End-to-End 3D Dense Captioning with Vote2Cap-DETR

Sijin Chen[1*] Hongyuan Zhu[2] Xin Chen[3] Yinjie Lei[4] Gang Yu[3] Tao Chen[1†]
[1]Fudan University [2]Institute for Infocomm Research (I^{2}R) & Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[3]Tencent PCG [4]Sichuan University

https://github.com/ch3cook-fdu/Vote2Cap-DETR

Abstract

*3D dense captioning aims to generate multiple captions localized with their associated object regions. Existing methods follow a sophisticated "detect-then-describe" pipeline equipped with numerous hand-crafted components. However, these hand-crafted components would yield sub-optimal performance given cluttered object spatial and class distributions among different scenes. In this paper, we propose a simple-yet-effective transformer framework Vote2Cap-DETR based on recent popular **DE**tection **TR**ansformer (DETR). Compared with prior arts, our framework has several appealing advantages: 1) Without resorting to numerous hand-crafted components, our method is based on a full transformer encoder-decoder architecture with a learnable vote query driven object decoder, and a caption decoder that produces the dense captions in a set-prediction manner. 2) In contrast to the two-stage scheme, our method can perform detection and captioning in one-stage. 3) Without bells and whistles, extensive experiments on two commonly used datasets, ScanRefer and Nr3D, demonstrate that our Vote2Cap-DETR surpasses current state-of-the-arts by 11.13% and 7.11% in CIDEr@0.5IoU, respectively. Codes will be released soon.*

1. Introduction

In recent years, works on 3D learning has grown dramatically for various applications [10, 11, 21, 41, 42]. Among them, 3D dense captioning [7, 13] requires a system to localize all the objects in a 3D scene and generate descriptive sentences for each object. This problem is challenging, given 1) the sparsity of point clouds and 2) the cluttered distribution of objects.

3D dense captioning can be divided into two tasks, object detection, and object caption generation. Scan2Cap [13], MORE [20], and SpaCap3D [39] propose well-designed relation reasoning modules to model relations among object proposals efficiently. [48] introduces contextual information from two branches to improve the caption. 3DJCG [4] and D3Net [7] study the correlation between 3D visual grounding and 3D dense captioning and point out that these two tasks promote each other. Additionally, χ-Trans2Cap [43] discusses how to transfer knowledge from additional 2d information to boost 3d dense captioning.

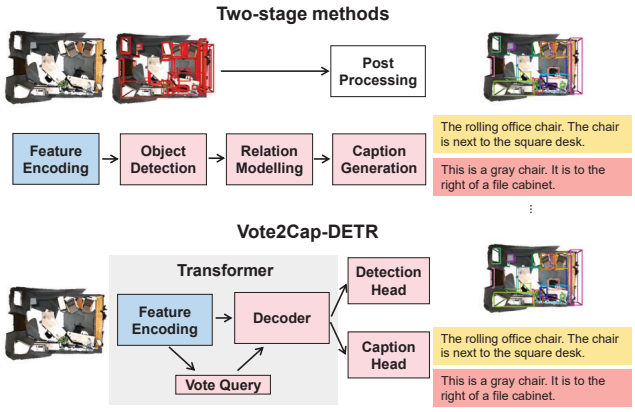

Figure 1. **Illustration of existing two-stage 3D dense captioning method (upper) and our Vote2Cap-DETR (bottom).** Existing methods adopt a two-stage pipeline that heavily depends on a detector's output. Therefore, we propose a transformer-based one-stage model, Vote2Cap-DETR, that frames 3D dense captioning as a set prediction problem.

Among existing methods, they all adopt a two-stage "detect-then-describe" pipeline [4, 7, 13, 20, 39, 48] (Figure 1). This pipeline first generates a set of object proposals, then decodes each object by a caption generator with an explicit reasoning procedure. Though these methods have achieved remarkable performance, the "detect-then-describe" pipeline suffers from the following issues: 1) Because of the serial and explicit reasoning, the captioning performance highly depends on the object detection performance, which limits the mutual promotion of detection and captioning. 2) The heavy reliance on hand-crafted components, e.g., radii, 3D operators, the definition of pro-

*Part of this work was accomplished under supervison by Dr. Hongyuan Zhu from A*STAR, Singapore.
†Corresponding author.

posal neighbors, and post-processing (non-maximum suppression [28]) introduces additional hyper-parameters, leading to a sub-optimal performance given the sparse object surfaces and cluttered object distributions among different indoor scenes. This inspires us to design a one-stage 3D dense captioning system.

To address the above issues, we propose Vote2Cap-DETR, a full transformer encoder-decoder architecture for one-stage 3D dense captioning. Unlike traditional "detect-then-describe" pipelines, we directly feed the decoder's output into the localization head and caption head in parallel. By casting 3D dense captioning as a set-to-set problem, each target instance and its language annotation is matched with a query in a one-to-one correspondence manner, enabling a more discriminative feature representation for proposals to identify each distinctive object in a 3D scene. Additionally, we also propose a novel vote query driven decoder to introduce spatial bias for better localization of objects in a cluttered 3D scene.

With fully attentional design, we resolve 3D dense captioning with the following innovations: 1) Our method treats the 3D dense captioning task as a set prediction problem. The proposed Vote2Cap-DETR directly decodes the features into object sets with their locations and corresponding captions by applying two parallel prediction heads. 2) We propose a novel vote decoder by reformulating the object queries in 3DETR into the format of the vote query, which is a composition of the embeddings of the seeds point and the vote transformation with respect to the seeds. This indicates the connection between the vote query in Vote2Cap-DETR with the VoteNet, but with better localization and higher training efficiencies; 3) We develop a novel query driven caption head, which absorbs the relation and attribute modeling into self- and cross-attention, so that it can look into both local and global contexts for better scene description. Extensive experiments on two commonly used datasets, ScanRefer and Nr3D, demonstrate that our approach surpasses prior arts with many hand-crafted procedures by a large margin, which demonstrates the superiority that fully transformer architecture with sophisticated vote head and caption head can inspire many 3D vision and language tasks.

