GAMES 003 科研素养课

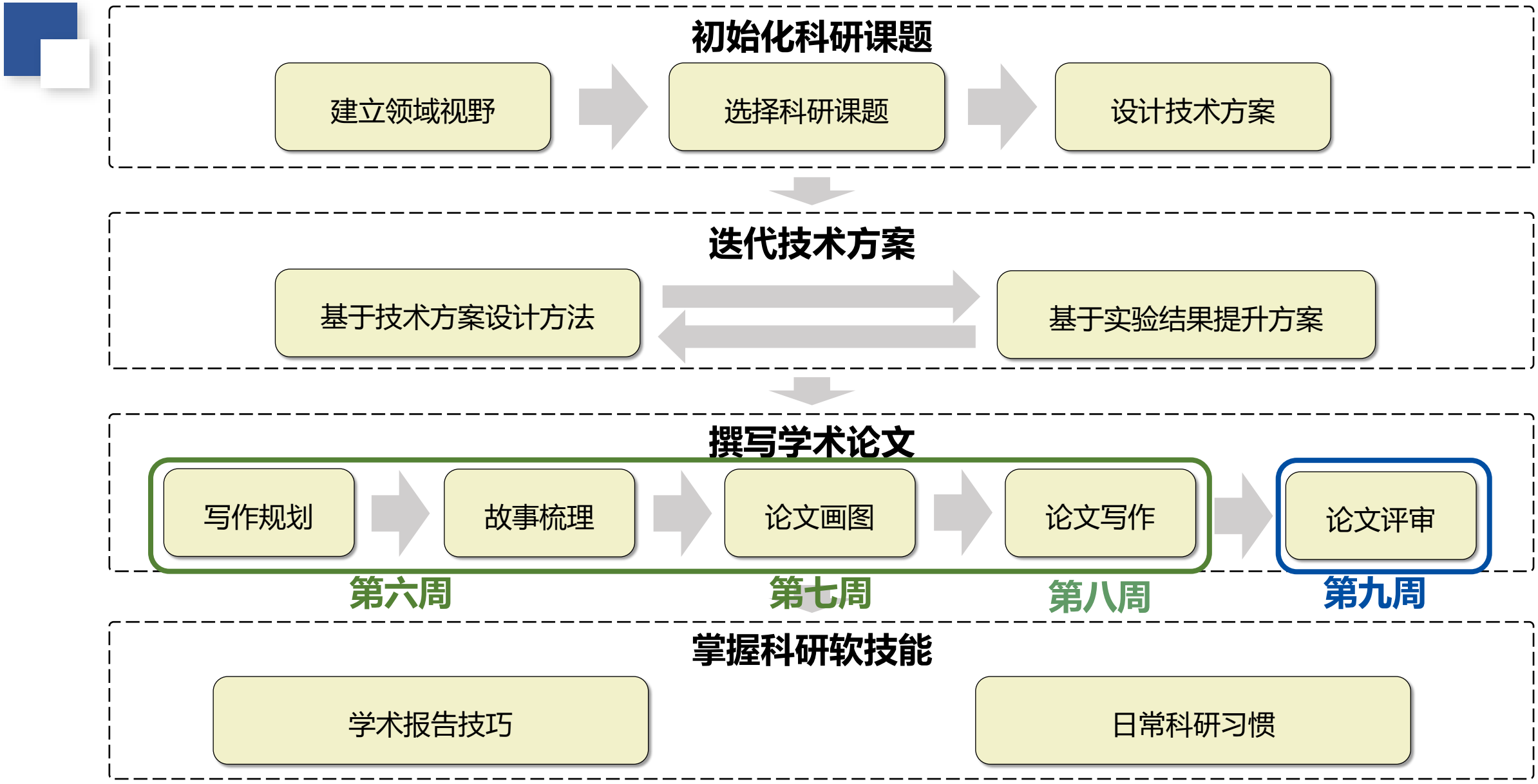
第九周：论文Rebuttal策略

Sida Peng

Jun Gao

Songyou Peng

Qianqian Wang


初始化科研课题
建立领域视野
选择科研课题
设计技术方案
迭代技术方案
基于技术方案设计方法
基于实验结果提升方案
撰写学术论文
写作规划
故事梳理
论文画图
论文写作
论文评审
第六周
第七周
第八周
第九周
掌握科研软技能
学术报告技巧
日常科研习惯

资源

除了和其他讲者（思达，高俊，倩倩）一起讨论的内容，本堂课的内容主要来源于：

- **Devi Parikh**: **How we write rebuttals**
- **Rana Hanocka**: Writing a rebuttal for SIGGRAPH
- **Fredo Durand**: Rebuttal advice
- **Aaron Hertzmann**: Technical Paper Rebuttals Arent Just For "Factual Errors"

为什么要Rebuttal?

论文 Rebuttal 为什么重要

- Rebuttal 就是帮助**澄清误解**并**说服读者**的一个方式，最终让文章可以被接受

- 需要说服的人：**评审人** 和 **领域主席 (AC)**

- **评审人**阅读过你的论文（阅读深度各有不同），通常都会对文章有一些**误解**
  - 可能因为文章写得不清楚
  - 可能因为评审人看得不仔细（每个评审人花在评审文章的时间都是有限的）

- **领域主席 (AC)** 对你的工作更不熟悉
  - 一个指导原则：假设他们只会阅读评审意见和你的回复

目标

- **评审人**：澄清疑问，回答问题，纠正误解，用好的方式反驳错误的描述，并以诚意接纳反馈，改进你的工作。
- **领域主席 (AC)**: 知道你做出的努力，列出评审意见的代表性总结，帮助他们了解评审人的问题是否得到了回应，指出恶意的评审行为，帮助他们做出决定。
- 让评审人和AC相信你的**论文优点足够多**！

绝大部分新人只关注评审人，而忽略领域主席 (AC)，但是他们才是最终的裁判

如何Rebuttal？

内容大部分来自于: How we write rebuttals. Devi Parikh

Rebuttal的指导思想

假设一位中立的**第三方**，是否能够**仅通过你的回复**（**无需**再次**阅读论文或评审意见**），判断评审人的问题**得到了准确的回应**，并**相信**你文章**有足够的优点**？

评审和Rebuttal方式 (两个典型)

一个PDF回复所有评审人

(CVPR, ICCV, SIGGRRAPH...)

CVPR #12552

CVPR #12552

Rebuttal - NeRF *On-the-go*: Exploiting Uncertainty for Distractor-free NeRFs in the Wild

We thank all reviewers ([R1, cdX4], [R2, njxJ], and [R3, ARpQ]) for their insightful feedback. The reviewers agree on the efficacy of our method, noting its *quite good results* (R1, R2, R3), its *simple yet effective* nature, and its *ease of integration with current methods* (R1, R3). Additionally, the *interesting and inspirational analysis* (R2, R3), along with R3's acknowledgment of our *interesting story and task*, further highlights the impact of our work. We will incorporate all feedback (additional comments and missing references) and address the main concerns in the following.

