Reply to the Reviewers

Re: Manuscript ID Sensors-85869-2025
"DSOSplat: Monocular 3D Gaussian SLAM with Direct Tracking"
Yi Zhou, Zhetao Guo, Dong Li, Runwei Guan, Yuxiang Ren, Hongyu Wang and Mingrui Li
IEEE Sensors Journal

Overview

On behalf of my co-authors, we thank you very much for allowing us to revise our manuscript "DSOSplat: Monocular 3D Gaussian SLAM with Direct Tracking" (Manuscript ID: sensors-85869-2025). We appreciate the editor and reviewers very much for acknowledging the innovative and technical quality of our work and for the positive and constructive feedback. In the revised version, we have addressed the concerns of the reviewers. The revision was marked in red fronts in the manuscript.

1. **Clarification of System Application Scenarios and Positioning** We clarified that the proposed monocular dense SLAM system is particularly advantageous for indoor applications such as AR/VR and human-computer interaction. Meanwhile, the system also demonstrates good generalization capabilities in outdoor scenarios. We further elaborated on its strengths in robustness, scale awareness, reconstruction accuracy, and computational efficiency. Related content has been added to Section **IV.A**.

2. **Justification for Baseline Method Selection** We provided the rationale behind selecting ESLAM, RTG-SLAM, and Loopy-SLAM as baseline methods. These approaches represent state-of-the-art techniques in neural radiance fields and 3D Gaussian Splatting-based SLAM. RTG-SLAM and Loopy-SLAM, being RGB-D systems with loop closure, emphasize the competitiveness and generalization of our method. Explanations were added to Section **IV.A**.

3. **Motivation and Necessity of the SC-AMVS Module** We thoroughly explained the motivation behind introducing the SC-AMVS module, which integrates Depth Anything V2's scale prior with multi-view geometric consistency to improve the continuity and precision of depth estimation. We also added ablation studies comparing SC-AMVS with DAV2-only and AMVS-only configurations to validate its effectiveness. Relevant discussion appears in Section **III.A**.

4. **Clarification of Key Technical Details and Reference Additions** We refined the technical descriptions of feature volume construction, the 3D convolutional network architecture, and the role of the learned weighting matrix. Moreover, we added the previously missing citations to strengthen the completeness and clarity of the methodology. These updates were applied in Sections **III.A**, **III.B**, and **IV.E**.

5. **Standardized Terminology and Model Definitions** We formally defined core concepts such as "Gaussian points," "uniform/multiple motion models," and "initial keyframe selection." These clarifications resolve potential ambiguities in the original manuscript and were added to Sections **III.C**, **III.D**, and related parts.

6. **Discussion of Limitations and Future Extensions** We expanded the conclusion with a detailed discussion of the system's limitations and potential extensions, including support for multi-robot collaborative mapping and multi-modal sensor fusion. This enhances the openness and forward-looking perspective of the manuscript. The discussion was added to Section **IV.H**.

Review 1

Q1. *For benchmarking, the authors use Replica and ScanNet datasets, which are well-known as indoor scenarios. Further, a self-collected outdoor data is also collected. The question is, for which applications is the proposed method better suited? Why? If the indoor application is perferred, the authors should explain which specific features of the proposed system make this kind of application more suited.*

We sincerely thank the reviewer for the valuable comments. Our proposed method primarily targets dense mapping and high-precision localization tasks based on monocular sensors, and demonstrates certain advantages across various scenarios (both indoor and outdoor). Specifically, compared to conventional monocular SLAM systems, our method exhibits the following four key features:

1. **Robust tracking module** We adopt DSO as the pose estimation module in our system. Compared to feature-based methods, DSO exhibits higher stability under conditions of low texture, illumination changes, and slight occlusions, which facilitates practical deployment.

2. **End-to-end dense mapping pipeline** By integrating our proposed SC-AMVS module with 3DGS, the system achieves a full pipeline optimization from image input to dense model output, making it well-suited for tasks requiring high-quality geometric reconstruction.

