *et al.* [23] proposed 3D dense captioning as a novel task for generating dense captions in 3D scenes, elevating the traditional dense captioning task from 2D to 3D and resulting in more customized and detailed descriptions. As a burgeoning vision-language-bridging task, 3D dense captioning targets to jointly localize and describe individual objects in 3D scenes leveraging commodity RGB-D sensors. The input for this task is the point cloud of 3D scenes, and the output includes descriptions for the underlying objects along with specific bounding boxes, as shown in Fig. 1. Furthermore, 3D dense captioning can be further divided into two sub-tasks: object detection and object caption generation.

Due to the success of the pioneering work [23], the research on 3D dense captioning has been rolled out comprehensively and rapidly in the following years. Several notable approaches have been proposed to address key challenges in 3D dense captioning [24], [25], [26], [27], [28], [29], [30]. For instance, MORE [24], SpaCap3D [25], and Scan2Cap [23] employed different relation reasoning modules to enhance the relationships among candidate objects. X-Trans2Cap [26] introduced a multi-modal knowledge transfer network with 2D priors to improve 3D dense captioning. 3DJCG [27] and D3Net [28] focused on the connection between 3D dense captioning and 3D visual grounding [31], [32], [33]. CM3D [29] aimed to incorporate contextual knowledge from point clouds into 3D dense captioning. In contrast, Vote2Cap-DETR [30] proposed a parallel approach that combines localization and captioning, deviating from the traditional *detect-then-describe* pipeline. The development timeline for 3D dense captioning is illustrated in Fig. 3.

Recent advancements highlight the active and ongoing research efforts, striving to overcome challenges and push the boundaries of 3D dense captioning. However, the existing literature lacks a comprehensive survey on the topic of 3D dense captioning, despite its increasing prevalence. Therefore, this paper aims to bridge this gap by providing a comprehensive and insightful overview that bridges past research with future prospects in the field of 3D dense captioning. The overview will encompass critical components, including task introduction, methodology review, and future outlook, with the purpose of offering valuable insights for researchers and practitioners. The major contributions can be summarized as the following three aspects:

- Comprehensive and insightful review: This paper presents the first known survey that offers a comprehensive and insightful review of 3D dense captioning. It covers various



(a) 2D Image

(b) 3D Point Cloud

Fig. 2: Visualization of 3D point cloud and 2D image. The inherent single viewpoint of 2D images inevitably triggers a misaligned, distorted, and obscured appearance. Compared to 2D images with uniform pixels and grids, 3D point clouds are represented by sparse and disordered points, providing comprehensive geometric information, including physical characteristics such as size and shape and prosperous spatial relationships from multiple rotatable perspectives.

aspects, such as task definition, architecture classification, datasets analysis, evaluation metrics, and multi-faceted discussions, providing a holistic understanding of 3D dense captioning.

- Critical analysis of existing architectures: Instead of a chronological order introduction, this paper categorizes the existing methods into different classes, enabling a more beneficial exploration and analysis of the differences and connections among the models. This critical analysis provides valuable insights into the strengths and limitations of current approaches, aiding researchers and practitioners in making informed decisions.
- Proposal of future directions: Drawing upon the challenges identified in the field, this paper proposes a series of promising future directions for 3D dense captioning. These directions are aligned with the developments in related tasks, aiming to inspire future research endeavors and drive further advancements in the field.

The paper is organized into four main sections to ensure a clear and organized presentation. In Section II, a targeted reading guideline is provided to assist readers with different backgrounds and purposes. Section III discusses the four related tasks, including image captioning, dense image captioning, dense video captioning, and 3D visual grounding. In Section IV, we synthesize task definition, main framework, and model classification, the most substantial and crucial components of the paper. Section V introduces 3D dense captioning regarding dataset analysis and evaluation metrics. In Section VI, we analyze the experimental details, including the loss function and model performance. Furthermore, the challenges of past 3D dense captioning techniques are discussed, and potential future innovations are proposed in Section VII. Finally, we conclude this survey in Section VIII.

## II. READING GUIDELINES

This paper aims to provide comprehensive insights into the field of 3D dense captioning, catering to readers with varying

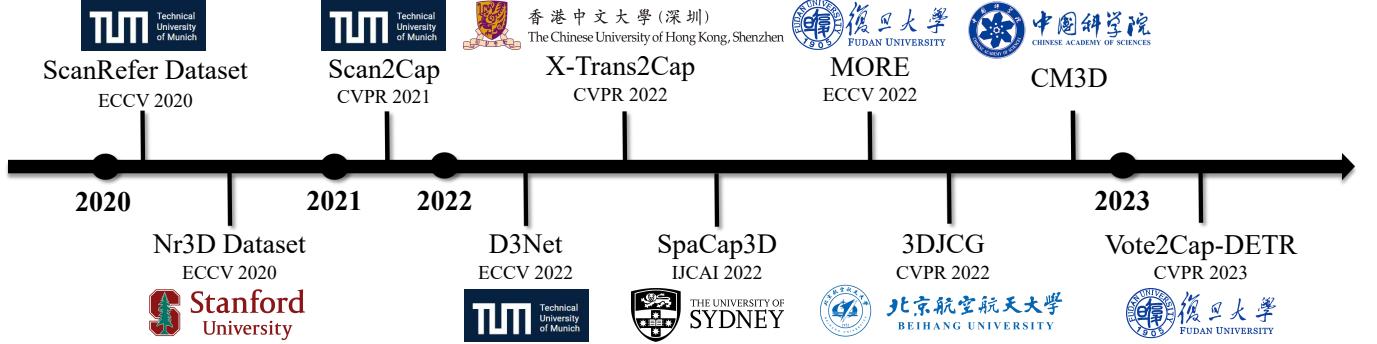


Fig. 3: Evolution of 3D dense captioning: datasets and models over time. Two fundamental datasets ScanRefer [31] and Nr3D [34], have played a pivotal role in shaping the field. The ScanRefer dataset was initially curated for the 3D visual grounding task, and Nr3D is a sub-dataset of ReferIt3D [34], comprising human-annotated 3D scenes. In subsequent years, the field witnessed a surge of novel models, including Scan2Cap [23], D3Net [28], X-Trans2Cap [26], SpaCap3D [25], MORE [24], 3DJCG [27], CM3D [29], and Vote2Cap-DETR [30]. Notably, the order of these models is based on the commit dates on the benchmark, as the exact appearance times of papers were ambiguous.

backgrounds and interests. To optimize the utilization of this paper, the following reading guidance is provided:

- Beginners in the 3D visual-language domain: If readers are new to the field and lack prior experience in related areas, we recommend reading the entire paper section by section. This approach will facilitate a comprehensive understanding of the landscape of 3D dense captioning, as the paper covers relevant literature in detail.
- Researchers familiar with related tasks: Readers with prior experience in tasks such as image captioning or 3D grounding may selectively skip some sections. For instance, the dataset analysis in Section V may be familiar to those working on 3D grounding tasks.
- Experienced researchers of 3D dense captioning: If readers are already well-versed in the field of 3D dense captioning, Sections IV and VII may be particularly meaningful to them. These sections summarize existing models and provide insights into future directions, which may be of interest to researchers with expertise in the field.
- Readers with specific goals: Readers with clear goals or motivations for reading this paper can directly jump to the corresponding section they want to focus on, as each section is relatively independent. This approach allows for quick access to the specific information needed without reading the entire paper.

We hope that this reading guideline will assist readers in maximizing the benefits of this paper and enhancing their understanding of 3D dense captioning, a multimodal learning task in the domain of 3D scene understanding.

### III. RELATED WORK

In this section, we briefly review the most relevant research on 3D dense captioning, especially those representative studies that involve captioning or 3D fields, including image captioning, dense image captioning, dense video captioning, and 3D visual grounding, as illustrated in Fig. 4.

#### A. Image Captioning

As a representative visual-language generation task [35], [36], image captioning aims to provide a descriptive sentence for an input image, with the purpose of facilitating 2D scene comprehension, particularly for individuals with visual impairments [37]. The development of image captioning can be delineated into two distinct phases: the traditional machine learning stage and the advent of deep learning approaches, as discussed in [16], [38]. In the earlier machine learning stage, template-based [39] and retrieval-based [40] methods were commonly employed. However, with the emergence of deep learning [16], researchers have shifted their focus towards leveraging advanced technologies, such as attention mechanisms [8], [6], [41], graph neural networks [42], [43], [44], convolutional networks [45], [46], [47], transformers [9], [48], [49], and Vision-Language Pre-training (VLP) models [50], [51]. Typically, the core encoder-decoder framework has been popularly used in existing image captioning approaches. Convolutional Neural Networks (CNNs) [52], [53], [54] were employed as encoders to map input images into feature vectors, while Recurrent Neural Networks (RNNs) [55], [56] were utilized as decoders to generate sentences from the feature vectors. Furthermore, some approaches [49], [57], [58], [59], [60] incorporated object detectors to enhance the extraction of visual features. Attention mechanisms [49], [61], [62] and graph neural networks [44], [63] have gained increasing popularity in recent years due to their ability to capture fine-grained visual details and contextual relationships. Notably, methods incorporating VLP models, such as ConZIC [64], have demonstrated efficient performance in image captioning. As a self-correction framework, ConZIC integrated BERT [10] and CLIP [50], enabling controllable zero-shot image captioning with significantly improved generation speed compared to previous works.

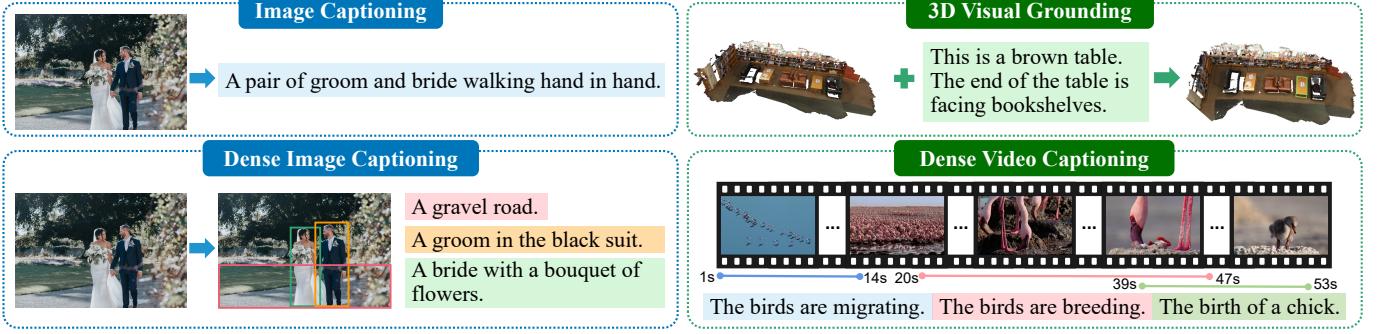


Fig. 4: Illustration of representative visual-language bridging tasks related to 3D dense captioning. Both image captioning and dense image captioning involve using a 2D image as input to generate natural language descriptions. However, they differ in their objectives. Image captioning aims to generate a sentence that describes the overall content of the input image. On the other hand, dense captioning, as a variation of image captioning, focuses on generating distinctive descriptions for prominent regions within an image, capturing fine-grained details. 3D visual grounding is a task that emphasizes localizing the object described by an input sentence within a 3D scene. It involves linking the language-based description to the corresponding 3D object in the scene, bridging the gap between visual perception and natural language understanding. Dense video captioning involves localizing significant events in an untrimmed video and generating captions for each event. Among the four tasks mentioned, dense image captioning exhibits the closest resemblance to 3D dense captioning. Conversely, 3D visual grounding stands in contrast to 3D dense captioning, emphasizing different aspects of visual understanding and language integration.