Paper contributions (R1, R2). We emphasize that our work pioneers a highly simple, yet versatile and robust module, designed for easy integration into any NeRF pipeline, as recognized by R1 & R3. We significantly enhance NeRF's applicability to casually captured data in various scenarios. As highlighted in Michael Black's article on novelty in science, we consider our method's simplicity and effectiveness itself as a key contribution. Within this, we have three key innovations: 1) Use DINO features for accurate uncertainty prediction; 2) Replace L2 with SSIM-based loss for enhanced uncertainty learning; 3) Dilated patch sampling for fast and effective distractor removal. Furthermore, R3's acknowledgment of our On-the-go dataset, highlighting its potential to *accelerate new avenues for NeRF* further underscores the innovative nature of our work.

Differences to NeRF-W (R1). We claim that our method

Figure 1. **Additional comparisons with HA-NeRF and S3IM sampling strategy on the Patio-High scene.** LPIPS metrics are included.

Additional baseline (R1). Thanks the reviewer for the suggestion. We additionally compare with Ha-NeRF in Fig. 1. More baseline comparisons will be provided in the paper.

Failure cases (R1, R3). Similar to baselines, we struggle in regions with strong view-dependent effects, see Fig. 2. Moreover, inherited from the limitation of our base model Mip-NeRF360, we also require sufficient training views. We will include discussions with more failure cases.

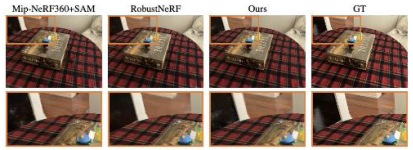


Figure 2. **Failure cases.**

S3IM discussion (R2). We respectfully clarify that our dilated SSIM strategy is distinct from S3IM. Firstly, S3IM is

单独回复每个评审人

(NeurIPS, ICLR...)

Response to Reviewer nyEb

NeurIPS 2021 Conference Paper2229 Authors *Songyou Peng (privately revealed to you)*

09 Aug 2021, 06:52 (modified: 10 Aug 2021, 03:45) NeurIPS 2021 Conference Paper2229 Official Comment Readers: Everyone Show Revisions

Comment:

We thank the reviewer for the constructive feedback. We appreciate that the reviewer finds our paper promising, novel, and well-written. We address additional comments below.

None of PSGN, 3D-R2N2 and AtlasNet models were designed for surface reconstruction from input unoriented points, so they were not optimized for this task

We propose SAP as a novel shape representation. Therefore, we find it important to compare SAP against different established shape representations with PSGN representing point clouds, 3D-R2N2 voxels and AtlasNet meshes. As all of these methods are encoder-decoder approaches, this comparison provides a fair analysis of the quality of each output representation. We further remark that other related works (e.g., Occupancy Networks) use a similar evaluation protocol.

Compare to Points2Surf, ECCV'20

We thank the reviewer for the suggestion. For this rebuttal, we evaluated Points2Surf on the ShapeNet test set with 3K input points (noise level=0.005) and obtained the following results which we will include in the paper:

	Chamfer-L1	F-Score	Normal C.
Points2Surf	0.068	0.808	0.857
Ours	**0.034**	**0.975**	**0.944**

Note that Points2Surf requires roughly 75 seconds for inference of a single shape at a resolution of 128^3, while our method requires only 0.07 seconds.

Remark: As training of Points2Surf requires over one week on 4 RTX 2080Ti and since the authors mention that "Points2Surf is patch-based and therefore independent from classes", we use the model provided by the authors for this experiment. For the final version of the paper we will also add the results of a model retrained on our data.

Concern is mainly equation 7 from the supplementary. Equations 5 and 6 partly justify it but do not account for the appearance of $\tilde{g}$. It is said that it is introduced to mitigate Gibbs phenomenon, but it is not obvious why the addition of $\tilde{g}$ to the solution of Poisson equation in frequency domain do not corrupt this solution (there will not be an identity if one put $\tilde{\chi}$ in FFT of Poisson equation)

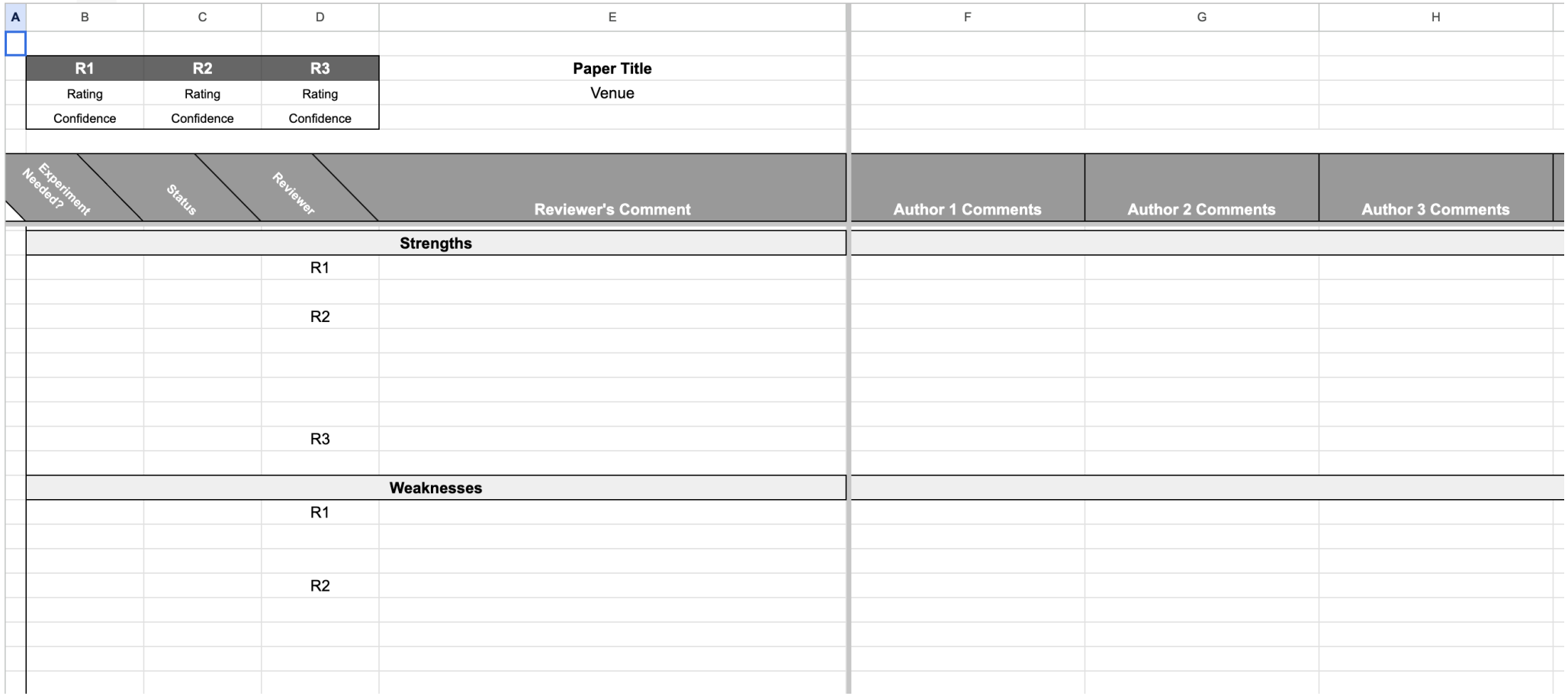
1. *Why we need a Gaussian*: The Gaussian serves as a regularizer to the smoothness of the solved implicit function. Not using a Gaussian is equivalent to using a Gaussian kernel with sigma = 0. Please refer to Fig. 3 in the supplementary which motivates the use of our sigma parameter. We will rephrase Eqn. 5, 6, 7 in the supplementary to further discuss and motivate the inclusion of the Gaussian term.
2. *Why using a Gaussian in the spectral domain*: First, the FFT of a Gaussian remains a Gaussian. Second, convolution of a Gaussian in the physical domain is equivalent to a dot product with a Gaussian in the spectral domain and a dot product in the spectral domain is more efficient than convolution in the physical domain: $O(NlogN)$ vs $O(N^2)$, where n is the resolution of a regular grid and $N = n^3$. We will clarify this in the paper.