To summarize, the main contributions of this work include:

- We propose a novel one-stage and fully attention driven architecture for 3D dense captioning as a set-to-set prediction problem, which achieves object localization and caption generation in parallel.
- Extensive experiments show that our proposed Vote2Cap approach achieves a new state-of-the-art performance on both Nr3D [1] (45.53% C@0.5) and ScanRefer [13] (73.77% C@0.5).

2. Related Work

We briefly summarize works on 3D and video dense captioning, and DETR-based methods for images and 3D point clouds. Additionally, we also introduce some methods for image captioning, which are closely related to our work.

3D and Video Dense Captioning. 3D dense captioning, a task that requires translating 3D scene information to a set of bounding boxes and natural language descriptions, is challenging and has raised great interest among scholars recent years. Scan2Cap [13] and MORE [20] build graph on a detector's [19, 32] box estimations with hand-crafted rules for complex relation reasoning among objects in a 3D scene. SpaCap3D [39] build a spatiality-guided transformer to model spatial relations among the detector's output. 3DJCG [4] and D3Net [7] study the joint promotion of 3D dense captioning and 3D visual grounding. χ-Trans2Cap [43] introduces additional 2D prior to complement information for 3D dense captioning with knowledge transfer. Recently, [48] shifts attention to contextual information for the perception of non-object information. Though these approaches have made great attempts at 3D dense captioning, they all follow a "detect-then-describe" pipeline, which heavily depends on a detector's performance. Our proposed Vote2Cap-DETR differs from existing works in that our method is a one-stage model that detects and generates captions in parallel and treats 3D dense captioning as a set prediction problem. Video dense captioning requires a model to segment and describe video clips from an input video. [40, 49] propose transformer architecture for end-to-end video dense captioning. In this paper, we design elements specially for 3D dense captioning, such as vote queries for better localization in sparse 3D space and the utilization of local contextual information through cross attention for informative object description.

DETR: from 2D to 3D. DEtection **Tr**ansformer(DETR) [5] is a transformer [37] based architecture that treats object detection as a set prediction problem and does not require non-maximum suppression [28] for post-processing. Though great results have been achieved, DETR suffers from slow convergence. Many follow-up works [9, 16, 18, 26, 44, 50] put efforts on speeding up DETR's training by introducing multi-scale features, cross attention designs, and label assignment techniques. Researchers also attempt to introduce transformer architectures to 3D object detection. GroupFree3D [24] learns proposal features from the whole point cloud through the transformer rather than grouping local points. 3DETR [27] analyzes the potential of the standard transformer model and generates proposals by uniformly sampling seed points from a 3D scene. In our work, we extend the DETR architecture for 3D dense captioning that makes caption generation and box localization fully interrelated with parallel decoding. Additionally, we propose

vote query for better performance and faster convergence.

Image Captioning. Image captioning requires a model to generate sentences describing key elements in an image, which has become a hot topic in computer vision. Existing image captioning works adopt an encoder-decoder architecture, where the decoder generates sentences from visual features extracted by the encoder. [2, 14, 17, 30] adopt a detector to extract region features as visual clues for the decoder, while [23,46] extract grid features directly from an image. Additionally, [29] generates captions from both region and grid visual features. Though these methods are effective in image captioning, they cannot be directly applied to 3D dense captioning since it requires describing each 3D object in a scene with respect to its surroundings. In contrast, our proposed caption head sufficiently leverages the rich context information in a 3D point cloud, receives visual clues from both the object query and its local context, and fuses them to achieve effective 3D dense captioning.

3. Method

As shown in Fig. 2, given a 3D scene, our goal is to localize objects of interest and generate informative natural language descriptions for each object. The **input** of our model is a point cloud $PC = [p_{in}; f_{in}] \in \mathbb{R}^{N\times(3+F)}$ representing an indoor 3D scene. Here, $p_{in} \in \mathbb{R}^{N\times 3}$ is the absolute locations for each point, and $f_{in} \in \mathbb{R}^{N\times F}$ is additional input feature for each point, such as *color*, *normal*, *height*, or *multiview feature* introduced by [6, 13]. The expected **output** is a set of box-caption pairs $(\hat{B}, \hat{C}) = \{(\hat{b}_1, \hat{c}_1), \cdots, (\hat{b}_K, \hat{c}_K)\}$, representing an estimation of K distinctive objects in this 3D scene.

Specifically, our system adopts 3DETR [27] encoder as our scene encoder and a transformer decoder to capture both object-object and object-scene interactions through the attention mechanism. Then, we feed the query feature to two parallel task-specific heads for object detection and caption generation.

3.1. 3DETR Encoder

Inspired by DETR [5], 3DETR [27] has made a successful attempt at bringing full transformer architecture to the 3D object detection task, which removes many hand-coded design decisions as the popular VoteNet and PointNet++ modules in most two-stage methods.

In 3DETR encoder, the input PC is first tokenized with a set-abstraction layer [33]. Then, point tokens are fed into a masked transformer encoder with a set-abstraction layer followed by another two encoder layers. We denote the encoded scene tokens as $[p_{enc}; f_{enc}] \in \mathbb{R}^{1,024\times(3+256)}$.

3.2. Vote Query

Though 3DETR has achieved initial success in 3D object detection, it suffers from certain limitations. 3DETR proposes box estimations around the query points (aka proposal centers) sampled from the scenes, which can make these predictions far away from real objects given the sparse object surfaces, resulting in slow convergence to capture discriminative object features with further miss detections.

Prior works on fast convergence DETR models [12, 26, 45] show that injecting more structured bias to initialize object queries, such as anchor points or content-aware queries, accelerates training. Therefore, we propose the vote query, which introduces both 3D spatial bias and content-related information, for faster convergence and performance improvement.