### B. Dense Image Captioning

As a specialized type of image captioning, dense image captioning focuses on generating separate descriptions for each prominent region or object within an image, closely related to 3D dense captioning [23]. The advent of [12] in 2016 marked a significant milestone in the field of dense image captioning. In contrast to global-based image captioning, which provides a single caption for the entire image, dense image captioning generates more detailed and region-level descriptions for multiple objects or regions in the image. To address the challenge of incorporating non-object context in prior work, Yang *et al.* proposed a context fusion module to generate more accurate and contextually relevant descriptions [13]. Context fusion module helps to improve the quality of the generated descriptions by considering the surrounding visual elements beyond individual objects. Furthermore, Song *et al.* presented a semantically symmetric LSTM [55] model combining dense image captioning with scene recognition models [65]. It leveraged the relationship between scene understanding and image captioning to generate more meaningful and coherent captions that capture both the local object-level details and the global scene context. More recently, there has been a surge of interest in multi-modal pre-trained methods for joint dense image captioning and object detection, leading to a new wave of research in this area. Prominent models [66], [67] integrated object detection and dense captioning into a unified framework for improved performance and computation efficiency. These approaches leveraged pre-trained models that incorporate multiple modalities, such as images and text, to jointly learn the representations for both tasks, leading to promising advancements in the field of dense image captioning.

### C. Dense Video Captioning

Building upon the concept of dense image captioning, Krishna *et al.* introduced a more challenging task known as dense video captioning [68]. Unlike traditional video captioning approaches [69], [70], [71], which generate a single caption for an entire video, dense video captioning [72], [73], [74], [68] involves describing multiple events within a minute-long video. This task proves beneficial for the search and indexing of untrimmed, large-scale videos. The majority of dense video captioning techniques [75], [74], [76], [77] follow a two-stage strategy. It begins with event localization [78], [79], where events of interest within the video are identified, followed by event captioning [80], [69] to generate descriptive captions for these localized events. These two-stage approaches bear a resemblance to the process employed in 3D dense captioning. To enhance the interaction between event localization and captioning, certain approaches [72], [81], [82] propose performing these subtasks jointly. In particular, Li *et al.* [72] employed a descriptiveness regression technique to dynamically adjust the temporal positions of individual event candidates and enable the integration of event proposal localization and event caption generation, resulting in a unified framework. Conversely, other works [83], [84] eliminate explicit event localization and instead ground each sentence in the video after generating a comprehensive video description paragraph. With the emergence of multi-modal pre-trained models [85], [86], [87], [70], Yang *et al.* have recently explored integrating end-to-end dense video captioning into these large-scale pre-trained models [88], leveraging their capabilities for improved performance and scalability.

### D. 3D Visual Grounding

3D visual grounding focuses on localizing objects in 3D scenes based on their textual descriptions. It is closely related

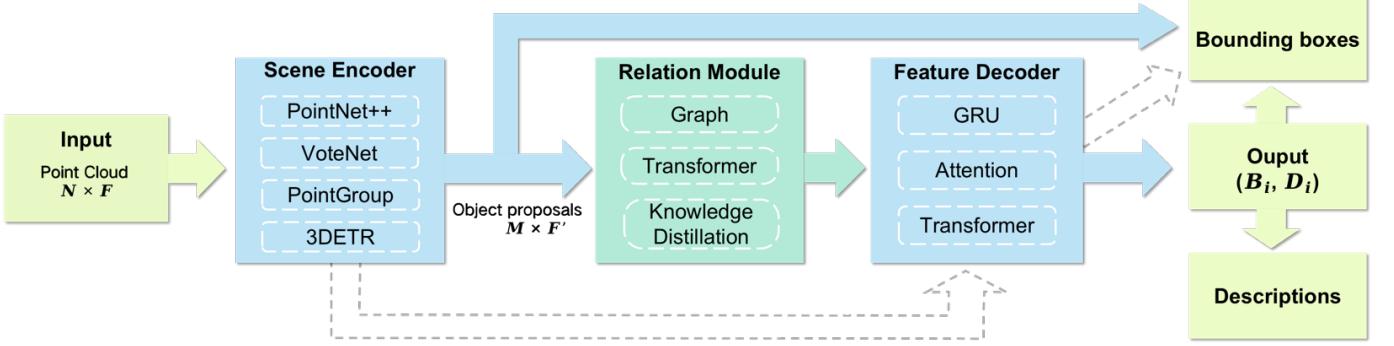


Fig. 5: The flowchart of the general framework for 3D dense captioning, which typically encompasses three main components: a scene encoder, a relation module, and a feature decoder. The relation module is a core component that is commonly employed in most existing works, rendering it an integral part of the encoder-decoder structure. Specifically, the input to the framework is a 3D point cloud containing  $N$  randomly sampled points, each characterized by  $F$ -dimensional features. The model then generates  $I$  pairs of bounding boxes along with multiple descriptions  $(B_i, D_i)$  as outputs. In the intermediate process, most models generate  $M$  object proposals with  $F'$ -dimensional feature representations using a scene encoder. These proposals are then utilized to learn the relationships between objects in the relation module. Finally, the feature decoder generates corresponding descriptions for the objects. Notably, a smaller portion of methods do not follow this approach and instead perform detection and captioning simultaneously in the final feature encoding step.

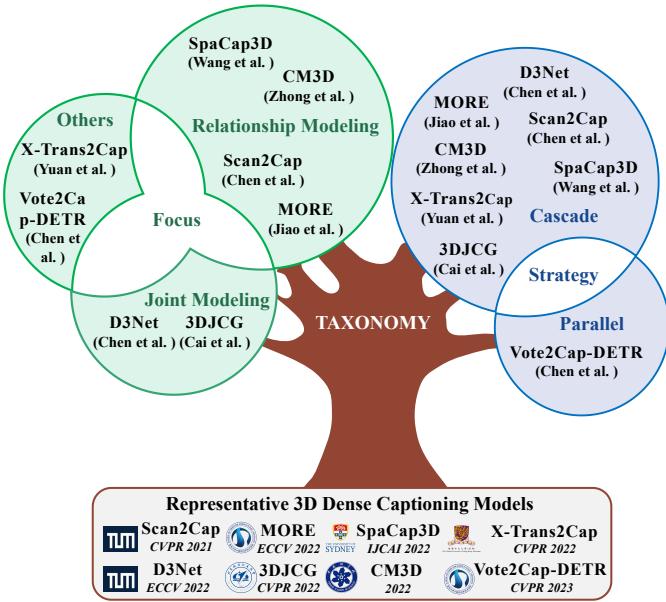


Fig. 6: Classification of models in the context of 3D dense captioning. The existing approaches can be categorized based on their research focus and research strategy. With regards to the specific research motives and foci, most methods can be broadly classified into three groups: relationship modeling that focuses on building inter-object or context relationships; research on joint modeling that combines two distinct tasks; and other approaches that do not fall into these two categories. Regarding the research strategy employed to tackle the 3D dense captioning task, existing models can be categorized into the “detection-then-captioning” cascade strategy and the “detection-and-captioning” parallel strategy.

to 3D dense captioning but with key differences. Unlike 3D dense captioning generating descriptions for each object in a point cloud, 3D visual grounding aims to locate the object described by a specific input sentence. In analogy to human listening and speaking, 3D visual grounding simulates the process of listening, while 3D dense captioning is more akin to speaking, as described in [28]. Dave *et al.* [31] introduced the dataset, the pioneer method, and the benchmarks for 3D visual grounding. Most contemporary 3D visual grounding methods consist of two phases: object detection [89], [90] and description matching [91], [92], [93], [32]. In the object detection phase, objects were detected and segmented under the related 3D scenes. Subsequently, in the description matching phase, the textual descriptions were matched with the detected objects to identify the referred object in the scene. Early methods for 3D visual grounding relied on object detectors for scene localization. However, Chen *et al.* [28] proposed a novel hierarchical attention model that enables end-to-end training for both detection and grounding [94], providing a unified and integrated pipeline. More recently, the technique of combining grounding tasks with dense captioning experience a growing trend [27], [28], which explored the potential synergies between related tasks and opened up new research directions.

## IV. ARCHITECTURE

### A. Task Definition

3D dense captioning is a task that aims to achieve object-level 3D scene understanding by generating natural language descriptions for objects in 3D scenes through the analysis of 3D visual data [23], [25], [29]. The input to the model is a 3D point cloud, which represents the geometric and appearance characteristics of objects in the scene, along with additional auxiliary features such as colors, normal vectors,

and multi-view features. The point cloud data can be mathematically represented as a matrix  $P \in \mathbb{R}^{N \times F}$ , where  $N$  denotes the number of randomly sampled points per scene (typically 40,000), and  $F$  represents the dimensionality of scene features, including point coordinates  $(x, y, z)$ , and other auxiliary features. Most existing methods generate  $M$  object proposals with  $F'$ -dimensional features prior to generating captions, where  $M$  is typically set to a default value of 256. The text data, comprising the captions, is tokenized using the SpaCy library [95] and represented as a matrix  $W \in \mathbb{R}^{T \times 300}$  using GloVe word embeddings [96], where  $T$  represents the number of tokens in the caption, and each word embedding vector has a dimension of 300. During the training phase, the ground truth caption represented by  $W$  is utilized to optimize the generated word token probabilities using a loss function (detailed discussion in Section VI). During the inference phase, only the point cloud data  $P$  is fed into the model. The expected output is a set of bounding boxes with corresponding captions  $(B_i, D_i)$ , where  $B_i$  denotes the bounding box coordinates and  $D_i$  represents the generated description for each object.

### B. Main Framework

The realm of 3D dense captioning has been predominantly influenced by the encoder-decoder architecture, which can be delineated into three principal components: scene encoder, relation module, and feature decoder, as depicted in Figure 5. The scene encoder is responsible for extracting initial scene details, such as object-level visual features and contextual information from input point clouds with various 3D object detection methods, such as PointNet++ [20], VoteNet [21], PointGroup [97], 3DETR [98], and others [89], [90]. The relationship module serves as a pivotal component in most 3D dense captioning models to model intricate connections within the scenes or cross-modal interactions. A well-designed relational module has been empirically shown to significantly enhance model performance, as substantiated by numerous ablation experiments. Feature decoder commonly employs sequential models such as GRU [99] or Transformer [9] to further decode the attribute information and relation tokens, thereby generating pertinent captions as well as bounding boxes for the target objects. The stage of decoding is essential for generating precise and meaningful captions that accurately describe the objects in the 3D scenes.