一个从零开始Rebuttal的四步法

1. 列出所有评审意见
2. 脑暴每个问题可能的回复
3. 撰写Rebuttal的草稿
4. 不断修改

1.列出所有评审意见

可以使用一个电子表格来整理每位评审人提出的每个点。尽早完成这一步，以便**及早识别任何必要的实验**



	R1	R2	R3	Paper Title
	Rating	Rating	Rating	Venue
	Confidence	Confidence	Confidence	

Experiment Needed?	Status	Reviewer	Reviewer's Comment	Author 1 Comments	Author 2 Comments	Author 3 Comments
Strengths						
		R1				
		R2				
		R3				
Weaknesses						
		R1				
		R2				

电子表格模版可以在这里获得

1.列出所有评审意见

可以使用一个电子表格来整理每位评审人提出的每个点。尽早完成这一步，以便**及早识别任何必要的实验**

	R1	R2	R3	NICE-SLAM: Neural Implicit Scalable Encoding for SLAM			
	Rating	Rating	Rating	CVPR 2022			
	WA	BL	WA	Paper ID: 6327			
	Reviewer	Experiment Needed?	Status	Reviewer's Comment	Zihan Comments	Songyou Comments	Viktor Comments
	R3			The paper proposed an elegant learning-based framework for RGB-D dense SLAM system. It shows potentials in dealing with challenges, such as real-time performance, scalable mapping, predictive power, and robustness, we face in developing SLAM systems working in real situations.			
				The basic representations using hierarchical, grid-based neural implicit encoding provide a practical approach to get trade-off between global consistency and local processing, a problem we should solve for general dynamic vision systems.			
				Weaknesses			
	R1	☐	☐	The main weakness in the reviewer's opinion is that using a hierarchical voxel grid encoding latent codes with fixed MLP as decoder, may be seen as an expected extension to IMAP, given prior work such as Local Deep SDF, PlenOctrees and many others, including the ones the authors mention themselves. Although the reviewer values that the authors have accomplished a real-time system of this nature.			
		☐	☐	Additionally, some design choices could be better justified. Why only 3 scales in the map? And why the middle occupancy layer is not a residual of the coarse one?			
		☑	☐	Other weaknesses include the lack of comparison with classical reconstruction methods, such a Kinect Fusion both qualitatively and quantitatively (in terms of reconstruction, not position, accuracy).			
		☑	☐	The authors stress many times the relevance neural methods have in predicting unobserved regions, however few qualitative experiments are presented showing whether this is actually true.			
		☑	☐	I believe some quantitative experiment could be devised to study this, for example in the context of depth completion one could render artificial views and compare the accuracy of the completed maps against the rendering of a classical method such as Kinect Fusion where incomplete depthmap maps are filled using a classical method such as the bilateral			

一个例子：**NICE-SLAM**

2. 脑暴每个问题可能的回复

每位作者自由的填写自己的**初步回复想法。**收集所有人意见，这里不用考虑篇幅，之后再做“减法”

Reviewer comment	Author 1	Author N
1. There existing several related papers discussion of the using human attention map in image captioning and visual question answering. For example, (1) Liu et al. Attention correctness in neural image captioning. (2) Qiao et al. Exploring human-like attention supervision in visual question answering. Please illustrate the differences with these papers.	The papers mentioned provide attention supervision over the attention layer. Our central argument for this and the next point will be that Grad-CAM is more faithful than attention. In order to show this I am planning on doing occlusion studies in the proposal space and compare that with the attention weights and the Grad-CAM proposal importance weights. Also with attention supervision only the layers before the attention layer can be updated, but with HINT all layer weights can be updated. Include lines from the paper. Att supervision doesn't work.	We should first very clearly say what you say in the first sentece of your response. And expand on that a bit if needed to make the point clearly. Does the paper differentiate our work from these work R2 cites or other such works? If so, we should clearly say in the response "As discussed in LXYZ-ABC..." You can then make the point about which layers can be updated. We can then additionally make the "central argument" point. But the direct response should be clear / not confused with the description of a new experiment and such.
2. It seems that the ground-truth attention map is used for the VQA task. For the captioning task, although no ground-truth attention map is used, the segmentation maps are used. As such compare with other methods, strong information about the image are incorporated, which should results in performance improvements.	Human attention or segmentation maps are used only during training and not during testing. While we agree that this is extra information used during training, we show why other approach fail to utilize this information to achieve improvements in performance during test time. Only a fraction of images in VQA have Human attention. Also if it is possible to such a good boost with just human attention, people would start collecting. Also HATs are important to know if models are making the right decision for the right reasons.	"we show why other approach fail to utilize this information to achieve improvements in performance during test time." You'll have to point to a specific experiment in the paper / lines in the paper / table in the paper and reproduce the curcial numbers here to support this claim. Then you can say this is only at training time, not at test time. (I think the reviewer already knows this. So starting with this response is not a strong start.)
The method to set the ground truth importance scores seems hacky especially for image captioning. As I can imagine there shall be multiple objects in the same category and the HINT supervision will highlight all of them during generating the word. For example, assuming there are 3 people in a park and only 1 person is throwing a frisbee. The ground truth caption is 'A man is throwing a frisbee.' It is not appropriate to highlight all of the 3 people.	I completely agree. This problem does exist due to the way we use annotations for captioning. Mention that this is a first step and such cases although infrequent would make the model look at more than correct regions. In future work we plan on addressing such scenarios, basically modifying the loss that makes the model get heavily penalized if it places mass on incorrect regions, and penalize it not so much if it misses some regions which exists in the segmentation. This would make us use the same amount of supervision but address such scenarios pointed by R3	We can also say that this allowed us to use existing annotations that were collected for a different task, which is nice.
The author clearly states the importance of aligning the important region, however, the reason why aligning the gradient-based explanation can be better is not clear and detailed analyzed.	The above experiment showing that simple attention is not entirely faithful to the model, and gradient based explanation is more faithful, will help answer this comment. Also I think its important to state that using Attn. Supervision, later layer parameters cannot be updated, but with HINT they can be, as Network Importance is a function of all the weights of the network.	If you think that experiment helps here more than in the earlier comment, maybe mention it here and not there so the earlier response is cleaner? Not sure.. your call. Or maybe it is better to club this and the earlier comment (and the next one) into one response (while being clear in the rebuttal that it is in response to all three).

