arXiv:2312.09031v1 [cs.CV] 14 Dec 2023

iComMa: Inverting 3D Gaussians Splatting for Camera Pose Estimation via Comparing and Matching

Yuan Sun[1], Xuan Wang[2], Yunfan Zhang[1], Jie Zhang[1], Caigui Jiang[1], Yu Guo[1], and Fei Wang[1]

[1]Xi'an Jiaotong University
[2]Ant Group

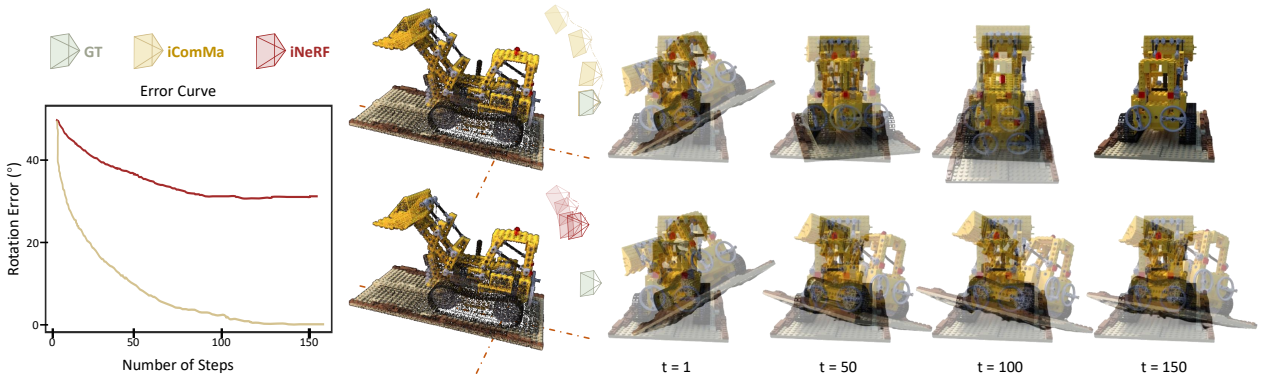

Figure 1. We present iComMa, a novel method for pose estimation achieved by inverting 3D Gaussians Splatting. In contrast to iNeRF, our approach demonstrates robust efficacy, particularly in challenging scenarios such as significant initial angular deviations. The upper diagram visually illustrates the outcomes, with the first row depicting the optimization process of iComMa and the second row representing iNeRF. The trajectory of predicted poses for iComMa and iNeRF, with gt denoting ground truth, is delineated. The error curve during the optimization process is presented on the right. Notably, iNeRF exhibits inaccuracies in challenging initial conditions, while our proposed method excels in achieving precise camera pose estimation.

Abstract

We present a method named iComMa to address the 6D pose estimation problem in computer vision. The conventional pose estimation methods typically rely on the target's CAD model or necessitate specific network training tailored to particular object classes. Some existing methods address mesh-free 6D pose estimation by employing the inversion of a Neural Radiance Field (NeRF), aiming to overcome the aforementioned constraints. However, it still suffers from adverse initializations. By contrast, we model the pose estimation as the problem of inverting the 3D Gaussian Splatting (3DGS) with both the comparing and matching loss. In detail, a render-and-compare strategy is adopted for the precise estimation of poses. Additionally, a matching module is designed to enhance the model's robustness against adverse initializations by minimizing the distances between 2D keypoints. This framework systematically incorporates the distinctive characteristics and inherent rationale of render-and-compare and matching-based approaches. This comprehensive consideration equips the framework to effectively address a broader range of intricate and challenging scenarios, including instances with substantial angular deviations, all while maintaining a high level of prediction accuracy. Experimental results demonstrate the superior precision and robustness of our proposed jointly optimized framework when evaluated on synthetic and complex real-world data in challenging scenarios.

1. Introduction

Six-degree-of-freedom (6DoF) object pose estimation is crucial in various areas like robotics, Simultaneous Localization and Mapping (SLAM), augmented reality, and virtual reality. Common pose estimation methods of-