1) *Scene Encoder*: The scene encoder is tasked with extracting initial scene details. Previous research efforts [23], [24], [25], [26], [27], [29] have primarily utilized VoteNet or modified-VoteNet as the feature extraction backbone to obtain comprehensive visual features. The composition of the visual features may vary depending on the specific approach. For instance, CM3D [29] captures background environmental details, while X-Trans2Cap [26] focuses on object attributes. In recent approaches, more robust detectors such as PointGroup [97] and 3DETR [98] have been employed as the feature extraction backbone. For example, D3Net [30] adopts PointGroup with U-Net architecture [100], while Vote2Cap-DETR [30] utilizes a full transformer structure for 3DETR. These

approaches have achieved notable performance in ablation studies, underscoring the significance of employing effective object detectors in the scene encoding stage. Additionally, most approaches[23], [24], [25], [26], [27], [28], [29] also aggregate the object proposals and bounding boxes at this stage, while some [30] handle this task in the decoding step, depending on the specific approach.

2) *Relation Module*: The relation module enables the modeling of intricate connections and cross-modal interactions within indoor scenes. The choice of relationship modeling approach depends on the specific requirements of the 3D dense captioning task and the characteristics of the scene understanding problem being addressed. There are several commonly used approaches for modeling relationships, including graph-based methods, transformer-based approaches, and knowledge distillation. Methods like Scan2Cap [23], MORE [24], and D3Net [28] utilize semantic scene graphs [101], [102] to capture inter-object spatial location relationships. In these methods, object proposals are treated as nodes in the graph, and the relationships between objects are modeled as edges connecting the nodes. For instance, relationships such as top, front, left, or center can be learned between objects with a graph structure. These approaches also leverage the transitive property of relationships, where the relationship between two non-adjacent nodes can be inferred from a common node. This allows for reasoning about relationships between objects that are not directly connected in the graph. 3DJCG [27] and SpaCap3D [25] employ transformer-based modules to build inter-object relationships. Specifically, 3DJCG utilizes a feature enhancement module comprising multi-head self-attention layers and fully connected layers to model inter-object relationships and enhance attribute characteristics. SpaCap3D uses a standard transformer encoder with a relation prediction head to capture object-to-object relations. These approaches leverage the self-attention mechanism of transformers to model relationships between objects and capture long-range dependencies in the scene. Knowledge distillation is another approach applied in some models, such as X-Trans2Cap [26], which applies a knowledge distillation framework with a cross-modal fusion module to facilitate interaction between 3D object features and multiple modalities.

3) *Feature Decoder*: Feature decoder involves generating bounding boxes and captions for candidate objects, incorporating the attribute and relationship features learned in previous stages. However, in addition to Vote2Cap-DETR, most of the models only perform caption generation at this stage. Scan2Cap, MORE, and D3Net employ a GRU-based decoder with attention mechanisms to generate descriptions for candidate objects. Other methods, such as SpaCap3D, X-Trans2Cap, 3DJCG, and Vote2Cap-DETR, apply a transformer-based decoder for caption generation. Specifically, X-Trans2Cap utilizes a transformer decoder for both the teacher and student frameworks, directly incorporating the transformer architecture for caption generation [26]. SpaCap3D introduces an object-centric decoder with an improved masked self-attention mechanism [25]. 3DJCG utilizes a multi-head cross-attention network following a fully connected layer with a language prediction module as the caption head [27]. Vote2Cap-DETR

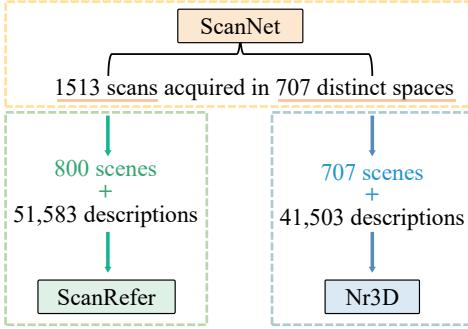


Fig. 7: Illustration of relationship between ScanNet [19], ScanRefer [31], and Nr3D [34]. ScanRefer and Nr3D are proposed with extra human-labeled descriptions based on the scenes in ScanNet.

employs two identical standard transformer decoders, a position embedding, and a linear classification head for generating descriptions [29]. Notably, CM3D stands out by stacking multiple decoder layers for caption generation. This allows CM3D to capture richer contextual information for generating more detailed and coherent captions.

### C. Model Classification

The existing approaches in the field of 3D dense captioning can be categorized based on their research focus and research strategy. With regard to the specific research motives and foci, these methods can be broadly classified into three groups: relationship modeling, joint modeling, and other approaches that do not fall into these two categories. Regarding the research strategy employed to tackle the 3D dense captioning task, existing models can be classified into two categories: the “*detection-then-captioning*” cascade strategy and the “*detection-and-captioning*” parallel strategy. The classification of models with the context of 3D dense captioning is illustrated in Fig. 6.

1) *Research Focus*: The research focus of 3D dense captioning has been explored from various perspectives, including research on relationship modeling, joint modeling, and other approaches.

**Relationship Modeling.** Scan2Cap is the pioneering method that introduces a graph-based attentive captioning framework to explore object relations. Building upon Scan2Cap, Jiao *et al.* [24] propose an improved model called MORE, which incorporates multi-order relation mining based on graphs. MORE captures more complex inter-object relationships through progressive encoding. In contrast, ScanCap3D focuses specifically on capturing spatial relativity in 3D scenes using a spatiality-guided encoder-decoder transformer architecture. ScanCap3D achieves this while maintaining a simpler size and higher computational efficiency compared to Scan2Cap and MORE. In contrast to previous works [23], [24], [25] that mainly investigate inter-object interactions, neglecting contextual details, CM3D [29] addresses this limitation by incorporating rich contextual information. CM3D considers non-object details, background environments, and generates more detailed de-

TABLE I: The statistics of the ScanRefer and Nr3D datasets. Both datasets comprise primarily 3D scenes, object labels, and manually annotated descriptions. However, the notable differences lie in their marking strategies. ScanRefer labels almost all objects in a scene, while Nr3D focuses on labeling specific categories that appear with higher frequency.

Number #	ScanRefer	Nr3D
Descriptions	51,583	41,503
Scenes	800	707
Objects / Contexts	11,046	5,878
Object classes	265	76
Objects / Contexts per scene	13.81	8.31
Descriptions per object	4.67	7
Average length of descriptions	20.27	11.4

scriptions, thus providing a more comprehensive understanding of the scene.

**Joint Modeling.** D3Net introduced unified networks for both 3D dense captioning and 3D visual grounding tasks in the context of 3D scenes. It originates from the same team that developed the Scan2Cap model for 3D dense captioning and utilized the ScanRefer dataset for 3D visual grounding [31]. In D3Net, the overlapping aspects of these two tasks are integrated to create a unified model. This integration is motivated by the challenges posed by limited vision-language data, which can lead to overfitting, as well as the difficulty of generating unique descriptions for similar objects. To tackle these challenges, a joint speaker-listener architecture is introduced, where the “speaker” corresponds to the captioning model and the “listener” corresponds to the grounding task. This architecture enables semi-supervised training and facilitates the generation of distinctive descriptions [28]. Similarly, 3DJCG [27] also explores the joint framework of 3D visual grounding and 3D dense captioning. The goal of this framework is to capture visual and textual information efficiently and effectively for generating comprehensive 3D captions [27]. By identifying the complementary relationship between the two tasks, 3DJCG proposes an innovative transformer-based approach that incorporates three key components: a 3D object detector, a feature enhancement module, and a grounding or captioning head.

**Others.** To address the challenge of enriching 3D point clouds with 2D information while minimizing computational overhead, Yuan *et al.* propose a transformer-based cross-modal teacher-student framework in X-Trans2Cap [26]. This framework utilizes knowledge distillation [103] to transfer rich appearance information, including texture awareness and color clues, from 2D images to 3D scenes. By employing the teacher-student paradigm, X-Trans2Cap effectively integrates 2D information into the 3D captioning process without imposing a significant computational burden [26]. In a more recent development, Vote2Cap-DETR adopts a fully end-to-end transformer encoder-decoder architecture to enable parallel detection and captioning [30]. This approach eliminates the need for separate detection and captioning stages, resulting in a more efficient and integrated processing framework.

2) *Research Strategy*: Regarding the research strategy employed to tackle the 3D dense captioning task, existing models can be categorized into two main categories. The first category is a cascade strategy, where object detection is followed by captioning, namely the “*detection-then-captioning*” cascade strategy. The second category is a parallel strategy, where object detection and captioning are addressed simultaneously, namely the “*detection-and-captioning*” parallel strategy.

TABLE II: The statistics of the standard split of ScanRefer dataset.

Number #	Train	Val	Test	Total
Descriptions	36,665	9,508	5,410	51,583
Scenes	562	141	97	800
Objects	7,875	2,068	1,103	11,046
Objects per scene	14.01	14.67	11.37	14.14
Descriptions per scene	65.24	67.43	55.77	65.68
Descriptions per object	4.66	4.60	4.90	4.64

**Cascade Strategy.** In the case of the cascade strategy, most methods follow a “detect-then-caption” paradigm. The model first generates a set of candidate objects with their bounding boxes and then performs feature enhancement and relationship inference around these candidate objects. Subsequently, corresponding descriptions are generated based on these object-centric features. However, this cascade strategy is not without its limitations [30]. On the one hand, the performance of captioning is heavily influenced by the accuracy of the detection results. On the other hand, the hand-crafted components designed in previous detectors may result in poor performance.

**Parallel Strategy.** To overcome these limitations, Vote2Cap-DETR takes a different approach by employing a parallel strategy. They reverse the order of captioning and detection by utilizing 3DETR [98], a successful transformer-based architecture for 3D object detection, as a feature encoder for generating scene tokens. Spatial bias and content-aware features are introduced to refine the initial object queries into more precise vote queries. Subsequently, two independent decoder heads are designed for object location and caption generation simultaneously. This one-step attention-driven strategy helps to mitigate over-reliance on detectors and reduce the hard-coded limitations of object detectors. Additionally, it derives more effective localization and more accurate captioning.

## V. DATASETS AND EVALUATION METRICS

### A. Datasets Analysis

The availability of large-scale data is crucial for achieving superior performance in vision and language tasks. However, capturing and annotating 3D data pose significant challenges, resulting in smaller RGB-D datasets [104], [105], [106], [18] for 3D scenes compared to their 2D counterparts. As a result, researchers resort to utilizing manufactured data to compensate for the absence of real-world data [107], [108]. To address this limitation, Dai *et al.* introduced the ScanNet dataset [19], which is a richly-annotated 3D indoor scan dataset of real-world environments captured employing a self-created RGB-D capture system.