3. 撰写Rebuttal的草稿

将表格中的共识**转化为具体的回复**, 保持简洁，但**确保涵盖每个要点**， 不要过于担心篇幅限制

Reviewer comment	Author 1	Author N
The method to set the ground truth importance scores seems hacky especially for image captioning. As I can imagine there shall be multiple objects in the same category and the HINT supervision will highlight all of them during generating the word. For example, assuming there are 3 people in a park and only 1 person is throwing a frisbee. The ground truth caption is 'A man is throwing a frisbee.' It is not appropriate to highlight all of the 3 people.	I completely agree. This problem does exist due to the way we use annotations for captioning. Mention that this is a first step and such cases although infrequent would make the model look at more than correct regions. In future work we plan on addressing such scenarios, basically modifying the loss that makes the model get heavily penalized if it places mass on incorrect regions, and penalize it not so much if it misses some regions which exists in the segmentation. This would make us use the same amount of supervision but address such scenarios pointed by R3	We can also say that this allowed us to use existing annotations that were collected for a different task, which is nice.

Thank you for your insightful feedback regarding our method for setting ground truth importance scores in image captioning. **We agree that** our current approach may not perfectly handle situations where multiple instances of the same object category are present……This limitation arises from…… We view our work as a first step toward integrating visual grounding into image captioning……

Specifically, we aim to design a loss that heavily penalizes the model when it assigns importance to incorrect regions……This approach would allow us to maintain the same level of supervision while improving the model's ability to focus on the most relevant regions……

Additionally, leveraging existing annotations collected for different tasks has the advantage of reducing the need for additional annotation efforts. This not only makes our approach more efficient but also demonstrates the versatility of these datasets in contributing to multiple areas of research……

4. 不断修改

不断**重新阅读草稿**和表格内容，确保所有问题都得到了回应。优先处理主要的要点，并开始逐步调整以满足篇幅限制。

Rebuttal的各种建议

开门见喜

先突出评审人对你工作的**正面评价**。虽然回复意见的重点通常是回应负面的看法，不要让领域主席（AC）在这个过程中忘记你工作的优点和亮点。

We thank all reviewers for their insightful comments. **R1** says the paper has *"a novel and interesting idea"* and *"very impressive results"*. **R2** says *"the contribution is interesting"* and *"has practical benefits"*. **R3** *"really likes the idea"* and and suggests it provides *"guidelines for future work"*. We will incorporate all feedback and suggested references. The following addresses other comments.

We thank reviewers for their insightful and positive feedback! We are encouraged that they find EmbodiedQA to be a novel task (R1, 2, 3), an important research problem (R1, 2), appropriately positioned w.r.t. prior work (R1, 3), the dataset thoughtfully created to avoid biases (R3) and of value to the community (R1, 2, 3), and the proposed methods reasonable (R3) and elegant (R2). One primary concern

问题引用

一个PDF回复所有评审人

(CVPR, ICCV, SIGGRRAPH...)

做法：简要概括问题

CVPR #12552

CVPR 2024 Submission #12552. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

CVPR #12552

Rebuttal - NeRF *On-the-go*: Exploiting Uncertainty for Distractor-free NeRFs in the Wild

We thank all reviewers ([R1, cdX4], [R2, njxJ], and [R3, ARpQ]) for their insightful feedback. The reviewers agree on the efficacy of our method, noting its *quite good results* (R1, R2, R3), its *simple yet effective* nature, and its *ease of integration with current methods* (R1, R3). Additionally, the *interesting and inspirational analysis* (R2, R3), along with R3's acknowledgment of our *interesting story and task*, further highlights the impact of our work. We will incorporate all feedback (additional comments and missing references) and address the main concerns in the following.

Paper contributions (R1, R2). We emphasize that our work pioneers a highly simple, yet versatile and robust module, designed for easy integration into any NeRF pipeline, as recognized by R1 & R3. We significantly enhance NeRF's applicability to casually captured data in various scenarios. As highlighted in Michael Black's article on novelty in science, we consider our method's simplicity and effectiveness itself as a key contribution. Within this, we have three key innovations: 1) Use DINO features for accurate uncertainty prediction; 2) Replace L2 with SSIM-based loss for enhanced uncertainty learning; 3) Dilated patch sampling for fast and effective distractor removal. Furthermore, R3's acknowledgment of our On-the-go dataset, highlighting its potential to *accelerate new avenues for NeRF* further underscores the innovative nature of our work.

Differences to NeRF-W (R1). We claim that our method

Figure 1. **Additional comparisons with HA-NeRF and S3IM sampling strategy on the Patio-High scene.** LPIPS metrics are included.

Additional baseline (R1). Thanks the reviewer for the suggestion. We additionally compare with Ha-NeRF in Fig. 1. More baseline comparisons will be provided in the paper.

Failure cases (R1, R3). Similar to baselines, we struggle in regions with strong view-dependent effects, see Fig. 2. Moreover, inherited from the limitation of our base model Mip-NeRF360, we also require sufficient training views. We will include discussions with more failure cases.

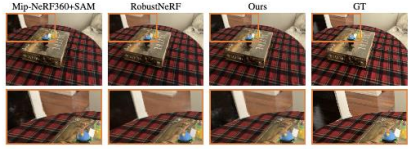


Figure 2. **Failure cases.**

S3IM discussion (R2). We respectfully clarify that our dilated SSIM strategy is distinct from S3IM. Firstly, S3IM is

单独回复每个评审人

(NeurIPS, ICLR...)