More specifically, we reformulate the object queries in 3DETR into the format of vote query as a composition of the embedding of the reference points and vote transformation around them. This helps build the connection between the object query in 3DETR and the vote set prediction widely studied in VoteNet [32].

The detailed structure is shown in Figure 3. Here, vote Δp_{vote} is predicted from encoded scene token feature f_{enc} with a **F**eed **F**orward **N**etwork (FFN) FFN_{vote} that learns to shift the encoded points to objects' centers spatially:

$$p_{vote} = p_{enc} + \Delta p_{vote} = p_{enc} + FFN_{vote}\left(f_{enc}\right). \quad (1)$$

Then, we sample 256 points p_{seed} from p_{enc} with farthest point sampling and locate each point's offset estimation for $p_{vq} = p_{seed} + \Delta p_{vote}$. Finally, we gather features from (p_{enc}, f_{enc}) for p_{vq} with a set-abstraction layer [33], to formulate the vote query feature $f_{vq} \in \mathbb{R}^{256\times 256}$. We represent vote query as (p_{vq}, f_{vq}).

Following 3DETR [27], our model adopts an eight-layer transformer decoder, and the i-th layer's input query feature f^i_{query} is calculated through

$$f^i_{query} = Layer_{i-1}\left(f^{i-1}_{query} + FFN\left(PE\left(p_{vq}\right)\right)\right), \quad (2)$$

where $f^0_{query} = f_{vq}$, and $PE(\cdot)$ is the 3D Fourier positional encoding function [35]. Experiments in later sections demonstrate that: 1) Vote query injects additional spatial bias to object detection and boosts the detection performance. 2) Encoding features from the point cloud as initial queries accelerates convergence.

3.3. Parallel Decoding

We adopt two task-specific heads for simultaneous object detection and caption generation. The two task heads are agnostic to each other's output.

Detection Head. Detecting objects in a 3D scene requires box corner estimation $\hat{B}$ and class estimation $\hat{S}$ (containing "no object" class) from each object query feature. Following 3DETR [27], box corner estimation is generated by learning spatial offset from a query point to an object's center and box size regression. All subtasks are implemented

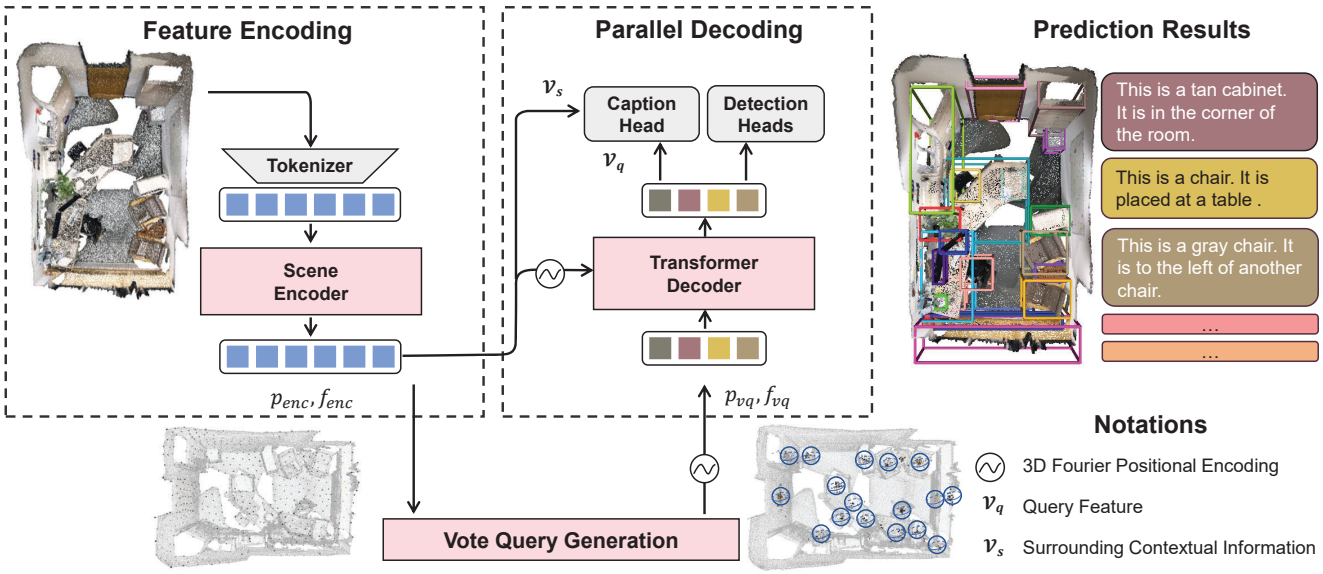

Figure 2. **Approach.** Vote2Cap-DETR is a one-stage transformer model that takes a 3D point cloud as its input, and generates a set of box predictions and sentences localizing and describing each object in the point cloud. The scene encoder first generates encoded scene tokens (p_{enc}, f_{enc}) from the input point cloud. Then, we generate vote query (p_{vq}, f_{vq}) from the encoded scene tokens, which introduce both spatial bias p_{vq} and content-aware feature f_{vq} to initial object queries. The transformer decoder decodes each vote query with two parallel task heads for captioning and detection. We optimize Vote2Cap-DETR with a set loss.

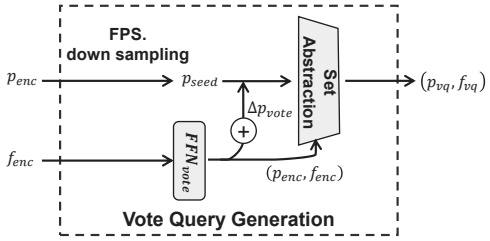

Figure 3. **Vote Query Generation.** Vote query p_{vq} contains spatial bias (Δp_{vote}) to initial object queries (p_{seed}), which are sampled from the scene with farthest point sampling (FPS) and gathered feature f_{vq} from the point cloud for each query.