The ScanNet dataset comprises 2.5 million views obtained from 1513 scans, which are obtained from 707 distinct spaces, making it larger in scale compared to other popular datasets [109], [110], [105], [111], [19], such as NYUV2 [109], SUN3D [105], and SceneNN [18]. However, it should be noted that ScanNet is primarily designed for tasks such as 3D object classification, semantic voxel labeling, and 3D object retrieval with instance-level category annotation in a 3D point cloud and may not be directly suitable for 3D dense captioning and 3D visual grounding tasks. To overcome this limitation, ScanRefer [31] and Nr3D [34] were developed as datasets specifically tailored for 3D dense captioning, building on top of ScanNet. The relationship between these datasets is depicted in Fig. 7. Both ScanRefer and Nr3D select partial point cloud scenes from ScanNet and provide additional human-labeled descriptions for the objects in each scene. Furthermore, they also offer online browsing sites for visualizing the data, making them valuable resources for developing and evaluating models in the field of 3D dense captioning and 3D grounding tasks.

**ScanRefer.** ScanRefer is widely recognized as a prominent dataset for 3D dense captioning tasks, offering a substantial number of natural language descriptions for objects in the scans from the ScanNet dataset [31]. Statistically, the dataset encompasses 51,583 detailed and diverse descriptions for 11,046 objects in 800 ScanNet scans, covering over 250 types of common indoor objects and including attributes such as color, size, shape, and spatial relationships. Approximately five manual annotations are provided for each object in each scene, resulting in a rich and comprehensive dataset. The detailed statistics of the ScanRefer are presented in the second column of Table I. An example of a typical scene from ScanRefer, specifically scene 0000\_00, is illustrated in Fig. 8. Following the official split of ScanNet [19], the dataset is partitioned into training, validation, and test sets quantitatively with 36,665, 9,508, and 5,410 samples, respectively. Since the unseeable testing split has not been officially released, most experiments are performed on the validation set [23]. The distribution statistics of the ScanRefer dataset are provided in Table II. Additionally, Fig. 9 presents the frequency of several major object categories in ScanRefer.

**Nr3D.** The 3D Natural Reference (Nr3D) dataset is one of the datasets from ReferIt3D [34], along with the synthetic language descriptions dataset Sr3D [34]. Fig. 10 presents a typical example of scene 0000\_00 from Nr3D. Nr3D focuses on fine-grained objects in 3D space, which makes it more challenging compared to ScanRefer as the reference object is of the same kind [26]. Nr3D consists of 41,503 natural, free-form utterances describing objects belonging to one of 76 fine-grained object classes within 5,878 communication contexts. The communication contexts are expressed as unique sets denoted as  $\{S, C\}$ , where  $S$  represents one of the 707 distinct ScanNet scenes, and  $C$  denotes the fine-grained class of  $S$ . In other words, a context represents the same object type in a scene. Similar to ScanRefer, Nr3D uses the official ScanNet splits for experiments. The statistics of Nr3D are presented in the right column of Table I.

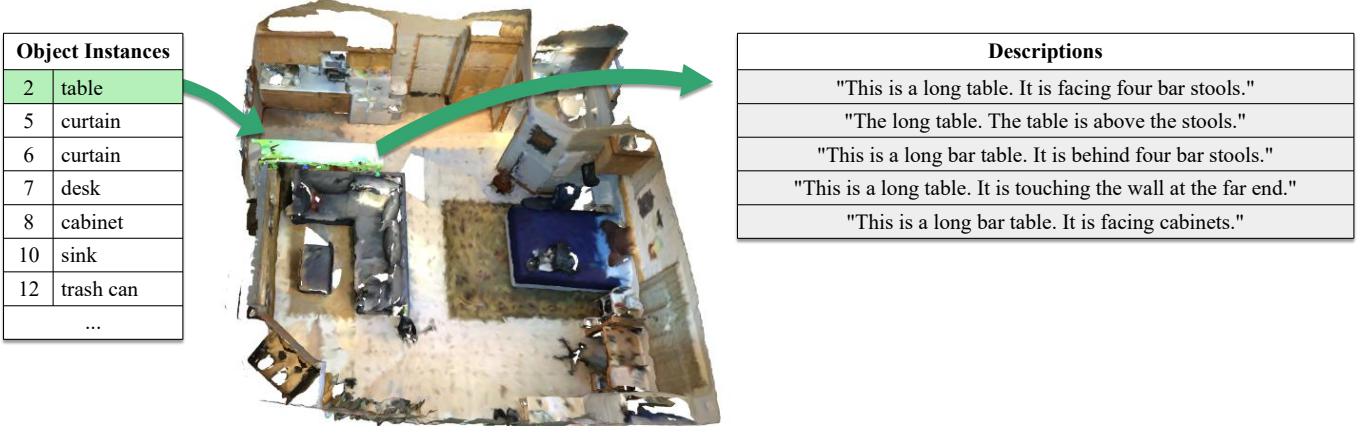


Fig. 8: Illustration of a representative example of scene 0000\_00 from the ScanRefer Dataset [31]. The ScanRefer dataset is known for its comprehensive object labeling approach, wherein nearly every object in the scene is annotated, accompanied by five corresponding descriptions.

TABLE III: Evaluation Metrics for 3D Dense Captioning. The table presents a comprehensive overview of various evaluation metrics employed for assessing the performance of 3D dense captioning models. Notably, CIDEr, being a metric that closely aligns with human assessments, is considered the most significant among the listed metrics.

Metric	Acronym	Based On	Original Task	Advantages	Drawbacks
BLEU-4	B-4	Precision	Machine translation	Simple and efficient	Grammar issues and limited by text length
ROUGE	R	Recall	Machine text summarization	Simple and orderly	Disregard synonyms
METEOR	M	Precision Recall Pen	Machine translation	Consider stems and synonyms	Reliance on external knowledge sources
CIDEr	C	TF-IDF	Image captioning	Better human assessments	Hard to optimize

### B. Evaluation Metrics

To rigorously assess the effectiveness of various methods for the 3D dense captioning task, performance evaluation is typically conducted from both captioning and detection perspectives, following established standards in the field. Captioning performance is evaluated using widely used text generation metrics, including CIDEr [112], BLEU-4 [113], METEOR [114], and ROUGE [?], denoted as  $C$ ,  $B-4$ ,  $M$ , and  $R$  respectively. Additionally, Chen *et al.* proposed a novel evaluation metric [23] to calculate Intersection-over-Union (IoU) scores between the ground-truth bounding boxes and the predicted bounding boxes, formulated as follows:

$$E@kIoU = \frac{1}{N} \sum_{i=0}^N E_i U_i ; U_i = \begin{cases} 1, & IoU \geq k \\ 0, & IoU < k \end{cases}$$

where  $E$  denotes the specific evaluation metrics, including CIDEr, BLEU-4, METEOR, and ROUGE (denoted as  $C$ ,  $B-4$ ,  $M$ , and  $R$ , respectively). The number of detected object bounding boxes is denoted as  $N$ . The value of  $U_i$  is utilized to determine the IoU threshold, and  $k$  is commonly set to 0.5 or 0.25. Additionally, the object detection metric employs the Mean Average Precision (mAP) thresholded by IoU, providing a comprehensive assessment of the detection performance in localizing objects.

In the past, image captioning evaluation heavily relied on machine translation evaluation metrics such as BLEU [113], ROUGE [115], and METEOR [114], which lack relevance to the human assessment. More recently, with the introduction of CIDEr [112] and SPICE [116] indicators designed explicitly for image captioning, this issue has been addressed. In 3D dense captioning, the metrics BLEU, ROUGE, METEOR, and CIDEr are selected for captioning assessment. We provide a detailed introduction to these four metrics below and organize them in Table III.

1) *BLEU*: Bilingual Evaluation Understudy (BLEU) is a well-established metric for evaluating machine translation outputs based on precision. It involves matching different parts of a candidate text with reference texts and then calculating the percentage of successfully matched sequences. However, the basic calculation of BLEU is limited to unigram precision, which matches individual words. Many modern approaches adopt modified n-gram precision, which considers sequences of  $n$  consecutive words. For instance, in the context of 3D dense captioning, BLEU-4 is commonly used, which considers four consecutive words in the generated text as one matching unit. One limitation of BLEU is that it may not be able to detect syntactic issues, as successful sequences of matches may appear in the wrong order. This can result in inaccurate

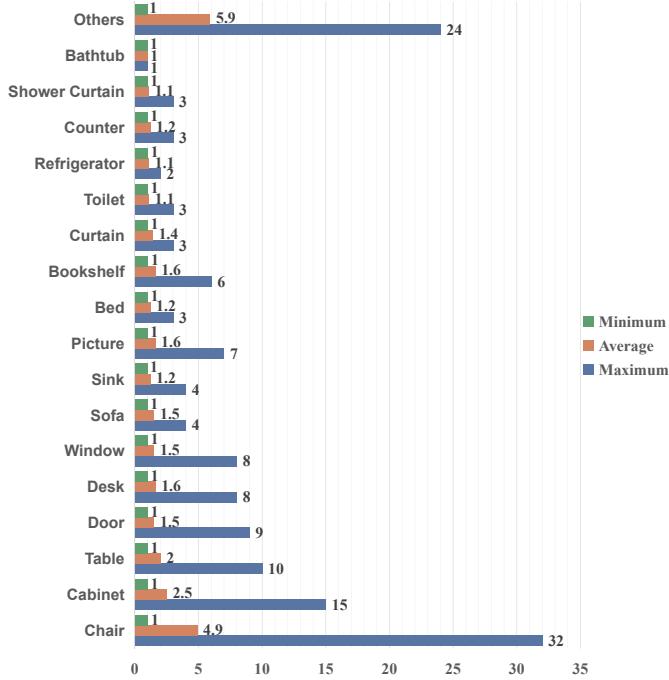


Fig. 9: Illustration of the manifestation of 18 broad categories comprised of 256 sub-categories in a 3D indoor scene, primarily encompassing furniture items such as chairs, tables, and cabinets. “Maximum”, “Average”, and “Minimum” indicate the maximum, average, and minimum number for a certain type of object that appears in a scene. For example, there are up to 32 chairs in a scene.

scores. Additionally, shorter generated texts may have higher chances of being incomplete, leading to unreliable BLEU scores [117], [118]. Therefore, BLEU tends to favor candidate texts that have a similar length to the reference captions to mitigate this issue.

2) *ROUGE*: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a generally employed metric for assessing the quality of text summarization generated by machine learning algorithms. It is based on recall and measures the similarity between the candidate and reference summaries with various n-gram-based and sequence-based statistics. ROUGE has several variants, including ROUGE-N, ROUGE-W, ROUGE-L, and ROUGE-S, built based on different considerations. ROUGE-N estimates the n-gram recall between the candidate and reference summaries, where n denotes the length of the n-gram. ROUGE-W considers the weighted longest common subsequence, considering the essence of words in the sequences. ROUGE-L focuses on calculating the sentence-level structure similarity, performing the statistics of the longest common subsequence and longest co-occurrence in the n-grams sequence. ROUGE-S incorporates skip-bigram co-occurrence statistics, allowing for measuring the similarity of non-consecutive words. It is worth mentioning that although ROUGE is widely adopted for evaluating text summarization tasks, it may not perform well in the context of multi-document

text summaries. Thus, it is essential to carefully consider the appropriate variant of ROUGE for the specific task being evaluated.