做法：直接引用整体问题

> **Q1: Given two views (which is the only input setting shown in the paper), the method still predicts per-pixel per-view 3DGS fields in a common coordinate (similar to Splatt3r). This is similar to Dust3r pointcloud variants where the pointclouds are predicted in a common coordinate space. I fail to see how this part is different from Dust3r-style prediction. Can you explain how this approach differs from those in the 2-view setting, if at all?**

The key difference lies in the **scene representation** itself. In our method, we predict 3DGS fields in a canonical coordinate space, whereas DUSt3R/MASt3R predicts a global point map. This distinction fundamentally impacts the focus and applications of these methods:

- DUSt3R/MASt3R is designed to output point maps for downstream tasks such as point matching and pose estimation.
- In contrast, our work investigates how canonical Gaussian prediction can advance novel view synthesis from unposed images, making it a distinct focus and contribution in this domain.

> **Q9: Table 4: why might be the Splatt3r numbers lower than those reported in the original paper?**

Similar to Q8, the discrepancy in Splatt3R's numbers arises from differences in the evaluation protocols. Specifically, we have chosen more varied overlap ratios, thus making the evaluation more challenging. As a result, the numbers reported for Splatt3R in Tab. 4 differ from those in their original paper.

Additionally Requested Experiments

> **1. On object-level, train the proposed model using the same data with the previously mentioned pose-free works and compare their performance. This also verifies whether MAST3R weight initialization generalizes to object-level data.**

Thank you for your suggestion. We agree that such an experiment can further verify the generalizability of our method to object-level reconstruction tasks.

The reviewer mentioned two pose-free object-level reconstruction works: PF-LRM and LEAP. Unfortunately, PF-LRM has not open-sourced its code or dataset, making a direct comparison infeasible. Therefore, we conducted a comparison with LEAP, the state-of-the-art open-source pose-free method for object-level reconstruction.

To ensure a fair comparison, we trained and evaluated both methods on the widely adopted Objaverse dataset, maintaining identical training iterations and input image resolutions. The results, shown in the table below, demonstrate that our approach significantly outperforms LEAP, with the PSNR improving from 20.559 to 28.378. This highlights the robust generalization capabilities of our method for object-level reconstruction. The qualitative comparison is presented in Fig. 9 of our revised manuscript.

	PSNR	SSIM	LPIPS
LEAP	20.559	0.853	0.144
Ours	28.378	0.935	0.053

对细节问题直接回应

对细节问题应该首先直接对其作出回应，接下来再提供细节、描述背景，或者解释你的立场。这样的回复更清晰有针对性。

Q: “Are these averaged across multiple runs?”
A: **Yes**, we averaged across 5 random seeds…

Q: “Are the segmentation masks used during training?”
A: **No,** they are only used to evaluate our results…

Q: “Why did you not compare to GMAP?“
A: **GMAP is prohibitively expensive in our setting**. Our environments have a significantly larger state-space…. it would take 128 GPUS for 3 months to evaluate GMAP.

就事论事

很多同学会答非所问，或者喜欢除了问题本身以外的一些不直接相关的点

Q: “表4中，为什么XX方法的数值比他们原始论文中的数值更低？”

回答: 差异来源于评估的数据的不同。虽然使用同一个基础数据集，但是我们选择了更加多样化的重叠比例，从而使评估更加具有挑战性，更能体现方法的有效性。

我们还发现，XX使用的测试集并不适合评估具有泛化能力的新颖视图合成，因为它们的目标视图内容与两个输入视图之间的重叠不够充分。相对而言，我们确保目标视图大多位于两个输入视图之间，并使用我们的测试集重新评估XX。为了在所有方法之间进行公平比较，我们使用自己的评估输入对所有方法进行测试。

BAD!

如果评审人没有提到某个问题，不要主动提出来！

回应问题背后的意图

有时候需要揣测评审意见中具体问题背后的意图，对背后意图进行作答

Q: “为什么你没有在A数据集上进行评估？”

背后的意图: 可能实质是在质疑你的实验设计。对此，你可以直接回答问题 (比如这个数据集为什么不好不适合我们的task，或者直接在这个数据集做实验)。接下来，应该说明你已经在 B, C 和 D 数据集上进行了评估，并且这些实验已经足够充分证明文章想表达的点。

这不仅是对问题的直接回应，也能提醒评审人和AC你已经进行了广泛的实验评估，从而避免他们因为这评论产生错误印象。

铺垫背景

如果所有评审人似乎都忽略了一个核心观点，可以简洁明了地重述这一要点

Recap: What is our goal? Enable researchers to use simulation for *evaluation* with confidence that their results will generalize to real robots.

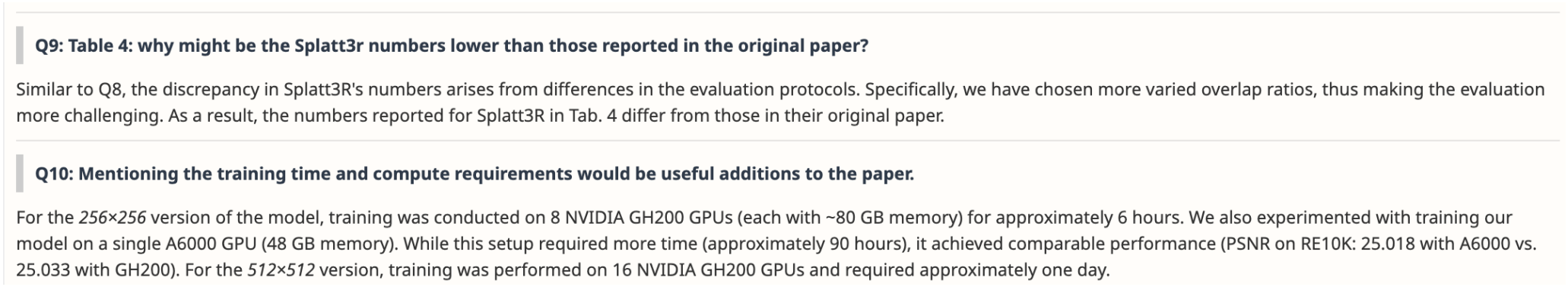
Recap: Why this goal? Evaluation on real robots is slow, dangerous, costly, and difficult to reproduce (L109-111).

Recap: What is *not* the goal? To improve the state-of-art on embodied navigation or to develop a new sim2real transfer method (L393-413). Those are important problems; but not the goal of this paper.