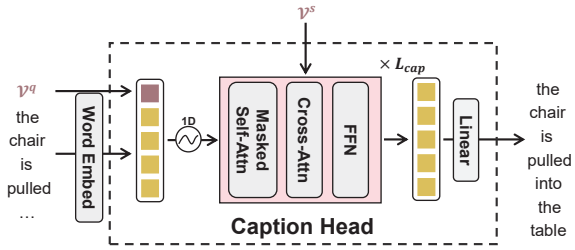

Figure 4. **Dual-Clued Captioner(DCC).** DCC is a lightweight transformer based caption head that uses vote query feature $\mathcal{V}_q$ as caption perfix to identify the region to be described, and contextual features $\mathcal{V}_s$ surrounding the vote query to complement with more surrounding information for descriptive caption generation.

by FFNs. In practice, the object localization head is shared through different layers in the decoder, following all existing works on DETR [5, 12, 26, 27].

Caption Head. 3D dense captioning requires attribute details on an object and its relation with its close surroundings. However, the vote query itself is agnostic to box predictions and fails to provide adequate attribute and spatial relations for informative caption generation. Therefore, the main difficulty is how to leverage sufficient surrounding contextual information without confusing the caption head.

To address the above issues, we propose **D**ual-**C**lued **C**aptioner(DCC), a lightweight transformer decoder-based caption head, for 3D dense captioning. DCC consists of a stack of 2 identical transformer decoder blocks, sinusoid position embedding, and a linear classification head. To generate informative captions, DCC receives two streams of visual clue $\mathcal{V} = (\mathcal{V}^q, \mathcal{V}^s)$. Here, $\mathcal{V}^q$ is the last decoder layer's output feature of a vote query, and $\mathcal{V}^s$ is contextual information surrounding the absolute location of each vote query. When generating a caption for a proposal, we substitute the standard **S**tart **O**f **S**eqenece('SOS') prefix with $\mathcal{V}^q$ to identify the object to be described following [39]. Since the vote query is agnostic of actual neighbor object proposals because of the parallel detection branch, we introduce the vote query's k_s nearest local context token features as its local surroundings $\mathcal{V}^s$ as keys for cross attention. During inference, we generate captions through beam search with a beam size of 5.

3.4. Set prediction loss for 3D Dense Captioning

Our proposed Vote2Cap-DETR requires supervision for vote query $(\mathcal{L}_{vq})$, detection head $(\mathcal{L}_{det})$, and caption head $(\mathcal{L}_{cap})$.

Vote Query Loss. We borrow vote loss from VoteNet [32] as $\mathcal{L}_{vq}$, to help the vote query generation module learn to

shift points p_{enc} to an object's center:

$$\mathcal{L}_{vq} = \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N_{gt}} \left\| p^i_{vote} - cnt_j \right\|_1 \cdot \mathbb{I}\left\{ p^i_{enc} \in I_j \right\}. \quad (3)$$

Here, $\mathbb{I}(\cdot)$ is an indicator function that equals 1 when the condition meets and 0 otherwise, N_{gt} is the number of instances in a 3D scene, M is the size of p_{vote}, and cnt_j is the center of jth instance I_j.

Detection Loss. Following 3DETR [27], we use the same Hungarian algorithm to assign each proposal with a ground truth label. Since 3D dense captioning cares much for the object localization ability, we apply a larger weight on the gIoU loss in set loss [27]:

$$\mathcal{L}_{set} = \alpha_1\mathcal{L}_{giou} + \alpha_2\mathcal{L}_{cls} + \alpha_3\mathcal{L}_{center-reg} + \alpha_4\mathcal{L}_{size-reg}, \quad (4)$$

where $\alpha_1 = 10$, $\alpha_2 = 1$, $\alpha_3 = 5$, $\alpha_4 = 1$ are set heuristically. The set loss $\mathcal{L}_{set}$ is applied to all $n_{dec-layer}$ layers in the decoder for better convergence.

Caption Loss. Following the standard practice of image captioning, we train our caption head first with standard cross-entropy loss (MLE training), and then fine-tune it with **S**elf-**C**ritical **S**equence **T**raining (SCST) [34]. During MLE training, the model is trained to predict the $(t+1)$th word c_i^{t+1}, given the first t words $c_i^{[1:t]}$ and the visual clue $\mathcal{V}$. The loss function for a T-length sentence is defined as:

$$\mathcal{L}_{c_i} = \sum_{i=1}^{T}\mathcal{L}_{c_i}(t) = -\sum_{i=1}^{T}\log \hat{P}\left(c_i^{t+1}|\mathcal{V}, c_i^{[1:t]}\right). \quad (5)$$

After the caption head is trained under word-level supervision, we fine-tune it with SCST. During SCST, the model generates multiple captions $\hat{c}_{1,\cdots,k}$ with a beam size of k and another $\hat{g}$ through greedy search as a baseline. The loss function for SCST is defined as:

$$\mathcal{L}_{c_i} = -\sum_{i=1}^{k}\left(R\left(\hat{c}_i\right) - R\left(\hat{g}\right)\right)\cdot\frac{1}{|\hat{c}_i|}\log\hat{P}\left(\hat{c}_i|\mathcal{V}\right). \quad (6)$$

Here, the reward function $R(\cdot)$ is the CIDEr metric for caption evaluation, and the log probability of caption $\hat{c}_i$ is normalized by caption length $|\hat{c}_i|$ to encourage the model to treat captions in different length with equal importance.