3) *METEOR*: Metric for Evaluation of Translation with Explicit ORdering (METEOR) is employed to measure the machine-translated language. The metric is based on the harmonic mean of unigram precision and recall. Moreover, it not only matches candidate sentences with standard reference captions but is also equipped with stemming and synonym-matching capabilities. To a certain extent, METEOR makes up for the shortcomings of the above two indicators, reflecting grammar and sentence fluency. However, it relies heavily on external knowledge sources and only considers unigram, which may not capture the nuances of higher-order language structures or semantic meanings.

4) *CIDEr*: In contrast to previous metrics, which often lack correlation with human consensus, CIDEr incorporates Term-Frequency Inverse Document Frequency (TF-IDF) weighting for each n-gram, resulting in more accurate human assessments [119]. Notably, CIDEr disregards n-grams that do not appear in the reference sentence and instead encodes the frequency of n-gram occurrences in the reference statement. N-grams that occur more frequently are assigned lower weights, as they are less likely to contain essential information, such as common phrases like “*this is a*”. As a result, CIDEr has been deemed as a pivotal metric for evaluating the quality of 3D dense captioning. Despite its significance, CIDEr does possess a limitation in terms of optimization. The optimization process for CIDEr can be challenging due to its intricate weighting schemes based on TF-IDF, and the scoring may not always align perfectly with human judgments.

## VI. EXPERIMENTAL DETAILS

### A. Loss Function

Most models incorporate three key components, namely detection loss, caption loss, and relative direction loss, in their loss functions. Although we use the same notation for convenience, it is crucial to recognize that these components are not necessarily identical. Each model customizes its loss function to suit its unique characteristics. For example, the 3DJCG model specifically emphasizes enhancing the detection loss. Further details regarding the adaptations of each model are provided below. In terms of training techniques, Maximum Likelihood Estimation (MLE) and Self-Critical Sequence Training (SCST) [120] are the primary categories. Models such as Scan2Cap, MORE, SpaCap3D, and 3DJCG follow the MLE training scheme, while other methods like X-Trans2Cap, D3Net, CM3D, and Vote2Cap-DETR integrate both MLE and SCST schemes. When comparing these two techniques, SCST consistently outperforms pure MLE. SCST leverages the output of test-time inference to normalize the rewards it receives, eliminating the need for reward signal estimation and normalization. These findings have been further supported by ablation experiments conducted with select methods [30], [28].

1) *Scan2Cap*: The loss function employed in Scan2Cap comprises three distinct components, namely, the detection loss denoted as  $\mathcal{L}_{\text{det}}$ , the relative orientation loss denoted

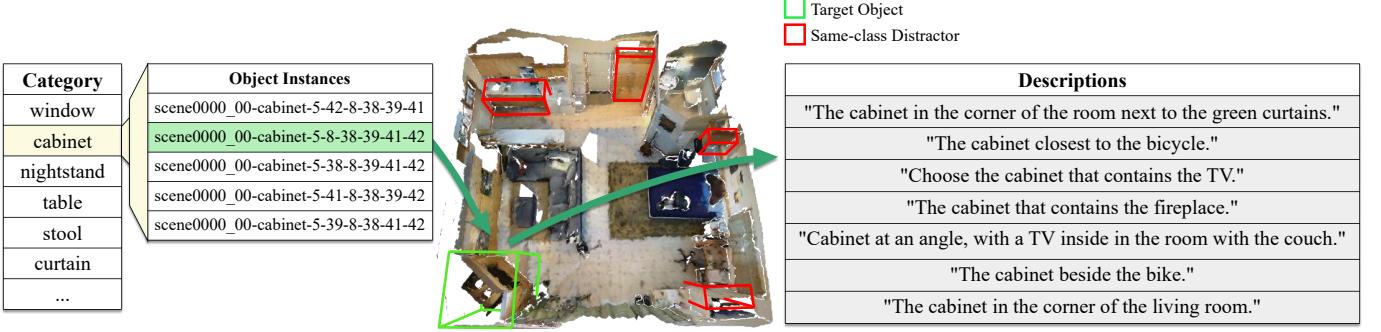


Fig. 10: Illustration of a typical example of scene 0000\_00 from Nr3D dataset [34]. The Nr3D dataset performs selective annotation of commonly occurring object categories, mainly encompassing multiple objects, with each object being associated with seven distinct descriptions.

as  $\mathcal{L}_{\text{rela}}$ , and the caption loss denoted as  $\mathcal{L}_{\text{cap}}$ . Specifically, the detection loss  $\mathcal{L}_{\text{det}}$  is formulated based on the approach proposed by Qi *et al.* [21], which encompasses four individual loss terms: vote regression loss, objectness binary classification loss, box regression loss, and semantic classification loss, as detailed in [21]. The relative orientation loss  $\mathcal{L}_{\text{rela}}$  is computed using a cross-entropy loss function, following the methodology employed in previous works [6], [5]. The caption loss  $\mathcal{L}_{\text{cap}}$  is determined using a conventional cross-entropy loss.

$$\begin{aligned} \mathcal{L} &= \alpha \mathcal{L}_{\text{det}} + \beta \mathcal{L}_{\text{rela}} + \gamma \mathcal{L}_{\text{cap}} \\ \mathcal{L}_{\text{det}} &= \theta_1 \mathcal{L}_{\text{vote}} + \theta_2 \mathcal{L}_{\text{obj}} + \theta_3 \mathcal{L}_{\text{box}} + \theta_4 \mathcal{L}_{\text{sem}} \end{aligned} \quad (1)$$

where the hyperparameters are set as  $\alpha = 10$ ,  $\beta = 1$ ,  $\gamma = 0.1$ ,  $\theta_1 = 1$ ,  $\theta_2 = 0.5$ ,  $\theta_3 = 1$ ,  $\theta_4 = 0.1$ , respectively.

2) *MORE*: The loss function utilized in MORE is similar to that of Scan2Cap, with the exception of the relative orientation loss  $\mathcal{L}_{\text{rela}}$  component.

$$\mathcal{L} = \alpha \mathcal{L}_{\text{det}} + \beta \mathcal{L}_{\text{cap}} \quad (2)$$

where the hyperparameters  $\alpha$  and  $\beta$  are set to specific values, i.e.,  $\alpha = 10$  and  $\beta = 0.1$ , respectively.

3) *X-Trans2Cap*: The loss function of X-Trans2Cap comprises two main components: the feature alignment loss, denoted as  $\mathcal{L}_{\text{align}}$ , and the captioning loss, denoted as  $\mathcal{L}_{\text{cap}}$ . The feature alignment loss  $\mathcal{L}_{\text{align}}$  is based on a Smooth-L1 regression loss, while the captioning loss  $\mathcal{L}_{\text{cap}}$  is a weighted combination of a cross-entropy loss function denoted as  $\mathcal{L}_{\text{cro}}$  and a reward based on the CIDEr-D score [8] denoted as  $\mathcal{L}_{\text{CIDEr}}$ .

$$\begin{aligned} \mathcal{L} &= \alpha \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{cap}} \\ \mathcal{L}_{\text{cap}} &= \beta \mathcal{L}_{\text{cro}} + \gamma \mathcal{L}_{\text{CIDEr}} \end{aligned} \quad (3)$$

where the hyperparameters are set as  $\alpha = 1$ ,  $\beta = 1$  and  $\gamma = 0.1$ , respectively.

4) *SpaCap3D*: The loss function of SpaCap3D is mathematically represented in Eq. 4, where the relational loss denoted as  $\mathcal{L}_{\text{rela}}$  is constructed on top of a three-class cross-entropy loss, which serves as a guide for spatial relation

learning. The components of  $\mathcal{L}_{\text{det}}$  and  $\mathcal{L}_{\text{cap}}$  in SpaCap3D closely align with the approach employed in Scan2Cap.

$$\mathcal{L} = \alpha \mathcal{L}_{\text{det}} + \beta \mathcal{L}_{\text{rela}} + \gamma \mathcal{L}_{\text{cap}} \quad (4)$$

where the hyperparameters  $\alpha$ ,  $\beta$ , and  $\gamma$  are set to specific values, i.e.,  $\alpha = 10$ ,  $\beta = 1$  and  $\gamma = 0.1$ , respectively.

5) *3DJCG*: The loss function of 3DJCG comprises both captioning and grounding components, as it is a unified task. The formulation of the loss function is outlined in Eq. 5. Notably, the grounding loss and captioning loss employ similar loss functions, with the former utilizing an average cross-entropy loss while the latter utilizes a conventional cross-entropy loss. It is worth mentioning that the detection loss has been enhanced following approach [21], and the bounding box loss has been replaced by the boundary regression loss [121] denoted as  $\mathcal{L}_{\text{bbox-reg}}$ .

$$\begin{aligned} \mathcal{L} &= \alpha \mathcal{L}_{\text{det}} + \beta \mathcal{L}_{\text{gro}} + \gamma \mathcal{L}_{\text{cap}} \\ \mathcal{L}_{\text{det}} &= \theta_1 \mathcal{L}_{\text{vote}} + \theta_2 \mathcal{L}_{\text{obj}} + \theta_3 \mathcal{L}_{\text{bbox-reg}} + \theta_4 \mathcal{L}_{\text{sem}} \end{aligned} \quad (5)$$

where the hyperparameters are set as  $\alpha = 1$ ,  $\beta = 0.3$ ,  $\gamma = 0.2$ ,  $\theta_1 = 10$ ,  $\theta_2 = 1$ ,  $\theta_3 = 200$  and  $\theta_4 = 1$ , respectively.

6) *D3Net*: The comprehensive loss function employed in D3Net is presented in Eq. 6. The detection loss  $\mathcal{L}_{\text{det}}$  is introduced from PointGroup [97]. It encompasses multiple components, namely, cross-entropy loss, L1 regression loss, the means of minus cosine similarities, and binary cross-entropy loss, each contributing to the overall loss. Similarly,  $\mathcal{L}_{\text{rela}}$  is fully adopted from Scan2Cap as the relative orientation loss for guiding spatial relation learning. The joint loss of grounding and captioning, denoted as  $\mathcal{L}_{\text{gro-cap}}$ , represents the most intricate and complex aspect of D3Net's loss function, which is trained using reinforcement learning techniques [120], [122]. It comprises three key components: a maximize reward function  $R(\hat{C}, I)$ , a baseline reward function  $R(C^*, I)$ , and a caption loss  $\mathcal{L}_{\text{cap}}$ . Specifically, the reward function  $R(\hat{C}, I)$  is a weighted sum of three terms: the CIDEr score of the sampled captioning  $R'(\hat{C}, I)$ , the localization loss  $\mathcal{L}_{\text{loc}}$ , and the language object classification loss  $\mathcal{L}_{\text{locl}}$ , with the last two being closely related to the grounding task and both

employing the cross-entropy loss function. Meanwhile, the loss  $\mathcal{L}_{\text{cap}}$  is a standard captioning loss derived from the MLE strategy.