3. **Ease of multi-sensor fusion** Since DSO is a direct method that models image intensity values, it naturally offers continuity and differentiability. This property makes it easier to integrate with other sensors (e.g., IMU, depth cameras) under monocular settings for joint optimization.

4. **Low computational requirements** Compared to NeRF-based approaches, our method employs Gaussian-based representation and rendering, resulting in higher runtime efficiency. This makes it more suitable for deployment on resource-constrained platforms such as mobile devices or robots.

Our system is particularly well-suited for applications such as AR, VR, and human-computer interaction in indoor environments. The proposed SC-AMVS module, combined with DAV2 for scale calibration, enables the generation of dense depth maps with real-world scale under purely monocular input, significantly enhancing the precision and fidelity of indoor reconstruction. Leveraging the detailed modeling capability of 3DGS, our system outperforms traditional sparse or mesh-based representations in both texture continuity and geometric accuracy, making it ideal for high-frequency interactive tasks in indoor scenarios. Furthermore, the system operates without the need for additional post-processing or lengthy optimization, providing strong real-time performance to meet the latency-sensitive requirements of AR/VR applications.

Although our method is primarily designed for indoor scenarios, it also demonstrates a certain level of adaptability to outdoor environments, especially when compared with traditional sparse SLAM systems. It is capable of producing dense reconstructions with structural continuity and high semantic fidelity, offering a stronger foundation for tasks such as path planning and object recognition. Additionally, our multi-view dense estimation preserves geometric consistency even without auxiliary sensor inputs, thereby enhancing its practicality in natural outdoor environments.

In summary, we believe that our system offers great advantages in indoor AR, VR, and human-computer interaction applications, while also exhibiting good extensibility to complex outdoor environments or collaborative operation with multimodal sensor systems. We have incorporated the relevant discussions into Section **IV.H** of the manuscript. We thank the reviewer for their insightful feedback and attention to our work.

Q2. *The authors should explain why they choose ESLAM, RTG-SLAM and Loppy-SLAM as Benchmarks. Are there any specific reasons?*

We thank the reviewer for raising this constructive question. The lack of sufficient explanation regarding the selection of baseline methods may indeed lead to confusion among readers. In the revised manuscript, we have provided a more detailed clarification:

ESLAM, RTG-SLAM, and Loopy-SLAM represent the state-of-the-art methods corresponding to radiance field-based and 3D Gaussian Splatting (3DGS)-based SLAM approaches. Notably, both RTG-SLAM and Loopy-SLAM are RGB-D systems equipped with loop closure capabilities, which further highlights the competitiveness and generalizability of our monocular-based method.

We have incorporated and revised the relevant discussion in Section **IV.A** of the manuscript. We thank the reviewer for helping us improve the rigor and clarity of the manuscript.

Review 2

Q1. *Depth Calibration Ablation Study: The manuscript proposes an adaptive multi-view depth estimation module (SC-AMVS) that leverages depth calibration based on depth maps provided by DAV2. However, the motivation for estimating depth using SC-AMVS is unclear, given that DAV2 appears to provide accurate absolute depth information directly as stated "Leveraging the real depth information provided by Depth Anything V2". Please discuss it clearly. An additional ablation study comparing the proposed SC-AMVS method against a baseline directly using the DAV2 depth maps would significantly clarify the necessity and benefit of employing SC-AMVS. Such a study would elucidate whether SC-AMVS contributes meaningful improvements beyond the raw DAV2 depth data.*

We thank the reviewer for the thorough review of our work and for pointing out this important issue. We understand the reviewer's concern regarding the motivation for introducing the SC-AMVS module, particularly given that we already use Depth Anything V2 (DAV2) to provide scale-aware depth as a reference. We have addressed this concern explicitly in the revised manuscript.

Although DAV2 is capable of generating depth estimates with absolute scale from monocular images, it remains fundamentally a single-image-based depth prediction network. As such, it exhibits several limitations when applied to continuous frames or multi-view scenarios. Since DAV2 performs inference on individual images independently, it ignores inter-frame geometric relationships, which often leads to inconsistent depth predictions in scenes with complex geometry or occlusions. Despite producing depth maps with absolute scale, the predictions may still suffer from scale bias when encountering scene variations (e.g., indoor vs. outdoor environments, lighting changes), due to its reliance on image-level priors. In particular, DAV2 can produce local depth discontinuities across adjacent frames, which adversely affects the stability of downstream fusion and tracking processes.