保持内容自成一体

假设评审人和AC对你的论文记忆不深 （几乎所有人过了一两个月都会忘文章细节），并且可能不会再次详细阅读。

- 重新介绍任何缩略词
- 提醒他们实验设置的相关细节
- 确保你的回复在直接阅读时也能被轻松理解 (例如前面提到的示例)
- 如果没有回复字数数量限制 (比如ICLR, NeurIPS)， 也可直接把问题完全粘贴过来



> **Q9: Table 4: why might be the Splatt3r numbers lower than those reported in the original paper?**

Similar to Q8, the discrepancy in Splatt3R's numbers arises from differences in the evaluation protocols. Specifically, we have chosen more varied overlap ratios, thus making the evaluation more challenging. As a result, the numbers reported for Splatt3R in Tab. 4 differ from those in their original paper.

> **Q10: Mentioning the training time and compute requirements would be useful additions to the paper.**

For the *256×256* version of the model, training was conducted on 8 NVIDIA GH200 GPUs (each with ~80 GB memory) for approximately 6 hours. We also experimented with training our model on a single A6000 GPU (48 GB memory). While this setup required more time (approximately 90 hours), it achieved comparable performance (PSNR on RE10K: 25.018 with A6000 vs. 25.033 with GH200). For the *512×512* version, training was performed on 16 NVIDIA GH200 GPUs and required approximately one day.

保持内容自成一体

反例：

Weaknesses:

Unclear descriptions of target feature processing in Sec 4.1

1. How do you know if a 3D point belongs to the object, or the
2. For the background features, you will get only a single featu
3. Also, isn't it making more sense to take per-pixel CLIP featu

Unclear descriptions of neural scene encoding in Sec 4.2

1. Related to the questions above. In this section you mention hierarchy. However, how exactly do you learn these two set
2. To learn MHE or NeRF in general, you need to actually shoo rendering. How do you make sure your features on the 3D

Unclear Topometric Mapping in Sec 4.3

1. Line 309, what is C_t, S_t? What are the differences to C_R, S_R

Official Comment by Authors

Official Comment by Authors — Everyone

Comment:

Thank you for showing interest in our novel idea and carefully reading to make this paper better. Sorry for the Ambiguity, we carefully apply proof-reading and formulation clarification on the methodology and the revised sections is highlighted red in rebuttal pdf.

For ambiguity problems, the pixel-wise encoding strategy is clarified in Section 4.1; neural encoding including MHE and separate heads are declared in Section 4.2; Topo-mapping pipeline and matching method is explained in detail in Section 4.3; formulation in the figure 2, 3 has been clarified; space is added before all ().

太宽泛

For other questions:

Bounding-box in text query localization: Actually, it's not the general bounding-box from object detection, we filter the points with similarity over threshold (0.6 in our practice), and simply draw a bounding box to cover these points for visualization. 'Samples' is the number of text queries, which is clarified in revised pdf. Ground truth comes from the object instance labels from Matterport3D. For table 3, we've mentioned in Section 5.2 that more than 40 images are sampled from each scene, ground truth comes from back-projecting image pixels into 3D according to ground truth pose and depth.

Image query localization: Table 3 shows weighted average distance among all samples in a scene, using similarity as weight. Given that few points in orange may appear in other rooms as noise, the max distance of a single point from these points (similarity would be 0.3~0.6) would be less than 6~8 m, which counts relative little.

指出在文章里已经包含的细节

- 如果评审人提到的内容已经在论文中给出，务必要指出这一点。
- 提供**具体的行号**、**表格**或**图表编号**
- 目的是向审稿人和AC 证明你的论文并未缺少重要细节

> **Q8: Both MVSplat and pixelSplat report higher PSNR numbers for RealEstate10k and ACID datasets. Since the same train-test split was used, why is there a huge discrepancy in the 'Average' column of the metrics?**

1. While we use the same train-test split as MVSplat and pixelSplat, the discrepancy in the 'Average' column arises from differences in the image pairs used for evaluation. Specifically, the image pairs in their evaluation have large overlaps, which simplifies the task and makes it difficult to distinguish the true capabilities of each method in novel view synthesis NVS.
2. To better assess each method's ability to handle varying degrees of camera overlap, we generate evaluation input pairs categorized by their overlap ratios: small (5%–30%), medium (30%–55%), and large (55%–80%). This categorization ensures a more rigorous evaluation, as described in L.319–L.321 of the paper.
3. The "Average" column in our results corresponds to the average performance across these three overlap settings. This explains why the performance appears lower compared to evaluations focused on large-overlap pairs.
4. To address any potential concerns, we have also reported results using the original evaluation set from pixelSplat and MVSplat in Tab. 8. Even under their evaluation settings, our method outperforms both pixelSplat and MVSplat, which require poses as input.

为评审人使用颜色编码（一页PDF rebuttal）

这样可以让评审人更容易找到与自己相关的回复，即使多个**回复合并** 在一起

We thank all reviewers for their insightful comments. R1 says the paper has *"a novel and interesting idea"* and *"very impressive results"*. R2 says *"the contribution is interesting"* and *"has practical benefits"*. R3 *"really likes the idea"* and and suggests it provides *"guidelines for future work"*. We will incorporate all feedback and suggested references. The following addresses other comments.

Design choices: a) why 3 scales b) why middle layer is not a residual to the coarse (R1, R3): For a), we show in Fig. 9 that using hierarchical grids leads to better convergence compared to a single level, and we find that the current design guarantees a good balance between the qual-

More comparisons (R1, R2): In Fig. 10, we compare to KinectFusion, i.e., TSDF-Fusion, using our camera poses for fair comparison. Tab. 6 shows that NICE-SLAM produces high-quality geometry with a low memory foot-

Paper contributions (R1, R2). We emphasize that our work pioneers a highly simple, yet versatile and robust module, designed for easy integration into any NeRF pipeline, as recognized by R1 & R3. We significantly enhance NeRF's applicability to casually captured data in various scenarios. Within this, we have three key innovations: 1) Use DINO features for accurate uncertainty prediction; 2) Replace L2 with SSIM-based loss for enhanced uncertainty learning; 3) Dilated patch sampling for fast and effective distractor removal. Furthermore, R3's acknowledgment of our On-the-go dataset, highlighting its potential to *accelerate new avenues for NeRF* further underscores the innovative nature of

数据胜于雄辩

- 与其与审稿人争论，不如用数据或结果来支持你的观点（已有结果或新实验）
- 用数据解决问题后，再提供直观的解释或补充论点，这样的回应更具说服力

More comparisons (R1, R2): In Fig. 10, we compare to KinectFusion, i.e., TSDF-Fusion, using our camera poses for fair comparison. Tab. 6 shows that NICE-SLAM produces high-quality geometry with a low memory footprint. DI-Fusion has poor camera tracking in 3 scenes, and even after removing them (Acc.=2.30, Comp.=6.24, Comp. Ratio=77.53), ours still outperforms overall.