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{gro-cap}} + \alpha \mathcal{L}_{\text{rela}} \\ \mathcal{L}_{\text{det}} &= \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{obj-reg}} + \mathcal{L}_{\text{obj-dir}} + \mathcal{L}_{\text{c-score}} \\ \mathcal{L}_{\text{gro-cap}}(\theta) &\approx \left( R(\hat{C}, I) - R(C^*, I) \right) \mathcal{L}_{\text{cap}} \\ \mathcal{L}_{\text{cap}}(\theta) &= - \sum_{t=1}^T \log p(\hat{c}_t | \hat{c}_1, \dots, \hat{c}_{t-1}; I, \theta) \\ R(\hat{C}, I) &= R'(\hat{C}, I) - \beta [\mathcal{L}_{\text{loc}}(\hat{C}) + \gamma \mathcal{L}_{\text{locl}}(\hat{C})]\end{aligned}\quad (6)$$

where the hyperparameters  $\alpha$ ,  $\beta$ , and  $\gamma$  are set to specific values, i.e.,  $\alpha=0.3$ ,  $\beta=0.1$ , and  $\gamma=1$ , respectively. The generated token at time step  $t$  is denoted as  $\hat{c}_t$ , and the visual signal is represented by  $I$ . The model parameters are denoted as  $\theta$ , and the generated description is denoted as  $\hat{C}$ , which can be expressed as  $\hat{C} = \{\hat{c}_1, \dots, \hat{c}_T\}$ .

7) *CM3D*: CM3D combines MLE training objectives with a SCST, which is formulated as follows.

$$\begin{aligned}\mathcal{L}_{\text{mle}} &= - \sum_{i=1}^T \log \hat{P}(c_i^{t+1} | \mathcal{V}, c_i^{[1:t]}) \\ \mathcal{L}_{\text{scst}} &= - \sum_{i=1}^k (R(\hat{c}_i) - R(\hat{g})) \cdot \frac{1}{|\hat{c}_i|} \log \hat{P}(\hat{c}_i | \mathcal{V})\end{aligned}\quad (7)$$

where  $c_i^{t+1}$  represents the  $(t+1)^{\text{th}}$  word, and the first  $t$  words are denoted as  $c_i^{[1:t]}$ . The visual clue is represented as  $\mathcal{V}$ . The function  $R(\cdot)$  denotes the CIDEr reward function, while  $\hat{c}_i$  represents the generated multiple captions  $c_i^1, \dots, c_i^k$  with a beam size of  $k$ , and  $\hat{g}$  serves as a baseline.

8) *Vote2Cap-DETR*: The loss function of Vote2Cap-DETR comprises three components: vote query loss  $\mathcal{L}_{\text{vq}}$ , detection loss  $\mathcal{L}_{\text{det}}$ , and caption loss  $\mathcal{L}_{\text{cap}}$ . The vote query loss  $\mathcal{L}_{\text{vq}}$  is adopted from VoteNet, while the caption loss is fine-tuned with the self-critical sequence training strategy after maximum likelihood estimation. As for the detection loss  $\mathcal{L}_{\text{det}}$ , it updates the weights on top of the original 3DETR loss to refine the model's performance in object detection. Specifically,  $\mathcal{L}_{\text{det}}$  consists of four parts  $\mathcal{L}_{\text{giou}}$ ,  $\mathcal{L}_{\text{cls}}$ ,  $\mathcal{L}_{\text{center-reg}}$ , and  $\mathcal{L}_{\text{size-reg}}$ , indicating the *gIoU* loss, the box size classification loss, the box center regressing loss, and the box size regressing loss, respectively.

$$\begin{aligned}\mathcal{L} &= \alpha \mathcal{L}_{\text{vq}} + \beta \sum_{i=1}^{n_{\text{dec-layer}}} \mathcal{L}_{\text{det}} + \gamma \mathcal{L}_{\text{cap}} \\ \mathcal{L}_{\text{det}} &= \theta_1 \mathcal{L}_{\text{giou}} + \theta_2 \mathcal{L}_{\text{cls}} + \theta_3 \mathcal{L}_{\text{center-reg}} + \theta_4 \mathcal{L}_{\text{size-reg}}\end{aligned}\quad (8)$$

where the hyperparameters are set as  $\alpha = 10$ ,  $\beta = 1$ ,  $\gamma = 5$ ,  $\theta_1 = 10$ ,  $\theta_2 = 1$ ,  $\theta_3 = 5$ ,  $\theta_4 = 1$ , respectively.

## B. Performance Analysis

The performance of the advanced methods on the ScanRefer and Nr3D datasets are summarized in Table IV and Table V, respectively. Furthermore, Table VI presents the performance of the Scan2Cap online benchmark, the currently only

benchmark that incorporates the test dataset of ScanRefer for 3D dense captioning. It is noteworthy to mention that we utilized the best outcomes reported in the original papers as our evaluation criterion, considering the presence of diverse data types and various training strategies presented for each approach.

Table IV summarizes the performance of various advanced methods on the ScanRefer dataset. The transformer-based framework Vote2Cap-DETR, equipped with innovative approaches, achieved state-of-the-art performance in terms of  $C$  and  $B-4$  index, regardless of whether the IoU threshold is set at 0.5 or 0.25. However, D3Net, another strong performer, outperformed Vote2Cap-DETR in terms of  $M$  and  $R$  index at IoU of 0.5 and achieved the best mAP result. It should be indicated that dual learning between 3D dense captioning and 3D visual grounding on larger datasets can significantly improve model performance. In contrast, 3DJCG, a transformer-based model for dual learning tasks, performed worse than D3Net's non-transformer architecture. This may be due to the fact that 3DJCG used VoteNet as a scene encoder instead of replacing it with a more powerful detector, as done in D3Net, highlighting the importance of a strong scene encoder. Transformer-based models such as CM3D, SpaCap3D, and X-Trans2Cap generally outperformed graph and GRU-based methods such as Scan2Cap and MORE, reflecting the dominance and potential of transformer-based approaches. Furthermore, incorporating a more comprehensive relationship module usually leads to performance improvement. For example, MORE, built on Scan2Cap, enhanced the relationship module and achieved a 3.78%  $C@0.5\text{IoU}$  improvement over Scan2Cap in 3D input. Additionally, incorporating additional 2D input, such as pre-trained multi-view features, can further boost performance, which is evident in almost every method.

Table V presents the performance of various methods on the Nr3D dataset, which is more challenging and focuses on fine-grained 3D object detection. The ranking trend of the methods' performance on Nr3D is similar to that on ScanRefer, with Vote2Cap-DETR remaining at the top of the list. However, the highest score achieved on Nr3D was only 45.53%  $C@0.5\text{IoU}$ , which is quantitatively 25.1% lower than the performance on ScanRefer. This indicates that existing models may struggle to provide distinctive descriptions when multiple instances of the same type of objects appear in a scene, which makes the task more challenging. Notably, 3DJCG showed superior performance compared to CM3D on the Nr3D dataset, suggesting that unified models that combine 3D dense captioning and 3D visual grounding could be more effective in dealing with complex scenes where fine-grained object detection is required. This highlights the importance of integrating different modalities and tasks to achieve better performance in challenging scenarios.

## VII. DISCUSSION

Although prior works had made significant progress on 3D dense captioning, there remain several challenges that could be the focus of future studies. Hence, we will discuss the challenges and future trends of 3D dense captioning from six

TABLE IV: Experimental results of different models on the ScanRefer Validation Set. The best-reported result from the original paper is used as the benchmark standard. “-” signifies that neither the original paper nor any subsequent works provide such results. The term “3D” refers to methods that solely utilize coordinate information, while “2D+3D” incorporates additional auxiliary features such as colors, normals, multi-view features, etc., along with point coordinates. “GA” denotes the use of a GRU (Gated Recurrent Unit) backbone with an attention mechanism, and “KD” stands for Knowledge Distillation. Caption metrics, including CIDEr, BLEU-4, METEOR, and ROUGE (denoted as C, B-4, M, and R, respectively), are adopted to evaluate the quality of captions, and mAP (mean Average Precision) is employed as the detection metric.

Method	Architecture	Data	IoU=0.5					IoU=0.25			
			C	B-4	M	R	mAP	C	B-4	M	R
Scan2Cap	VoteNet Graph GA	3D	35.20	22.36	21.44	43.57	29.13	53.73	34.25	26.14	54.95
		2D+3D	39.08	23.32	21.97	44.78	32.21	56.82	34.18	26.29	55.27
MORE	VoteNet Graph GA	3D	38.98	23.01	21.65	44.33	31.93	58.89	35.41	26.36	55.41
		2D+3D	40.94	22.93	21.66	44.42	33.75	62.91	36.25	26.75	56.33
X-Trans2Cap	VoteNet KD	3D	41.52	23.83	21.09	44.97	34.68	58.81	34.17	25.81	54.10
		2D+3D	43.87	25.05	22.46	45.28	35.31	61.83	35.65	26.61	54.70
SpaCap3D	VoteNet Transformer	3D	42.53	25.02	22.22	45.65	34.44	58.06	35.30	26.16	55.03
		2D+3D	44.02	25.26	22.33	45.36	36.64	63.30	36.46	26.71	55.71
3DJCG	VoteNet Transformer	3D	47.68	31.53	24.28	51.08	-	60.86	39.67	27.45	59.02
		2D+3D	49.48	31.03	24.22	50.80	-	64.70	40.17	27.66	<b>59.23</b>
CM3D	VoteNet Transformer	3D	50.29	25.64	22.57	44.71	35.97	-	-	-	-
		2D+3D	54.30	27.24	23.30	45.81	42.77	-	-	-	-
D3Net	PointGroup Graph GA	3D	-	-	-	-	-	-	-	-	-
		2D+3D	62.64	35.68	<b>25.72</b>	<b>53.90</b>	<b>53.95</b>	-	-	-	-
Vote2Cap-DETR	3DERT Transformer	3D	<b>73.77</b>	<b>38.21</b>	<b>26.64</b>	<b>54.71</b>	-	<b>84.15</b>	<b>42.51</b>	<b>28.47</b>	<b>59.26</b>
		2D+3D	<b>70.63</b>	<b>35.69</b>	25.51	52.28	-	<b>86.28</b>	<b>42.64</b>	<b>28.27</b>	59.07

TABLE V: Experimental results of different models on the Nr3D validation set. The best result reported in the original paper is taken as the comparison standard.

Method	Data	IoU=0.5			
		C	B-4	M	R
X-Trans2Cap	3D	30.96	18.70	22.15	49.92
SpaCap3D	3D	31.43	18.98	22.24	49.79
CM3D	3D	<b>35.86</b>	<b>20.73</b>	<b>22.86</b>	<b>51.23</b>
X-Trans2Cap	2D+3D	33.62	19.29	22.27	50.00
SpaCap3D	2D+3D	33.71	19.92	22.61	50.50
CM3D	2D+3D	37.37	20.96	22.89	51.11
3DJCG	2D+3D	38.06	22.82	23.77	52.99
D3Net	2D+3D	38.42	22.22	24.74	54.37
Vote2Cap-DETR	2D+3D	<b>45.53</b>	<b>26.88</b>	<b>25.43</b>	<b>54.76</b>

different aspects in this session, including datasets, external 2D knowledge, framework, generator module, vision-language pretraining technique, unified networks, and downstream applications.