To address these limitations, we introduce the SC-AMVS module. The core motivation lies in integrating multi-view geometric consistency with DAV2's absolute scale information to achieve high-accuracy, real-scale dense depth estimation. By constructing an adaptive cost volume and incorporating multi-scale feature aggregation, our approach provides superior spatial consistency compared to single-frame predictors. DAV2 is employed as an external source of scale supervision, forming a depth-scale calibration term that effectively compensates for the scale ambiguity commonly encountered in traditional multi-view depth estimation. Notably, the module does not require ground-truth depth for training and can be trained on synthetic data while still generalizing well to complex real-world scenarios, demonstrating its practicality and effectiveness as an intermediate depth fusion module.

Therefore, we argue that SC-AMVS does not merely replicate DAV2's functionality. Instead, through a complementary design, it achieves an organic integration of multi-view geometric constraints and single-image depth priors, yielding significant improvements in both tracking robustness and mapping accuracy. We have added a detailed discussion of the motivation and ablation experiments (including comparisons using DAV2 only, AMVS only, and the combined approach) in Section **III.A** of the manuscript to support this viewpoint. We thank the reviewer for raising this critical question, which has helped us further refine the theoretical coherence and methodological logic of our work.

Q2. *Citations and References: Several important algorithmic steps lack proper citations. Please provide clear citations or references for key technical approaches, especially in Section **III.A** and Section **III.B**, to ground your work clearly within existing research.*

Thank you for the reviewer's suggestions. The reason the Section **III.A** (SC-AMVS) and Section **III.B** (Visual Odometry) lacked proper citations is that we had previously referenced related papers but did not cite them in those sections. Here, we have rechecked Section **III.A** and Section **III.B**, added the relevant citations and highlighted them in red in the text. The references are listed below.[1,2]

Q3. *Feature Volume V_{σ_k}: Clarify explicitly how the feature volume is obtained and what specific transformations or operations are involved.*

We thank the reviewer for their attention to our work and for the valuable suggestions. In accordance with the comments, we have provided a more detailed and explicit description of the generation process of V_k^σ in the revised manuscript.

The generation of V_k^σ involves multi-scale feature extraction and depth-hypothesis-based geometric transformation. Specifically, each keyframe image I_k is processed by a shared-weight convolutional network to extract multi-scale feature maps $F_k^\sigma \in \mathbb{R}^{C^\sigma \times H^\sigma \times W^\sigma}$, where σ denotes the scale level. At each scale σ, we construct a discrete set of depth hypotheses $D_{\text{hyp}}^\sigma \in \mathbb{R}^{L^\sigma}$, where L^σ is the number of hypothetical depth planes. This set is designed to converge progressively across scales.

For each pixel (h, w) in the reference image, we back-project it into 3D space under each depth hypothesis d_l, and then project the resulting 3D point onto the target image I_k using the known relative camera pose T_k. By performing bilinear interpolation at the projected locations on the feature map F_k^σ of the target image, we obtain a feature volume $V_k^\sigma \in \mathbb{R}^{L^\sigma \times H^\sigma \times W^\sigma \times C^\sigma}$ that encodes feature information from different depth levels under the given view. This feature volume serves as the foundational input to the subsequent adaptive cost aggregation module, thereby enhancing the accuracy and robustness of depth consistency evaluation across multiple views.

We have incorporated and revised the relevant explanation in Section **III.A** of the manuscript. We thank the reviewer for helping us improve the clarity of technical details and the overall rigor of our presentation.