数据胜于雄辩

- 与其与审稿人争论，不如用数据或结果来支持你的观点（已有结果或新实验）
- 用数据解决问题后，再提供直观的解释或补充论点，这样的回应更具说服力

审稿人的要求
(额外4个实验)

3. As there are not a lot directly comparable works on scene-level, I require the authors to include additional five experiment results.

- On object-level, train the proposed model using the same data with the previously mentioned pose-free works and compare their performance. This also verifies whether MAST3R weight initialization generalize to object-level data.
- On scene-level, train with only RealEstate10K+ACID for comparing with MVSplat.
- On scene-level, train a variant of the proposed model, which is also conditioned on the camera poses of inputs (using the plucker ray representation rather than 6D pose representation). If the performance gap is small enough, the model has a strong capability of correlating the two images with the missing poses.
- On scene-level, ablate the weight initialization by training with no weight initialization (from MAST3R/DUST3R/Croco) using the current training set (ACID+RealEstate10K+DL3DV). This experiment ablates whether the pose-free inference capability comes from initialization or your model learning process. Please include results for both pose-free and pose-conditioned variants of your model.

我们的回答

1. On object-level, train the proposed model usi
MAST3R weight initialization generalizes to obj

Thank you for your suggestion. We agree that such a

The reviewer mentioned two pose-free object-level re
infeasible. Therefore, we conducted a comparison wi

To ensure a fair comparison, we trained and evaluate
results, shown in the table below, demonstrate that
capabilities of our method for object-level reconstruc

	PSNR	SSIM	LPIPS
LEAP	20.559	0.853	0.144
Ours	28.378	0.935	0.053

3. On scene-level, train a variant of the proposed mod
(using the plucker ray representation rather than 6D
the model has a strong capability of correlating the t

Thank you for your suggestion. Following your recommend
ray representation and concatenating it with the RGB imag
pose conditioning slightly improves performance compare
indicates our method's robust capability in correlating imag

Init Weight	pose condition	PSNR	SSIM	LPIPS
MASt3R	Yes	25.080	0.844	0.158
	No	25.033	0.838	0.160
Random	Yes	23.708	0.788	0.173
	No	23.487	0.779	0.189

4. On scene-level, ablate the weight in
MAST3R/DUST3R/Croco) using the curr
the pose-free inference capability com
both pose-free and pose-conditioned

Thank you for the insightful suggestion.

Init Weights	PSNR	SSIM	LPIPS
MASt3R	25.033	0.838	0.160
CroCo-v2	24.559	0.818	0.171
DINOv2	24.094	0.812	0.176
Random	23.487	0.779	0.189

这个审稿人从 3 (reject)
提分到 8 (accept)

不要只承诺，直接行动

- 与其说“我们将在论文中讨论XX工作” ，不如在回复中直接进行讨论
- 然后补充说明你会在论文中添加这些内容

@R1, dialog-level evaluation: Thanks for the suggestion! Using Recall@5 to define round-level ‘success’, our best discriminative model MN-QIH-D gets 7.01 rounds out of 10 correct, while generative MN-QIH-G gets 5.37. Further, the mean first-failure-round (under $R@5$) for MN-QIH-D is 3.23, and 2.39 for MN-QIH-G. Fig. 1c and Fig. 1d show plots for all values of k in $R@k$. We will add this analysis.

保持开放和理性

- 与其争论评审人的建议为什么可能行不通，想想是不是确实可行，然后尝试一下，或许真的可行！
- **给你的审稿人credit**! 如果真的有帮助，好好感谢你的审稿人给出的建议

Use F-Score

It is a great suggestion, thanks! Consistent with the other metrics, our method outperforms SPSR (which requires normals as input) also in terms of F-Score. We will add this evaluation to the paper:

Synthetic Room	SPSR	SPSR (trimmed)	Ours-2D ($128^2 \times 3$)	Ours-3D (32^3)	Ours-3D (64^3)
F-Score	0.810	0.892	0.948	0.941	**0.964**

ScanNet	SPSR	SPSR (trimmed)	Ours-3D (64^3)
F-Score	0.731	0.847	**0.886**

Matterport3D run on each room individually

Thanks for this comment which inspired us to also implement a fully convolutional version of our model that scales to any size by running on overlapping crops of the point cloud in a sliding window fashion. The overlap is determined by the size of the receptive field to ensure correctness of the results. We will update the paper and release code for both variants.

> **Q2: The main advantage seems to come from finetuning the Mast3r backbone (improve overall model performance. But this orthogonal to the motivation of t**

Thank you for your insightful comment.

1. **We agree with the reviewer** that using photometric loss alone to fine-tune a ViT methods like MASt3R and DUSt3R. This extension offers several notable benefits:
   - High-quality novel view synthesis
   - More accurate pose estimation
   - Broader dataset applicability (we can train purely on posed RGB image sequer
2. At the same time, we would like to note that our motivation in the paper is still val limitations of previous approaches, such as the dependency on ground-truth pose resolve these constraints.

Convolutional Occupancy Networks
(两个审稿人都从borderline 提到 weak accept)

@R1, models with attention over image: We trained a variant of our best-performing MN-QIH with attention over image, and got MRR 0.531, $R@10$ 80.97 for discriminative and MRR 0.443, $R@10$ 59.91 for generative, which outperforms our earlier approaches (by $\sim$1.4% $R@10$)! We thank R1 for this and will definitely include these results!

保持开放和理性

- 但如果审稿人真的完全理解错你的文章了，先别急着生气然后怼回去
- 理性地强调TA理解错的点的事实，还可以再用其他新的内容去进一步支撑

一个例子：DynIBaR: Neural Dynamic Image-Based Rendering **(CVPR'23 Best Paper Honorable Mention)**

一个人误认为文章只能做video stabilization，但还不比: The paper does not mention video stabilisation techniques, as if this field would not exist… No comparisons to video stabilisation techniques…

回答思路:
1. 强调文章目标是做新视角生成，不是video stabilization
2. 这其实让我们方法可以做很多新的应用，比如控制时间和视角，dolly zoom, video bokeh，等等
3. 这些video stabilization的方法都做不到
4. 但即便如此，我们也新做了和video stabliization方法的比较，比他们还要好很多

Reject 提分到 Weak Accept, 最终拿了最佳论文荣誉提名

揭示评审人不公平的行为（极少数情况下使用）

- 有些时候评审人没有认真对待评审，有很强的偏见，这时需要让其他评审人和AC意识到这一点，并适当降低该评审意见的权重
- 此外（如果适用），你还可以通过对AC的保密评论进一步说明问题

@R1 – *"8 percentage points improvement with HINT does not count too much"*: We respectfully disagree (and to be honest, suspect that most other researchers would disagree). Visual question answering under changing priors (VQA-CP) is challenging and prior published work on the task has pushed performance from 39.74% to 41.17% (Adv-Reg [23]). In contrast, we increase performance by 8%, significantly improving over existing work which both R2 and R4 recognize as significant. **If R1 had provided any explanation for the opinion that 8% is insignificant we could have perhaps addressed those concerns.**

@R1 – *"It is almost common sense that as the % of the dataset with HINT supervision gets higher, the performance will definitely get higher"*: We disagree. It is not obvious that access to additional, even if relevant, information during training will necessarily improve performance **when generalizing to test instances without this information** – how the additional information is used is paramount. In fact, our experiments show that directly supervising attention masks with human attention fails to yield improvements (UpDn+Attn. Align in Tab. 1).