**From Datasets:** The limited availability of diverse and large-scale datasets for 3D dense captioning is a significant challenge. Existing datasets are based on indoor scenes and require human-annotated labels, which increases the cost of collecting data and introduces linguistic priors. Specifically, there are only two datasets [31], [34] for 3D dense captioning, while the number of image captioning datasets [123], [124], [125], [126], [127], [128], [129], [130], [131] exceeds ten quantity. Moreover, both datasets are based on indoor scenes, which

inevitably carry more fixed relationships than random and complex outdoor scenes. In addition, most current models use supervised learning methods that require human-annotated labels for training, which greatly increases the cost of collecting the dataset. Moreover, it not only demands the accuracy of the dataset but also brings the problem of over-reliance on linguistic priors [28]. Therefore, future research can focus on developing unsupervised and reinforcement learning methods that rely less on labeled data and also explore ways to increase the diversity and size of datasets.

**From External 2D Knowledge:** The use of additional 2D features, such as those extracted by pre-trained ENet [132], is crucial for generating high-quality captions in 3D dense captioning. However, this can result in a computational burden [26]. Developing robust migration models that effectively integrate 2D and 3D features without compromising efficiency [133] is a challenge that needs to be addressed.

**From Framework:** Most of the previous work retains the detector, which causes the quality of the generated captions to be severely limited by the detection results. Although [30] creates the precedent of the detector-free full end-to-end framework and obtains state-of-the-art performance, constructing a stronger detector-free model still needs exploration. Additionally, we can clearly observe from Table IV that the techniques we used are slightly homogeneous, which is either only graph-based or only transformer-based. Inspired by image captioning [16], we can attempt to incorporate a diverse range of technologies in the future, such as combinations of graph-based and transformer-based techniques, and build more robust

TABLE VI: The performance of the Scan2Cap online benchmark, the currently only benchmark that incorporates the test dataset of ScanRefer for 3D dense captioning. The superscripts in each column indicate the models’ rankings based on different metrics. The first sorting criterion used is the CIDEr metric (C@0.5IoU). The overall ranking results are largely consistent with those in Table IV and Table V. However, a slight difference is observed where the individual D3Net-Speaker model is not as effective as the combined D3Net model.

Method	Captioning					Detection mAP@0.5	Submission date
	C@0.5IoU	B@0.5IoU	R@0.5IoU	M@0.5IoU	DCmAP		
Vote2Cap-DETR[30]	<b>0.3128</b> <sup>1</sup>	<b>0.1778</b> <sup>1</sup>	<b>0.2842</b> <sup>1</sup>	<b>0.1316</b> <sup>1</sup>	<b>0.1825</b> <sup>1</sup>	<b>0.4454</b> <sup>1</sup>	17 Nov, 2022
CFM	0.2360 <sup>2</sup>	0.1417 <sup>2</sup>	0.2253 <sup>2</sup>	0.1034 <sup>2</sup>	0.1379 <sup>5</sup>	0.3008 <sup>5</sup>	4 Nov, 2022
CM3D[29]	0.2348 <sup>3</sup>	0.1383 <sup>3</sup>	0.2250 <sup>4</sup>	0.1030 <sup>3</sup>	0.1398 <sup>4</sup>	0.2966 <sup>7</sup>	27 Sep, 2022
Forest-xyz	0.2266 <sup>4</sup>	0.1363 <sup>4</sup>	0.2250 <sup>3</sup>	0.1027 <sup>4</sup>	0.1161 <sup>10</sup>	0.2825 <sup>10</sup>	6 Oct, 2022
D3Net-Speaker[28]	0.2088 <sup>5</sup>	0.1335 <sup>6</sup>	0.2237 <sup>5</sup>	0.1022 <sup>5</sup>	0.1481 <sup>3</sup>	0.4198 <sup>2</sup>	25 Aug, 2022
3DJCG(Captioning)[27]	0.1918 <sup>6</sup>	0.1350 <sup>5</sup>	0.2207 <sup>6</sup>	0.1013 <sup>6</sup>	0.1506 <sup>2</sup>	0.3867 <sup>3</sup>	12 Sep, 2022
REMAN	0.1662 <sup>7</sup>	0.1070 <sup>7</sup>	0.1790 <sup>7</sup>	0.0815 <sup>7</sup>	0.1235 <sup>8</sup>	0.2927 <sup>9</sup>	11 Sep, 2022
NOAH	0.1382 <sup>8</sup>	0.0901 <sup>8</sup>	0.1598 <sup>8</sup>	0.0747 <sup>8</sup>	0.1359 <sup>6</sup>	0.2977 <sup>6</sup>	28 Sep, 2022
SpaCap3D[25]	0.1359 <sup>9</sup>	0.0883 <sup>9</sup>	0.1591 <sup>9</sup>	0.0738 <sup>9</sup>	0.1182 <sup>9</sup>	0.3275 <sup>4</sup>	31 Aug, 2022
X-Trans2Cap[26]	0.1274 <sup>10</sup>	0.0808 <sup>11</sup>	0.1392 <sup>11</sup>	0.0653 <sup>11</sup>	0.1244 <sup>7</sup>	0.2795 <sup>11</sup>	29 Aug, 2022
MORE-xyz[24]	0.1239 <sup>11</sup>	0.0796 <sup>12</sup>	0.1362 <sup>12</sup>	0.0631 <sup>12</sup>	0.1116 <sup>12</sup>	0.2648 <sup>12</sup>	11 Sep, 2022
SUN+	0.1148 <sup>12</sup>	0.0846 <sup>10</sup>	0.1564 <sup>10</sup>	0.0711 <sup>10</sup>	0.1143 <sup>11</sup>	0.2958 <sup>8</sup>	28 Sep, 2022
Scan2Cap[31]	0.0849 <sup>13</sup>	0.0576 <sup>13</sup>	0.1073 <sup>13</sup>	0.0492 <sup>13</sup>	0.0970 <sup>13</sup>	0.2481 <sup>13</sup>	25 Aug, 2022

and effective models.

**From Generator Module:** Most current generator decoders in 3D dense captioning generate sequential sentences word-by-word, which the previous word can influence. Exploring parallel word generation techniques, such as diffusion models, can enable bidirectional textual message interaction and potentially improve the generation process. It is worth mentioning that the diffusion model [134] has made a significant breakthrough in visual content generation by generating the words in parallel and enabling bidirectional textual message interaction. Most recently, Chen *et al.* confirmed the validity of the diffusion model on image captioning [135], [136], which may be able to be transferred to future 3D dense captioning tasks.

**From Vision-Language Pretraining Technique:** It is an undeniable fact that large-scale vision-language pre-training (VLP) models hold unparalleled advantages over other models, and their implementation in various fields has been steadily increasing. The recent release of GPT-4 [137] has particularly stirred the research community, creating a seismic impact. Notably, image captioning and dense image captioning [138], [139], [140], [141] have witnessed a proliferation of relevant studies with remarkable results, leveraging models such as CLIP [50], BERT [10], and GPT-2 [51]. In light of these developments, it is imperative to bridge the gap in VLP for 3D dense captioning, which we believe will usher this research field into a new era.

**From Unified Networks and Downstream Applications:** Currently, an increasing number of researchers have expanded their scope beyond single-task exploration and placing greater emphasis on the integration of multiple tasks [66], [67], [66], [67]. Notably, the combination of 3D dense captioning with 3D visual grounding has been shown to possess significant potential [28], [27]. Consequently, the exploration of joint models for related tasks is emerging as a promising research

direction, presenting both opportunities and challenges for the future and potentially giving rise to new 3D multimodal tasks. Furthermore, there is a notable gap in the existing literature in terms of the limited attention given to downstream applications. The application of 3D dense captioning to assist individuals with visual impairments, for instance, holds greater relevance to real-world scenarios compared to 2D-based tasks. Furthermore, there are still unexplored opportunities and untapped potential in this area, awaiting further exploration and investigation.

## VIII. CONCLUSIONS

This paper presents a comprehensive review of the field of 3D dense captioning, which includes analyzing the main framework, datasets, related tasks, evaluation metrics, and future directions. By critically examining the strengths and weaknesses of existing models and considering the development trends in other domains, this paper has identified the current challenges and limitations of 3D dense captioning. The motivation of this paper is not only to provide a thorough understanding of the task for scholars who may be unfamiliar with it but also to serve as a source of inspiration and guidance for future research. It is our hope that this review will stimulate further investigations and advancements in the field of 3D dense captioning, ultimately leading to breakthroughs and advancements in this emerging research area.

## REFERENCES

- [1] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona, “What do we perceive in a glance of a real-world scene?” *Journal of vision*, vol. 7, no. 1, pp. 10–10, 2007.
- [2] T. Xian, Z. Li, Z. Tang, and H. Ma, “Adaptive path selection for dynamic image captioning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5762–5775, 2022.

- [3] S. Cao, G. An, Z. Zheng, and Z. Wang, "Vision-enhanced and consensus-aware transformer for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 7005–7018, 2022.
- [4] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467–4480, 2020.
- [5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
- [7] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.
- [8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [11] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.
- [12] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4565–4574.
- [13] L. Yang, K. Tang, J. Yang, and L.-J. Li, "Dense captioning with joint inference and visual context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2193–2202.
- [14] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, and J. Shao, "Context and attribute grounded dense captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6241–6250.
- [15] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6271–6280.
- [16] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *arXiv preprint arXiv:2201.12944*, 2022.
- [17] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1534–1543.
- [18] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, M.-K. Tran, L.-F. Yu, and S.-K. Yeung, "Scenenn: A scene meshes dataset with annotations," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 92–101.
- [19] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.
- [20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [21] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.
- [22] J. Hou, A. Dai, and M. Nießner, "Revealnet: Seeing behind objects in rgb-d scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2098–2107.
- [23] Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, "Scan2cap: Context-aware dense captioning in rgb-d scans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3193–3203.
- [24] Y. Jiao, S. Chen, Z. Jie, J. Chen, L. Ma, and Y.-G. Jiang, "More: Multi-order relation mining for dense captioning in 3d scenes," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*. Springer, 2022, pp. 528–545.
- [25] H. Wang, C. Zhang, J. Yu, and W. Cai, "Spatiality-guided transformer for 3d dense captioning on point clouds," *arXiv preprint arXiv:2204.10688*, 2022.
- [26] Z. Yuan, X. Yan, Y. Liao, Y. Guo, G. Li, S. Cui, and Z. Li, "X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8563–8573.
- [27] D. Cai, L. Zhao, J. Zhang, L. Sheng, and D. Xu, "3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16464–16473.
- [28] N. Takahashi and Y. Mitsufuji, "D3net: Densely connected multilidited densenet for music source separation," *arXiv preprint arXiv:2010.01733*, 2020.
- [29] Y. Zhong, L. Xu, J. Luo, and L. Ma, "Contextual modeling for 3d dense captioning on point clouds," *arXiv preprint arXiv:2210.03925*, 2022.
- [30] S. Chen, H. Zhu, X. Chen, Y. Lei, T. Chen, and G. YU, "End-to-end 3d dense captioning with vote2cap-detr," *arXiv preprint arXiv:2301.02508*, 2023.
- [31] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*. Springer, 2020, pp. 202–221.
- [32] Z. Yuan, X. Yan, Y. Liao, R. Zhang, S. Wang, Z. Li, and S. Cui, "Instancerefere: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1791–1800.
- [33] L. Zhao, D. Cai, L. Sheng, and D. Xu, "3dvg-transformer: Relation modeling for visual grounding on point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2928–2937.
- [34] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 422–440.
- [35] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 43–51, 2022.
- [36] A.-A. Liu, Y. Zhai, N. Xu, W. Nie, W. Li, and Y. Zhang, "Region-aware image captioning via interaction learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3685–3696, 2022.
- [37] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.
- [38] H. Senior, G. Slabaugh, S. Yuan, and L. Rossi, "Graph neural networks in vision-language image understanding: A survey," *arXiv preprint arXiv:2303.03761*, 2023.
- [39] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11*. Springer, 2010, pp. 15–29.
- [40] S. Wu, J. Wieland, O. Farivar, and J. Schiller, "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service," in *proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 2017, pp. 1180–1192.
- [41] T. Yu, J. Yu, Z. Yu, and D. Tao, "Compositional attention networks with two-stream fusion for video question answering," *IEEE Transactions on Image Processing*, vol. 29, pp. 1204–1218, 2019.
- [42] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1969–1978.
- [43] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.