Q4. *3D Convolutional Network: The description "using a shallow 3D convolutional network to generate a weighting matrix " is overly vague. Please clearly define this network's architecture, parameters, inputs, and outputs.*

We sincerely thank the reviewer for the valuable comments. We acknowledge that the original description of the "shallow 3D convolutional network" was not sufficiently specific, which may have hindered the reader's understanding of the working mechanism of this module. We have now added a detailed definition of the network structure, parameter configuration, and input/output specifications in the revised manuscript, as outlined below:

In our proposed SC-AMVS module, a shallow 3D convolutional network is introduced to enable adaptive weighting across different viewpoints when constructing the cost volume. This network is used to generate the weight matrix w_k^σ, which reflects the reliability of each view under different depth hypotheses. The input to the network is the feature volume $V_k^\sigma \in \mathbb{R}^{L^\sigma \times H^\sigma \times W^\sigma \times C^\sigma}$ of keyframe I_k at scale σ, representing a collection of feature maps under various depth hypotheses. We transpose this tensor to $\mathbb{R}^{C^\sigma \times L^\sigma \times H^\sigma \times W^\sigma}$ to conform with the standard input format for 3D convolutions. The output is a weight matrix $w_k^\sigma \in \mathbb{R}^{L^\sigma \times H^\sigma \times W^\sigma}$, which is used to perform per-voxel adaptive weighting of contributions from different viewpoints in the cost

volume.

We have explicitly included this architectural description and added corresponding annotations in Section **III.A** of the manuscript. In future work, we also plan to explore the potential of deeper networks to further enhance the robustness of adaptive weighting.

We thank the reviewer for the detailed review and constructive feedback that have helped us improve the clarity and completeness of our work.

Q5. *Learning Weight Matrix: There is potential ambiguity regarding whether the learned weighting matrix effectively differs from uniform weighting (i.e.,). Include a brief discussion or experimental evidence verifying that the learned weights meaningfully deviate from uniform weighting.*

We thank the reviewer for raising this critical question. We acknowledge that the original manuscript did not explicitly clarify whether the learned weight matrix w_k^σ significantly deviates from uniform weighting, which may lead to ambiguity in understanding.

To address this issue, we have included a brief verification in the revised version to demonstrate that the learned weights indeed exhibit adaptive variability in practice and are effectively distinguished from simple uniform averaging. Specifically, we conducted an ablation study comparing two settings: one using the learned weights and the other using fixed uniform weights, in terms of depth estimation accuracy. The experimental results show that the learning-based weighting scheme consistently achieves lower errors across multiple public datasets, confirming the practical advantage of adaptive weighting.

Based on this analysis, we conclude that the learned weight matrix not only numerically deviates from a uniform distribution but also plays a critical role in improving the overall performance of the system. The relevant supplementary discussion has been added to Section **IV.E** of the manuscript. We sincerely thank the reviewer for the thorough review and constructive guidance.

Q6. *Definition and Computation of Feature Volume: Clearly describe how the feature volume is derived from extracted image features. Explicitly detail the steps involved in generating these volumes, as it is critical for understanding the adaptive aggregation process.*

We thank the reviewer for pointing out this important issue and fully agree that it is a key component for understanding our proposed adaptive cost aggregation process. To address this, we have clarified and elaborated on the definition and construction procedure of V_k^σ in the revised manuscript.

In our method, V_k^σ refers to a four-dimensional voxel structure constructed at scale σ by spatially aligning multi-view image features according to a set of predefined depth hypotheses.

All keyframe images I_k are first processed by a shared-weight feature extraction network to obtain scale-specific feature maps $F_k^\sigma \in \mathbb{R}^{C^\sigma \times H^\sigma \times W^\sigma}$. At each scale σ, we define a discrete set of depth hypotheses $D_{\text{hyp}}^\sigma \in \mathbb{R}^{L^\sigma}$ that spans the depth range from near to far, which serves as the basis for the voxel dimension. For each pixel location (h, w) in the reference frame, we back-project it into 3D space using each depth hypothesis $d_l \in D_{\text{hyp}}^\sigma$; then, using the known relative pose T_k, we project the 3D point into the target frame I_k to compute the corresponding 2D coordinates. By performing bilinear interpolation at these locations on the feature map F_k^σ of the target frame, we obtain the corresponding feature values.

