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

Sijin Chen\textsuperscript{1}\textsuperscript{*}  Hongyuan Zhu\textsuperscript{2}  Xin Chen\textsuperscript{3}  Yinjie Lei\textsuperscript{4}  Gang Yu\textsuperscript{3}  Tao Chen\textsuperscript{1}\textsuperscript{†}

\textsuperscript{1}Fudan University  \textsuperscript{2}Institute for Infocomm Research (I\textsuperscript{2}R) & Centre for Frontier AI Research (CFAR), A\textsuperscript{*}STAR, Singapore  \textsuperscript{3}Tencent PCG  \textsuperscript{4}Sichuan University

\texttt{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 suboptimal 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 DETection TRansformer (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\cite{10, 11, 21, 41, 42}. Among them, 3D dense captioning\cite{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\cite{13}, MORE\cite{20}, and SpaCap3D\cite{39} propose well-designed reasoning modules to model relations among object proposals efficiently.\cite{48} introduces contextual information from two branches to improve the caption. 3DJCG\cite{4} and D3Net\cite{7} study the correlation between 3D visual grounding and 3D dense captioning and point out that these two tasks promote each other. Additionally, \(\chi\)-Trans2Cap\cite{43} discusses how to transfer knowledge from additional 2d information to boost 3d dense captioning.

Among existing methods, they all adopt a two-stage “detect-then-describe” pipeline\cite{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-

\textsuperscript{*}Part of this work was accomplished under supervision by Dr. Hongyuan Zhu from A\textsuperscript{*}STAR, Singapore.

\textsuperscript{†}Corresponding author.
positional neighbors, and post-processing (non-maximum suppression [28]) introduces additional hyper-parameters, leading to 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 spatially-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 Transformer(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, 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. 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 query is predicted from encoded scene token feature $f_{enc}$ with a Feed Forward Network (FFN) $FFN_{vote}$ that learns to shift the encoded points to objects’ centers spatially:

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

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

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

$$f_{query}^i = Layer_{i-1} \left( f_{query}^{i-1} + FFN \left( PE(p_{vote}) \right) \right), \quad (2)$$

where $f_{query}^0 = f_{query}$, 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 $B$ and class estimation $S$ (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...
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 **Dual-Clued Captioner (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 clues $V = (V^v, V^s)$. Here, $V^v$ is the last decoder layer’s output feature of a vote query, and $V^s$ is contextual information surrounding the absolute location of each vote query. When generating a caption for a proposal, we substitute the standard Start Of Sequence (‘SOS’) prefix with $V^q$ 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 $V^s$ 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 ($L_{vq}$), detection head ($L_{det}$), and caption head ($L_{cap}$).

**Vote Query Loss.** We borrow vote loss from VoteNet [32] as $L_{vq}$, to help the vote query generation module learn to
shift points $p_{enc}$ to an object’s center:

$$
L_{vq} = \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N_{gt}} \| p_{vq}^{i} - cnt_{j} \|_1 \cdot I \{ p_{vq}^{i} \in I_{j} \} .
$$

(3)

Here, $\| \cdot \|$ 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 $p_{vq}$, and $cnt_{j}$ is the center of $j$th 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]:

$$
L_{set} = \alpha_{1} L_{gious} + \alpha_{2} L_{cls} + \alpha_{3} L_{center-reg} + \alpha_{4} L_{size-reg},
$$

(4)

where $\alpha_{1} = 10$, $\alpha_{2} = 1$, $\alpha_{3} = 5$, $\alpha_{4} = 1$ are set heuristically. The set loss $L_{set}$ 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 Self-Critical Sequence Training (SCST) [34]. During MLE training, the model is trained to predict the $(t + 1)$th word $c_{t+1}$, given the first $t$ words $c_{1:t}$ and the visual cue $\mathcal{V}$. The loss function for a $T$-length sentence is defined as:

$$
L_{c} = \sum_{t=1}^{T} L_{c_{t}}(t) = - \sum_{t=1}^{T} \log \hat{P} \left( c_{t+1}^{i} | \mathcal{V}, c_{1:t}^{i} \right).
$$

(5)

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

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

(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 $L_{c_{i}}$ on all annotated instances in a batch to compute the caption loss $L_{cap}$. To balance loss for different tasks, our loss function is defined as:

$$
L = \beta_{1} L_{vq} + \beta_{2} \sum_{i=1}^{n_{dec-layer}} L_{set} + \beta_{3} L_{cap},
$$

(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 $(b_{i}, \hat{c}_{i})$, containing box corner prediction $b_{i}$ 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]:

$$
m@kIoU = \frac{1}{N} \sum_{i=1}^{N} m(\hat{c}_{i}, C_{i}) \cdot \mathbb{1} \{ IoU \left( b_{i}, b_{i} \right) \geq k \} .
$$

(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 $\mathcal{P} \in \mathbb{R}^{10,000 \times 10}$ 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, 11128
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.

We performed ablation studies on different combinations of content feature $f_{vq}$ and position $p_{vq}$ in vote query. The baseline model $(p_{query}, f_{vquery}) = (p_{accept}, 0)$ 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

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, 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 SCST in an additive manner, while Scan2Cap and D3Net achieve 70.63% C@0.5 comparing to 62.64% (7.99% C@0.5↑) for current state-of-the-art 3DNet [7].

In Table 2, we list results on the Nr3D [1] dataset with additional 2D input following [39]. Since Scan2Cap [13]
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.

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.

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

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\(^{-4}\)) 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.

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.
4.4. Qualitative Results

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.

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


[47] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with improved denoising anchor boxes for end-to-end object detection, 2022. 2