Set to Set Training for 3D Dense Captioning. We propose an easy-to-implement set-to-set training strategy for 3D dense captioning. Given a 3D scene, we randomly sample one sentence from the corpus for each annotated instance. Then, we assign language annotations to the corresponding number of proposals in the corresponding scene with the same Hungarian algorithm. During training, we average losses for captions $\mathcal{L}_{c_i}$ on all annotated instances in a batch to compute the caption loss $\mathcal{L}_{cap}$. To balance loss for different tasks, our loss function is defined as:

$$\mathcal{L} = \beta_1\mathcal{L}_{vq} + \beta_2\sum_{i=1}^{n_{dec-layer}}\mathcal{L}_{set} + \beta_3\mathcal{L}_{cap}, \quad (7)$$

where $\beta_1 = 10$, $\beta_2 = 1$, $\beta_3 = 5$ are set heuristically.

4. Experiments

We first present the datasets, metrics, and implementation details for 3D dense captioning (section 4.1). Then, we provide comparisons with all state-of-the-art methods (section 4.2). We also provide studies on the effectiveness of different parts in our model (section 4.3). Finally, we visualize several qualitative results to address the effectiveness of our method (section 4.4).

4.1. Datasets, Metrics, and Implementation Details

Datasets. We analyze performance on ScanRefer [6] and Nr3D [1], both of which are built on 3D scenes from ScanNet [15]. ScanRefer/Nr3D contains 36,665/32,919 free-form language annotations describing 7,875/4,664 objects from 562/511 out of 1201 3D scenes in ScanNet for training and evaluates on 9,508/8,584 sentences for 2,068/1,214 objects from 141/130 out of 312 3D scenes in ScanNet.

Evaluation Metrics. Following [4, 13, 20, 39], we first apply NMS on object proposals to drop duplicate object predictions. Each object proposal is a box-caption pair $(\hat{b}_i, \hat{c}_i)$, containing box corner prediction $\hat{b}_i$ and generated caption $\hat{c}_i$. Then, each annotated instance is assigned an object proposal with the largest IoU among the remaining proposals. Here, we use (b_i, C_i) to represent an instance's label, where b_i is an instance's box corner label, and C_i is the corpus containing all caption annotations for this instance. To jointly evaluate the model's localization and caption generation capability, we adopt the $m@kIoU$ metric [13]:

$$m@kIoU = \frac{1}{N}\sum_{i=1}^{N} m\left(\hat{c}_i, C_i\right)\cdot\mathbb{I}\left\{IoU\left(\hat{b}_i, b_i\right) \geq k\right\}. \quad (8)$$

Here, N is the number of total annotated instances in the evaluation dataset, and m could be any metric for natural language generation, such as CIDEr [38], METEOR [3], BLEU-4 [31], and ROUGE-L [22].

Implementation Details. We offer implementation details of different baselines. "w/o additional 2D" means the input $\mathcal{PC} \in \mathbb{R}^{40,000\times 10}$ contains absolute location as well as *color*, *normal* and *height* for $40,000$ points representing a 3D scene. "additional 2D" means we replace color information with 128-dimensional *multiview* feature extracted by ENet [8] from 2D images following [13].

We first pre-train the whole network without the caption head on ScanNet [15] for $1,080$ epochs (163k iterations,

$\sim$34 hours) using an AdamW optimizer [25] with a learning rate decaying from 5×10^{-4} to 10^{-6} by a cosine annealing scheduler, a weight decay of 0.1, a gradient clipping of 0.1, and a batch size of 8 following [27]. Then, we jointly train the full model from pre-trained weights with the MLE caption loss for another 720 epochs (51k/46k iterations for ScanRefer/Nr3D, $\sim$11/10 hours). To prevent overfitting, we fix the learning rate of the detector as 10^{-6}, and set that of the caption head decaying from 10^{-4} to 10^{-6} using another cosine annealing scheduler. During SCST, we tune the caption head with a batch size of 2 and freeze the detector for 180 epochs because of high memory cost (50k/46k iterations for ScanRefer/Nr3D, $\sim$14/11 hours) with a fixed learning rate of 10^{-6}. We evaluate the model every $2,000$ iterations during training for consistency with existing works [13, 39], and all experiments mentioned above are conducted on a single RTX3090 GPU.

4.2. Comparison with Existing Methods

In this section, we compare performance with existing works on metrics **C**, **M**, **B-4**, **R** as abbreviations for CIDEr [38], METEOR [3], BLEU-4 [31], Rouge-L [22] under IoU thresholds of 0.25, 0.5 for ScanRefer (Table 1) and 0.5 for Nr3D (Table 2). In both tables, "-" indicates that neither the original paper nor any follow-up works provide such results. We make separate comparisons for MLE training and SCST since different supervisions on the caption head have huge influence on the captioning performance. Among all the listed methods, experiments other than D3Net [7] and 3DJCG [4] utilize the standard VoteNet [32] detector. Meanwhile, D3Net adopts PointGroup [19], a 3D instance segmentation model, for better object detection. 3DJCG substitute the proposal head with an FCOS [36] head to improve VoteNet's localization performance. Additionally, 3DJCG and D3Net are trained on 3D dense captioning as well as 3D visual grounding to study the joint promotion of both tasks. Among methods listed under SCST, χ-Trans2Cap [43] combines MLE training with standard SCST in an additive manner, while Scan2Cap and D3Net [7] adopt the same reward that combines CIDEr score with a listener's [47] grounding loss by weighted summation. Meanwhile, our method adopts the standard SCST with CIDEr reward.

Table 1 reports comparisons on ScanRefer [6] validation dataset. Our Vote2Cap-DETR surpasses current state-of-the-art methods. Under MLE training with additional 2D inputs, Vote2Cap-DETR achieves 59.32% C@0.5 while 3DJCG [4] achieves 49.48% (9.84% C@0.5↑) with additional training data. Under SCST, our Vote2Cap-DETR achieves 70.63% C@0.5 comparing to 62.64% (7.99% C@0.5↑) for current state-of-the-art D3Net [7].