For each keyframe I_k, this alignment and sampling process is repeated over all depth hypotheses, resulting in a feature volume $V_k^\sigma \in \mathbb{R}^{L^\sigma \times H^\sigma \times W^\sigma \times C^\sigma}$, where L^σ is the number of depth planes, H^σ and W^σ are the spatial dimensions, and C^σ denotes the number of feature channels. This feature volume is then passed into the cost volume construction module, where it is combined with the learned view-dependent weight matrix to perform adaptive cost aggregation. This approach effectively improves the stability and accuracy of depth estimation in multi-view scenarios.

We have added the detailed explanation of these steps to Section **III.A** of the manuscript. We thank the reviewer for helping us identify and refine the expression of this critical component.

Q7. *Initialization and Motion Models: The terms "unity speed motion model" and "multiple motion models" are unclear. Explicitly clarify the concrete models used, and provide rationale or citations supporting their selection and effectiveness.*

We thank the reviewer for raising this constructive question. We acknowledge that the original manuscript presented certain conceptual ambiguities regarding the terms "unity speed motion model" and "multiple motion models." To improve the clarity and accuracy of our paper, we have thoroughly revised and explicitly clarified this section in the updated manuscript:

For the **Unity Speed Motion Model**, we adopt a constant velocity assumption during the initial pose estimation phase. Specifically, we infer the initial pose of the current frame by extrapolating from the motion observed between the previous two frames. This strategy is equivalent to assuming that the current frame follows the same velocity and direction as the preceding one along the camera trajectory.

For the **Multiple Motion Models**, we consider the presence of challenging conditions in real-world scenes, such as abrupt motion changes, insufficient inter-frame parallax, or inconsistent lighting. To improve the robustness of initial pose estimation under such conditions, we introduce several alternative motion hypotheses, including a double-speed model, half-speed model, perturbed rotation model, and static model. During runtime, we generate a set of candidate poses based on these motion hypotheses and evaluate each of them using forward photometric error and convergence criteria during optimization. The motion model that leads to the most reliable convergence is then selected as the initialization for the current frame. This multi-model strategy significantly reduces the risk of tracking failure, especially in dynamic, blurry, low-texture, or large-motion environments.

We have incorporated these clarifications and revisions into Section **III.C** of the manuscript. Additionally, we present a comparative analysis in the experimental section to demonstrate the improvement in tracking stability brought by the introduction of the multi-model initialization strategy. We sincerely thank the reviewer for the insightful feedback and careful attention to the details of our work, which helped enhance the overall clarity and academic rigor of the paper.

Q8. *Definition of Gaussian Points: The manuscript frequently refers to "Gaussian points" without clear definition. Please provide a concise yet rigorous definition of "Gaussian points," including their properties, roles, and relevance within your framework.*

We thank the reviewer for raising this important point. We acknowledge that although the term "Gaussian points" is frequently used throughout the manuscript, the original version lacked a clear and formal definition, which may hinder readers' understanding. To address this, we have added the following concise and rigorous explanation in the revised manuscript:

Using the 3D Gaussian point cloud representation, we achieve continuous and smooth dense modeling. A set of 3D Gaussian primitives $\{P_k\}_{k=1}^{N}$ is defined in the scene, with each primitive P_k parameterized by a mean position $\mu_k \in \mathbb{R}^3$, a covariance matrix $\Sigma_k \in \mathbb{R}^{3\times3}$, an opacity value $\alpha_k \in [0,1]$, and a color vector $c_k \in \mathbb{R}^3$. The density function of a single Gaussian is given by:

$$P_k(\sigma) = \exp\left(-\frac{1}{2}(\sigma-\mu_k)^\top \Sigma_k^{-1}(\sigma-\mu_k)\right)$$

The spatial covariance matrix Σ_k defines the shape and orientation of the ellipsoidal support of the Gaussian, allowing for anisotropic spatial distribution. It is factorized as:

$$\Sigma_k = R_k S_k S_k^\top R_k^\top$$

where $R_k \in \mathbb{R}^{3\times3}$ denotes the orientation matrix and $S_k = \mathrm{diag}(s_k) \in \mathbb{R}^{3\times3}$ defines the scale along each principal axis.