揭示评审人不公平的行为（极少数情况下使用！）

一个例子：Cavali et al.: **Handcrafted outlier detection revisited**. ECCV 2020

R2（初步评分 strong reject)：重点批评在于文章组织非常混乱，未能总结出论文的贡献和优点

作者回复思路：a) 其他审稿人并没有这样的顾虑 b) 列出了文章贡献的所有讨论段落

R2（最终评分 strong reject)：这些都属于文章的组织问题，我认为小修补无法解决这些问题。

作者给AC的信：R2 声称"由于文章组织不佳"，因此未能总结出论文的贡献和优点。我们对此表示十分困惑，尤其是R2并未提供任何理由来支持这一说法，同时R3却指出"文章写得很好，易于阅读"。R2表示"尽管作者在引言中总结了贡献，但我未能找到"，这显得非常奇怪，因为一些贡献很容易找到。例如，第83-89行的内容可以在表1和表2中直接看到。同时，其他评审者似乎并没有遇到理解贡献和优点的问题。总体而言，我们对R2评审的质量及其理解我们论文的努力表示担忧。

AC的决定：尽管R2建议强烈拒稿，但两位AC认为R2的论点缺乏依据。本文确实重新审视了手工设计的离群点过滤，证明了其仍然具有重要价值。R3也提到论文写得很好，AC对此表示同意。AC进一步认同作者……因此，R2的评审意见被完全弃用。
……. 尽管评审意见大多为负面，AC认为论文中有足够有趣的结果和想法，不能简单地拒稿，因此建议接受该论文

1 strong reject, 1 weak reject, 1 borderline reject -> accept as poster

感谢评审人的努力

如果某位评审人非常认真并提供了建设性的意见，一定要对他们表示感谢。例如：

- 如果评审人真的提到一个实际的问题，感谢，承认问题，并说明如何改正

- 提供了错别字清单 (typo list)？感谢！
- 指出相关的研究工作？感谢！
- 对未来研究方向提供了详细的思考？感谢！

一些额外的建议

- **预想可能的实验和问题**：在rebuttal之前就先把可能被问的点先做实验，以免一周不够
- **顺序很重要**：从**最重要的**且**你有充分答案**的问题开始。 逐步过渡到较难明确回应的问题和次要点。这有助于在一开始就建立说服力，并让读者对你的回复留下积极印象
- **保持对话式的友善的语气**：注意前面所有例子，对话式能够展现你的合作态度，并提升回复的说服力
- 不要害怕在回复中**加粗**或者**大写**来强调重点，We do NOT need blabla.
- **保持透明**: 如果询问某个趋势的直觉解释，而你目前没有好的答案, 坦诚地说你已经考虑过，但暂时没有好的想法，并表示会继续研究。如果他们要求的实验你因为 GPU 不足而无法完成？如实说明
- **保持一致**: 所有评审者都可以看到你的回复，因此你的回应必须一致且没有矛盾

Q&A：回答同学们的一些问题

问题 1：如何在rebuttal之前预想reviewer会提问的问题并在supp里面补充对应的实验

回答：

- 自己最清楚文章可能被诟病的点，努力先把那些实验做了
- 可以和同学老师们讨论，他们觉得什么是欠缺的

Q&A：回答同学们的一些问题

问题 2：如果reviewer提的实验量很大，在rebuttal期间无法跑完，应该怎么做？

回答：

- 一方面，认准最重要的实验，比如两三个reviewer都说你要加一个实验，那肯定要做
- 另外，如果确实是跑不完，可以诚实的说，时间和资源有限，我们没法做所有的实验，但是我们还是做了1 2 3等实验
- 看看合作者能不能帮忙

Q&A：回答同学们的一些问题

问题 3：Rebuttal期间让reviewer涨分最关键的因素是什么呢

回答：

- 要找出这个reviewer真正在乎的1到3个诟病点是什么，缺乏实验？没有创新点？写作不好？然后对症下药
- **清晰**以及**令人信服**的**解决**了**所有或者绝大多数**这个**reviewer**的问题

Q&A：回答同学们的一些问题

问题 4：以7天为标准，如何规划rebuttal的宝贵时间呢？每个阶段应该注意哪些重点呢？

回答：分情况，但是大概可以是以下：

- **第一天**：把所有reviewer的所有问题都过一遍，列出来，然后写一些初步的回复想法，以及确认需要添加的实验。**这一定要尽早开始**
  - **列出所有关键问题（影响论文接受决策的问题），进一步思考**
- **第二天到第五天**：做实验，和合作者讨论确定回答思路，然后已经写了初稿
- **第六七天**：可以不断修改优化回答，让更资深的人过一遍

Q&A：回答同学们的一些问题

问题 5：怎么判断是否要rebuttal，或者放弃re直接撤稿？

回答：

- 除非真的是文章有硬伤，觉得回复都没有机会可以改变审稿人想法的
- 或者这个会议可能不是很适合，转投另外一个更适合 (比如你的其实很偏ML，投Siggraph得到很差的意见，那么可以转投ICLR, NeurIPS)

我还是建议，尽可能还是要努力做一下rebuttal，不要放过任何一个机会

Q&A：回答同学们的一些问题

问题 6：如果审稿人不负责任回答和评分，比如使用AI审稿，应该怎么应对

回答：

- 还是要平静下来，分析到底有没有道理，没有道理也要心平气和的好好rebuttal
- 如果审稿人还是很不负责的评分，不理你的，考虑给AC写信

Q&A：回答同学们的一些问题

问题 7：如果给AC写信的话，应该从什么角度出发呢？如果审稿人没有明显问题+这篇工作确实是一篇平庸的工作。

回答：AC写信只能在非常极端的例子使用！！如果你的文章本身平庸，那确实可能没有机会，想想怎么提升文章本身吧！

谢谢！

Sida Peng

Jun Gao

Songyou Peng

Qianqian Wang