In Table 2, we list results on the Nr3D [1] dataset with additional 2D input following [39]. Since Scan2Cap [13] has not reported results on Nr3D, we adopt the best-reported result from [4]. Our Vote2Cap-DETR also surpasses current state-of-the-art methods (5.78%/7.11% C@0.5↑ for MLE training/SCST).

4.3. Ablation Study

Since 3D dense captioning concerns both localization and caption generation, we perform ablation studies to understand the effectiveness of different components.

Does the vote query improve 3DETR? We performed ablation experiments in Table 3 and Figure 5 to see if the vote query can improve 3DETR's localization and convergence. We notice that introducing position features p_{vq} alone helps improve detection performance (0.97% mAP50↑). However, it (green line in Figure 5) converges slower in the earlier training procedure than the 3DETR baseline (blue line in Figure 5), inferring the vote query generation module is not well learned to predict accurate spatial offset estimations at early training epochs. Introducing additional content feature f_{vq} in vote query features results in another boost in both detection performance (2.98% mAP50↑) and training speed (red line in Figure 5). The overall localization performance of Vote2Cap-DETR is about 7.2% mAP higher than the popular VoteNet.

Figure 5. **Vote query and convergence.** We take out convergence study on a different combination of content feature f_{vq} and position p_{vq} in vote query. The baseline model $(p_{query}, f^0_{query}) = (p_{seed}, \mathbf{0})$ downgrades to 3DETR. Introducing p_{vq} boosts performance but decelerates training since FFN_{vote} requires time to converge, and f_{vq} accelerates training.

Does 3D context feature help captioning? Since the performance of 3D dense captioning is affected by both localization and caption capability, we freeze all parameters other than the caption head and train with 3D only input and standard cross entropy loss (MLE training) for a fair evaluation. We use object-centric decoder [39] as our baseline, which is a decoder that generates captions with object feature as a caption's prefix. In Table 4, "-" refers to the object-centric decoder baseline, "global" means naively including all context tokens extracted from the scene encoder in the decoder, "local" is our proposed caption head that

Method	$\mathcal{L}_{des}$	w/o additional 2D input								w/ additional 2D input							
		IoU = 0.25				IoU = 0.50				IoU = 0.25				IoU = 0.50			
		C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
Scan2Cap [13]	MLE	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
MORE [20]		58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
SpaCap3d [39]		58.06	35.30	26.16	55.03	42.76	25.38	22.84	45.66	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
3DJCG [4]		60.86	**39.67**	27.45	59.02	47.68	31.53	24.28	51.80	64.70	**40.17**	27.66	**59.23**	49.48	31.03	24.22	50.80
D3Net [7]		-	-	-	-	-	-	-	-	-	-	-	-	46.07	30.29	24.35	51.67
Ours		**71.45**	39.34	**28.25**	**59.33**	**61.81**	**34.46**	**26.22**	**54.40**	**72.79**	39.17	**28.06**	**59.23**	**59.32**	**32.42**	**25.28**	**52.53**
χ-Trans2Cap [43]	SCST	58.81	34.17	25.81	54.10	41.52	23.83	21.90	44.97	61.83	35.65	26.61	54.70	43.87	25.05	22.46	45.28
Scan2Cap [13]		-	-	-	-	-	-	-	-	-	-	-	-	48.38	26.09	22.15	44.74
D3Net [7]		-	-	-	-	-	-	-	-	-	-	-	-	62.64	35.68	**25.72**	**53.90**
Ours		**84.15**	**42.51**	**28.47**	**59.26**	**73.77**	**38.21**	**26.64**	**54.71**	**86.28**	**42.64**	**28.27**	**59.07**	**70.63**	**35.69**	25.51	52.28

Table 1. **Evaluating Vote2Cap-DETR on ScanRefer [6].** We compare Vote2Cap-DETR with all published state-of-the-art 3D dense caption methods on the ScanRefer dataset. Though our method does not depend on hand-crafted NMS [28] to drop overlapped boxes, we follow the standard evaluation protocol from [13] for fair comparison and provide evaluation without NMS in Table 7. Our proposed Vote2Cap-DETR achieves new state-of-the-art under both MLE training and SCST.

Method	$\mathcal{L}_{des}$	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap [13]	MLE	27.47	17.24	21.80	49.06
SpaCap3d [39]		33.71	19.92	22.61	50.50
D3Net [7]		33.85	20.70	23.13	53.38
3DJCG [4]		38.06	22.82	23.77	52.99
Ours		**43.84**	**26.68**	**25.41**	**54.43**
χ-Tran2Cap [43]	SCST	33.62	19.29	22.27	50.00
D3Net [7]		38.42	22.22	24.74	54.37
Ours		**45.53**	**26.88**	**25.43**	**54.76**

Table 2. **Evaluating Vote2Cap-DETR on Nr3D [1].** Likewise, we perform the standard evaluation on the Nr3D dataset, and our proposed Vote2Cap-DETR surpasses prior arts.

p_{query}	f^0_{query}	IoU=0.25		IoU=0.50		1st layer IoU=0.50	
		mAP↑	AR↑	mAP↑	AR↑	mAP↑	AR↑
VoteNet Baseline		63.42	82.18	44.96	60.65	-	-
p_{seed}	**0**	67.25	84.91	48.18	64.98	34.80	55.06
p_{vq}	**0**	67.33	85.60	49.15	66.38	30.23	58.44
p_{vq}	f_{vq}	**69.61**	**87.20**	**52.13**	**69.12**	46.53	66.51

Table 3. **Vote query and performance.** We provide quantitative results for Figure 5. Introducing p_{vq} as query positions improves detection, and gathering f_{vq} from content further boosts performance.

includes a vote query's k_s ($k_s = 128$ empirically) nearest context tokens extracted from the scene encoder.