These Gaussian primitives serve as the core building blocks of the scene representation in our SLAM framework. They support differentiable rendering, enable continuous geometry reconstruction, and facilitate efficient gradient-based optimization.

We have incorporated the above definitions along with illustrative figures into Section **III.D** of the manuscript to help readers better understand the modeling approach, parameter structure, and functional role of Gaussian points within the overall SLAM system. We thank the reviewer for this critical comment, which helped us further improve the clarity and logical rigor of our paper.

Q9. *Initial Keyframe Selection: Explain explicitly how the initial keyframe is chosen, including specific criteria or initialization methods utilized.*

We thank the reviewer for pointing out this critical issue. We acknowledge that the original manuscript did not provide a sufficiently detailed explanation of the initial keyframe selection process, which may have led to ambiguity. In the revised manuscript, we have added the following clarification regarding our keyframe initialization strategy:

At the system initialization stage, a keyframe candidate is evaluated at fixed intervals, and one frame is selected as the initial keyframe. This strategy is inspired by the fixed time window approach used in the DSO system, which helps avoid robustness issues arising from insufficient visual parallax in the early stages due to a lack of reliable features. Unlike traditional feature-based SLAM systems, DSO does not rely on discrete keypoint detection and selection. Instead, it jointly optimizes camera poses using a set of sparse points observed across multiple frames in a sliding window.

During initialization, we synthesize a representative image frame—selected based on favorable lighting and structural characteristics—to serve as the first keyframe input, along with the corresponding initialization of sparse point maps and camera pose. Although the sparse DSO back-end offers high computational efficiency, it is susceptible to noise in low-texture or early initialization stages. To improve the robustness of the initialization process, we incorporate dense depth maps to enhance image alignment in the front-end. For each pixel in the current frame, if a sparse DSO depth estimate is available, it is used; otherwise, the pixel is filled with a rendered depth value. This strategy enables the construction of an approximately dense depth map, which is then used for direct image alignment between consecutive frames.

We have included the above details in Section **III.C** of the manuscript to improve the completeness and clarity of our method description. We sincerely thank the reviewer for this valuable suggestion, which has helped us refine the presentation of key methodological components in our system.

Q10. *Discussion on limitation and generalizability is necessary for research work.*

We sincerely thank the reviewer for pointing out this issue. We fully agree that a systematic discussion of future research directions is essential for any scholarly work. In the revised manuscript, we have added the following content to the conclusion section to elaborate on the applicability and potential extensions of our proposed method:

Despite the aforementioned computational challenges, we believe DSOSplat still demonstrates promising scalability and cross-scenario adaptability, particularly in two critical XR directions: multi-terminal immersive interaction and multimodal perception fusion. Regarding the former, the Gaussian distribution-based scene structure constructed by DSOSplat inherently possesses distributional continuity, making it naturally suitable for partitioning into multiple optimizable regions with synchronized updates across devices. In collaborative XR applications, where AR devices typically share overlapping fields of view and spatial anchoring relationships, our approach can leverage these relative pose constraints to achieve efficient and consistent parallel mapping and spatial alignment. This significantly enhances spatial consistency and real-time performance in multi-user immersive interactions. As for the latter, although the current system primarily relies

on monocular input and DAV2-assisted depth estimation, its modular decoupling design allows for straightforward extension to versions incorporating multi-source perception data such as RGB-D, IMU, or LiDAR. This upgrade would enable the system to maintain robust spatial awareness and target tracking capabilities in complex XR environments. Moreover, core components like the depth estimation module (SC-AMVS) and joint optimization mechanisms can be further extended to support adaptive fusion and dynamic updating of multi-source observational data.

We have incorporated the above discussion into Section **V** of the manuscript under "Future Work." We thank the reviewer for the thoughtful suggestions and in-depth understanding of our work, which have helped us strengthen the forward-looking aspects of the paper.

References

[1] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024.

[2] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017.