- [44] X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-encoding scene graphs for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 685–10 694.
- [45] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [46] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [47] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.
- [48] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image captioning: Transforming objects into words,” *Advances in neural information processing systems*, vol. 32, 2019.
- [49] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Mesched-memory transformer for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.
- [50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [51] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [55] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [56] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [57] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, “Attention on attention for image captioning,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.
- [58] L. Huang, W. Wang, Y. Xia, and J. Chen, “Adaptively aligned image captioning via adaptive attention time,” *Advances in neural information processing systems*, vol. 32, 2019.
- [59] L. Wang, Z. Bai, Y. Zhang, and H. Lu, “Show, recall, and tell: Image captioning with recall mechanism,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 176–12 183.
- [60] T. Yu, J. Yu, Z. Yu, Q. Huang, and Q. Tian, “Long-term video question answering via multimodal hierarchical memory attentive networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 931–944, 2020.
- [61] Y. Pan, T. Yao, Y. Li, and T. Mei, “X-linear attention networks for image captioning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 971–10 980.
- [62] I. Schwartz, A. Schwing, and T. Hazan, “High-order attention models for visual question answering,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [63] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.
- [64] Z. Zeng, H. Zhang, Z. Wang, R. Lu, D. Wang, and B. Chen, “Conzic: Controllable zero-shot image captioning by sampling-based polishing,” *arXiv preprint arXiv:2303.02437*, 2023.
- [65] X. Song, B. Wang, G. Chen, and S. Jiang, “Much: Mutual coupling enhancement of scene recognition and dense captioning,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 793–801.
- [66] Y. Long, Y. Wen, J. Han, H. Xu, P. Ren, W. Zhang, S. Zhao, and X. Liang, “Capdet: Unifying dense captioning and open-world detection pretraining,” *arXiv preprint arXiv:2303.02489*, 2023.
- [67] Y. Gao, X. Hou, Y. Zhang, T. Ge, Y. Jiang, and P. Wang, “Capon-image: Context-driven dense-captioning on image,” *arXiv preprint arXiv:2204.12974*, 2022.
- [68] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, “Dense-captioning events in videos,” in *International Conference on Computer Vision (ICCV)*, 2017.
- [69] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, “Swinbert: End-to-end transformers with sparse attention for video captioning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.
- [70] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li, J. Li, T. Bharti, and M. Zhou, “Univl: A unified video and language pre-training model for multimodal understanding and generation,” *arXiv preprint arXiv:2002.06353*, 2020.
- [71] 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*, 2022, pp. 17 959–17 968.
- [72] 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 Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7492–7500.
- [73] 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*, 2019, pp. 6588–6597.
- [74] V. Iashin and E. Rahtu, “Multi-modal dense video captioning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 958–959.
- [75] V. Lashin and E. Rahtu, “A better use of audio-visual cues: Dense video captioning with bi-modal transformer,” *arXiv preprint arXiv:2005.08271*, 2020.
- [76] 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 conference on computer vision and pattern recognition*, 2018, pp. 7190–7198.
- [77] 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.
- [78] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, “Turn tap: Temporal unit regression network for temporal action proposals,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3628–3636.
- [79] F. C. Heilbron, J. C. Niebles, and B. Ghanem, “Fast temporal activity proposals for efficient detection of human actions in untrimmed videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1914–1923.
- [80] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, “Video captioning with attention-based lstm and semantic consistency,” *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.
- [81] 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.
- [82] 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 conference on computer vision and pattern recognition*, 2018, pp. 8739–8748.
- [83] 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*, 2021, pp. 234–243.
- [84] W. Zhu, B. Pang, A. Thapliyal, W. Y. Wang, and R. Soricut, “End-to-end dense video captioning as sequence generation,” *arXiv preprint arXiv:2204.08121*, 2022.
- [85] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, “Large-scale adversarial training for vision-and-language representation learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6616–6628, 2020.
- [86] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.
- [87] G. Li, N. Duan, Y. Fang, M. U.-V. Gong, and D. U.-V. Jiang, “A universal encoder for vision and language by cross-modal pre-training. arxiv 2019,” *arXiv preprint arXiv:1908.06066*.
- [88] 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*, 2023, pp. 10 714–10 726.

- [89] H. Sheng, S. Cai, N. Zhao, B. Deng, M.-J. Zhao, and G. H. Lee, “Pdr: Progressive depth regularization for monocular 3d object detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [90] C. Wang, J. Deng, J. He, T. Zhang, Z. Zhang, and Y. Zhang, “Long-short range adaptive transformer with dynamic sampling for 3d object detection,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [91] D. He, Y. Zhao, J. Luo, T. Hui, S. Huang, A. Zhang, and S. Liu, “Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2344–2352.
- [92] P.-H. Huang, H.-H. Lee, H.-T. Chen, and T.-L. Liu, “Text-guided graph neural networks for referring 3d instance segmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1610–1618.
- [93] Z. Yang, S. Zhang, L. Wang, and J. Luo, “Sat: 2d semantics assisted training for 3d visual grounding,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1856–1866.
- [94] J. Chen, W. Luo, X. Wei, L. Ma, and W. Zhang, “Ham: Hierarchical attention model with high performance for 3d visual grounding,” *arXiv preprint arXiv:2210.12513*, 2022.
- [95] M. Honnibal and I. Montani, “spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing,” *To appear*, vol. 7, no. 1, pp. 411–420, 2017.
- [96] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [97] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, “Pointgroup: Dual-set point grouping for 3d instance segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, 2020, pp. 4867–4876.
- [98] I. Misra, R. Girdhar, and A. Joulin, “An end-to-end transformer model for 3d object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2906–2917.
- [99] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [100] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [101] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph cnn for learning on point clouds,” *Acm Transactions On Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.
- [102] M. Feng, Z. Li, Q. Li, L. Zhang, X. Zhang, G. Zhu, H. Zhang, Y. Wang, and A. Mian, “Free-form description guided 3d visual graph network for object grounding in point cloud,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3722–3731.
- [103] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [104] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, “Scene coordinate regression forests for camera relocalization in rgb-d images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2930–2937.
- [105] J. Xiao, A. Owens, and A. Torralba, “Sun3d: A database of big spaces reconstructed using sfm and object labels,” in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1625–1632.
- [106] S. Song and J. Xiao, “Deep sliding shapes for amodal 3d object detection in rgb-d images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 808–816.
- [107] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, “3d shapenets: A deep representation for volumetric shapes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.
- [108] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, “Shapenet: An information-rich 3d model repository,” *arXiv preprint arXiv:1512.03012*, 2015.
- [109] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images.” *ECCV (5)*, vol. 7576, pp. 746–760, 2012.
- [110] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of rgbd slam systems,” in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.
- [111] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, “Pigraphs: learning interaction snapshots from observations,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–12, 2016.
- [112] 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*, 2015, pp. 4566–4575.
- [113] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
- [114] 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*, 2005, pp. 65–72.
- [115] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.
- [116] P. Anderson, B. Fernando, M. Johnson, and S. Gould, “Spice: Semantic propositional image caption evaluation,” in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 382–398.
- [117] O. Callison-Burch, “Koehn, re-evaluation the role of bleu in machine translation research,” in *11th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2006.
- [118] C.-Y. Lin and F. J. Och, “Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics,” in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 2004, pp. 605–612.
- [119] S. Robertson, “Understanding inverse document frequency: on theoretical arguments for idf,” *Journal of documentation*, 2004.
- [120] S. J. Rennie, E. Marcheret, Y. Mrkuech, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.
- [121] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.
- [122] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, “Discriminability objective for training descriptive captions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6964–6974.
- [123] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [124] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [125] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *International journal of computer vision*, vol. 123, pp. 32–73, 2017.
- [126] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, “Stylenet: Generating attractive visual captions with styles,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3137–3146.
- [127] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, “Captioning images taken by people who are blind,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 417–434.
- [128] P. Sharma, N. Ding, S. Goodman, and R. Soricul, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.
- [129] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kanali, S. Popov, M. Mallochi, A. Kolesnikov *et al.*, “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale,” *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

- [130] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, “Nocaps: Novel object captioning at scale,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8948–8957.
- [131] A. Mathews, L. Xie, and X. He, “Senticap: Generating image descriptions with sentiments,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.
- [132] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” *arXiv preprint arXiv:1606.02147*, 2016.
- [133] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, “Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders,” *arXiv preprint arXiv:2212.06785*, 2022.
- [134] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [135] T. Chen, R. Zhang, and G. Hinton, “Analog bits: Generating discrete data using diffusion models with self-conditioning,” *arXiv preprint arXiv:2208.04202*, 2022.
- [136] J. Luo, Y. Li, Y. Pan, T. Yao, J. Feng, H. Chao, and T. Mei, “Semantic-conditional diffusion networks for image captioning,” *arXiv preprint arXiv:2212.03099*, 2022.
- [137] OpenAI, “Gpt-4 technical report,” 2023.
- [138] R. Mokady, A. Hertz, and A. H. Bermano, “Clipcap: Clip prefix for image captioning,” *arXiv preprint arXiv:2111.09734*, 2021.
- [139] X. Hu, X. Yin, K. Lin, L. Zhang, J. Gao, L. Wang, and Z. Liu, “Vivo: Visual vocabulary pre-training for novel object captioning,” in *proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 2, 2021, pp. 1575–1583.
- [140] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.
- [141] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, “Unified vision-language pre-training for image captioning and vqa,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.