Results show that the caption generation performance benefits from the introduction of additional contextual information. Additionally, compared with naively introducing contextual information from the whole scene, the introduction of local context could be more beneficial. This demonstrates our motivation that close surroundings matter when describing an object.

key	IoU=0.25				IoU=0.5			
	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
-	68.62	38.61	27.67	58.47	60.15	34.02	25.80	53.82
global	70.05	39.23	27.84	58.44	61.20	34.66	25.93	53.79
local	**70.42**	**39.98**	**27.99**	**58.89**	**61.39**	**35.24**	**26.02**	**54.12**

Table 4. **Different keys for caption generation.** We provide a comparison on different keys used in caption generation. Introducing contextual information relates to more informative captions generated. Since 3D dense captioning is more object-centric, introducing vote queries' local contextual feature is a better choice.

Do set-to-set training benefit dense captioning? To analyze the effectiveness of set-to-set training, we use a smaller learning rate (10^{-6}) for all parameters other than the caption head and freeze these parameters during SCST. We name the traditional training strategy as "Sentence Training" adopted in previous works [13, 39], which traverses through all sentence annotations in the dataset. As is shown in Figure 7, our proposed "Set-to-Set" training achieves comparable results with the traditional strategy during MLE training and converges faster because of a bigger batch size on the caption head, which also benefits SCST.

Training	$\mathcal{L}_{des}$	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Sentence	MLE	61.21	**35.35**	26.12	**54.52**
Set-to-Set		**61.81**	34.46	**26.22**	54.40
Sentence	SCST	71.39	37.57	26.01	54.28
Set-to-Set		**73.77**	**38.21**	**26.64**	**54.71**

Table 5. **Set to Set training and performance.** We compare our proposed set-to-set training with traditional "Sentence Training", which traverses through all sentence annotations. We achieve comparable performance with MLE training, and 2.38% C@0.5 improvement with SCST.

End to end training from scratch. Our Vote2Cap-DETR also supports end-to-end training from scratch for 3D dense captioning. However, both ScanRefer and Nr3D are annotated on limited scenes (562/511 scenes) for training; thus, directly training Vote2Cap-DETR from scratch will underperform given to satisfy two objectives simultaneously. Experiments on Scanrefer in Table 6 show that the greedy strategy we choose by pre-training detection head as a good pre-requisite for captioning achieves better performance.

pretrain/end2end	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑	AP@0.5↑	AR@0.5↑
end2end	52.15	28.87	24.68	49.76	46.68	62.17
pretrain+end2end	**62.03**	**34.90**	**26.06**	**54.33**	**51.26**	**67.57**

Table 6. **Ablation study for training strategies.** The greedy strategy we choose by pre-training detection head as a good pre-requisite for captioning achieves better performance than directly end to end training from scratch.

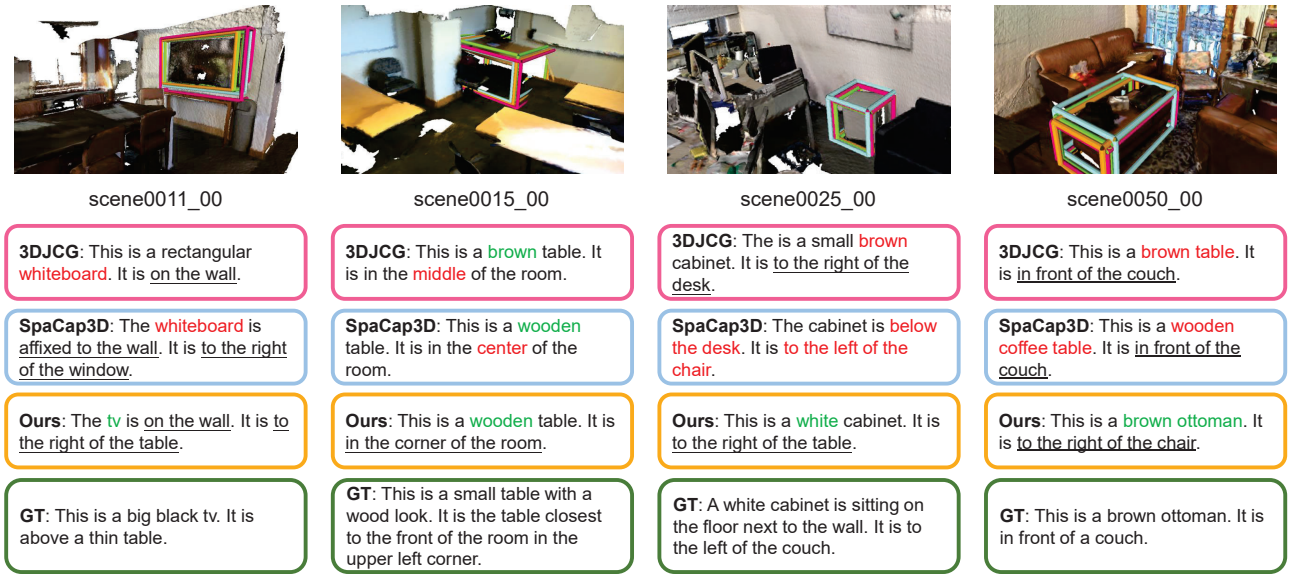

Figure 6. **Qualitative Comparisons.** We compare qualitative results with two state-of-the-art "detect-then-describe" methods, 3DJCG [4] and SpaCap3D [39]. We underline phrases describing spatial locations, and mark correct attribute words in green and wrong description in red. Our Vote2Cap-DETR produces tight bounding boxes close to the ground truth and accurate descriptions.

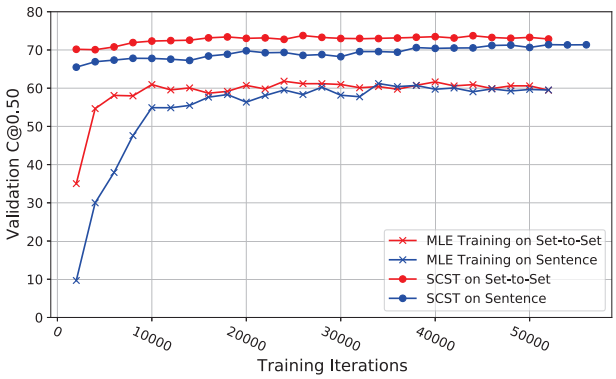

Figure 7. **Set-to-Set training and convergence.** Convergence speed analysis of two different training strategies with MLE training as well as SCST. Set-to-Set training enables a larger batch size for the caption head and accelerates convergence.

Is Vote2Cap-DETR robust to NMS? Similar to other DETR works, the set loss encourages the model to produce compact predictions. We compare performance on both 3D dense caption (C@0.5) and detection (mAP50, AR50) in Table 7. Since the $m@kIoU$ metric (Eq. 8) does not contain any penalties on redundant predictions, getting rid of NMS [28] results in performance growth in C@0.5. Results show that Vote2Cap-DETR is more stable compared to VoteNet based methods with the absence of NMS.

Models	w/ NMS			w/o NMS		
	C@0.5↑	mAP50↑	AR50↑	C@0.5↑	mAP50↑	AR50↑
SpaCap3D	43.93	37.77	53.96	51.35	23.30	64.14
3DJCG	50.22	47.58	62.12	54.94	30.03	68.69
Vote2Cap-DETR	70.63	52.79	66.09	71.57	52.82	67.80

Table 7. **Effect of NMS.** We analyze whether the absence of NMS affects the 3D dense captioning performance (C@0.5) as well as detection performance (mAP50, AR50).

4.4. Qualitative Results

We compare qualitative results with two state-of-the-art models, SpaCap3D [39] and 3DJCG [4] in Figure 6. One can see that our method produces tight bounding boxes close to the ground truth as well as accurate descriptions of object attributes, classes, and spatial relationships.

5. Conclusion

In this work, we present Vote2Cap-DETR, a transformer based one-stage approach, for 3D dense captioning. The proposed Vote2Cap-DETR adopts a full transformer encoder-decoder architecture that decodes a set of vote queries to box predictions and captions in parallel. We show that by introducing spatial bias and content-aware features, vote query boosts both convergence and detection performance. Additionally, we develop a novel lightweight query-driven caption head for informative caption generation. Experiments on two widely used datasets for 3D dense captioning validate that our proposed one-stage Vote2Cap-DETR model surpasses prior works with heavy dependence on hand-crafted components by a large margin.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (No. U1909207, 62071127, and 62276176), Shanghai Natural Science Foundation (No. 23ZR1402900), Zhejiang Lab Project (No. 2021KH0AB05), and in part by A*STAR AME Programmatic Funding A18A2b0046, RobotHTPO Seed Fund under Project C211518008, and EDB Space Technology Development Grant under Project S22-19016-STDP.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, pages 422–440. Springer, 2020. 2, 5, 6, 7

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei 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*, pages 6077–6086, 2018. 3

[3] Satanjeev Banerjee and Alon 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*, pages 65–72, 2005. 5, 6

[4] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong 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*, pages 16464–16473, 2022. 1, 2, 5, 6, 7, 8

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 3, 4

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pages 202–221. Springer, 2020. 3, 5, 6, 7

[7] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. *arXiv preprint arXiv:2112.01551*, 2021. 1, 2, 6, 7

[8] Jintai Chen, Biwen Lei, Qingyu Song, Haochao Ying, Danny Z Chen, and Jian Wu. A hierarchical graph network for 3d object detection on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 392–401, 2020. 5

[9] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 2

[10] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 1

[11] Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (TOG)*, 41(1):1–17, 2021. 1

[12] Xiaokang Chen, Fangyun Wei, Gang Zeng, and Jingdong Wang. Conditional detr v2: Efficient detection transformer with box queries. *arXiv preprint arXiv:2207.08914*, 2022. 3, 4

[13] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 1, 2, 3, 5, 6, 7

[14] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587, 2020. 3

[15] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5

[16] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3630, 2021. 2

[17] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643, 2019. 3

[18] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022. 2

[19] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 2, 6

[20] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. *arXiv preprint arXiv:2203.05203*, 2022. 1, 2, 5, 7

[21] Yongbin Liao, Hongyuan Zhu, Yanggang Zhang, Chuangguan Ye, Tao Chen, and Jiayuan Fan. Point cloud instance segmentation with semi-supervised bounding-box mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10159–10170, 2021. 1

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5, 6

[23] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 3

[24] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[26] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang.

Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 2, 3, 4

[27] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 2, 3, 4, 5, 6

[28] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006. 2, 7, 8

[29] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. *arXiv preprint arXiv:2207.09666*, 2022. 3

[30] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020. 3

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5, 6

[32] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 3, 4, 6

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[34] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017. 5

[35] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 3

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 6

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5, 6

[39] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022. 1, 2, 4, 5, 6, 7, 8

[40] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 2

[41] Chuangguan Ye, Hongyuan Zhu, Yongbin Liao, Yanggang Zhang, Tao Chen, and Jiayuan Fan. What makes for effective few-shot point cloud classification? In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1829–1838, 2022. 1

[42] Fukun Yin, Zilong Huang, Tao Chen, Guozhong Luo, Gang Yu, and Bin Fu. Dcnet: Large-scale point cloud semantic segmentation with discriminative and efficient feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[43] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen 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*, pages 8563–8573, 2022. 1, 2, 6, 7

[44] Chi Zhang, Lijuan Liu, Xiaoxue Zang, Frederick Liu, Hao Zhang, Xinying Song, and Jindong Chen. Detr++: Taming your multi-scale detection transformer. *arXiv preprint arXiv:2206.02977*, 2022. 2

[45] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 3

[46] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15465–15474, 2021. 3

[47] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 6

[48] Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*, 2022. 1, 2

[49] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 2

[50] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2