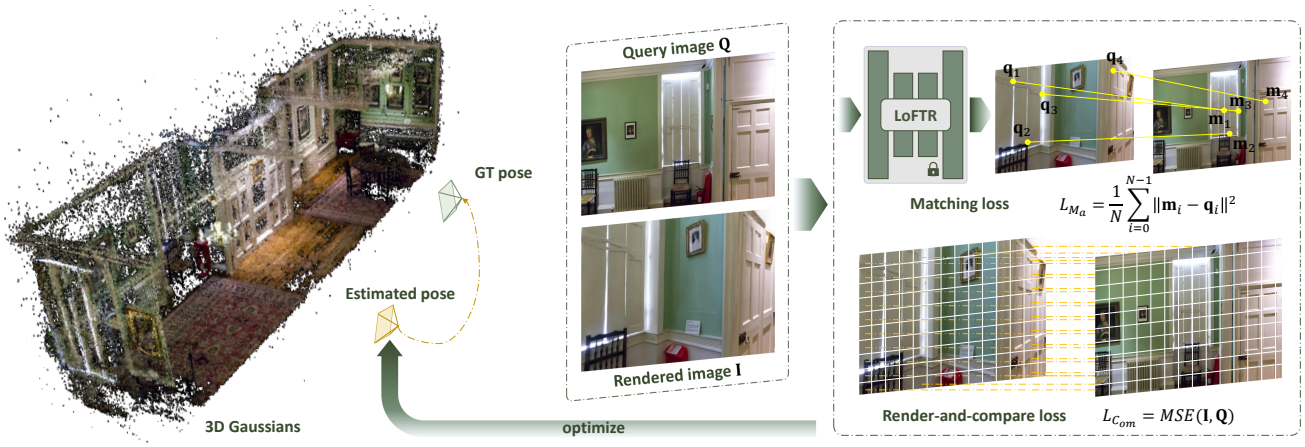

Figure 2. **Overview of iComMa.** Given an initial camera pose, we first utilize 3D Gaussian splatting to render an image corresponding to the current pose. Subsequently, we feed this into the joint optimization framework with the query image to calculate the loss function. Using LoFTR[33], we achieve matching between image pairs, calculating the positional discrepancies of matched points in the image coordinates as L_{M_a}. Simultaneously, we use the mean square error to compare the pixel differences between the two images, deriving the $L_{C_{om}}$. We holistically consider these two distinct criteria, continuously optimizing the camera pose through backpropagation.

ten rely on detailed geometric models related to the target object[18, 28, 36, 39]. In recent years, there has been a shift in scholarly focus towards category-level pose estimation[1, 7, 23, 24]. This change aims to reduce dependence on CAD models for target objects or scenes, leading to notable advancements. Category-level pose estimation involves learning from various instances within a category to find similarities in appearance and shape. Despite their effectiveness, these methods require separate data collection and training for each category, adding complexity to pose estimation. Some approaches also use traditional Perspective-n-Point (PnP) algorithms, combining them with neural networks to match three-dimensional and two-dimensional points[13, 29, 31, 34], ultimately determining relative poses. The accuracy of these methods depends on the effectiveness of the matching process and a precise understanding of the point cloud. Addressing the crucial issue of fully utilizing information from both 2D and 3D domains, along with their interconnections, is of paramount importance. This concern significantly impacts the precision and capability of pose estimation models.

Represented by Neural Radiance Fields (NeRF)[38], the application of differentiable rendering methods to articulate three-dimensional scenes has evinced substantial efficacy. The concept of target pose estimation grounded in this paradigm has garnered noteworthy scholarly attention[20, 22, 41]. These methods utilize neural radiance fields to articulate spatial information in three dimensions. Subsequently, they prognosticate the object's camera pose by establishing correspondences between the two-dimensional image and the implicit representation of the neural field, or by discerning disparities between the rendered image and the query image. In contradistinction to conventional matching-based methods, this approach iteratively refines the pose based on the extant three-dimensional scene representation, obviating the necessity for supplementary network training tailored to a singular target category. This iterative refinement culminates in heightened accuracy and efficacy within certain scenarios. However, its applicability is markedly circumscribed by initial conditions. In exigent cases, such as substantial angle deviations, accurate gradient information eludes generation through pixel disparities in the image, thereby impeding the fulfillment of the pose estimation task.

We thoroughly assess the merits and drawbacks of render-and-compare and matching methods, aiming to propose a more rational framework that aligns better with the general requisites of pose estimation tasks. Our objective is to devise an approach more aptly applicable to existing scene information, facilitating the pose estimation task. In contrast to preceding methods, we presume solely sparse point cloud information for the target scene, utilizing 3D Gaussian splatting[17] for initial scene reconstruction. Given our aim to present an optimization framework reliant solely on gradient information without network training, we initially explore leveraging the gradient information of the camera pose to complete the entire optimization process under the 3D Gaussian expression. Specifically, we employ a render-and-compare strategy to generate the gradient of the camera pose and utilize it for updating the camera's pose. Building on this foundation, we acknowledge the advantages of traditional pose estimation matching methods in addressing challenging scenarios. Consequently, we introduce a matching module, specifically an

end-to-end matching loss, which penalizes inaccurate pose predictions based on the degree of correspondence between 2D keypoints. This adaptation enhances the applicability of our proposed approach in complex and challenging environments. Moreover, it remains differentiable, ensuring seamless integration into our method. iComMa assimilates the strengths of both matching and comparing, optimizing them jointly, showcasing promising capabilities. In summary, we present the following contributions.

- We propose a novel approach of 6D pose estimation by inverting 3D Gaussian splitting.
- An end-to-end matching module is designed to address complex initialization scenarios, enhancing the robustness of camera pose estimation.
- We integrate the matching loss and comparing loss into a unified framework for joint optimization. The synergistic interplay between these two components enhances the precision and robustness of the 6D pose estimation.

2. Related Work

2.1. Matching-Based Pose Estimation

The pose estimation methods based on feature matching is a prevalent technique in the field of computer vision[9, 10, 12, 21, 26, 35]. These methods relies on matching identical feature points or feature descriptors between different image frames or 3D models, followed by utilizing these matched points to calculate the camera's pose. Initially, this involved manually designed matching features. In recent years, remarkable progress has been achieved by introducing neural networks for feature extraction[6, 16, 27]. Architectures such as SuperGlue[30], LoFTR[33], and MatchFormer[37] leverage distinct Transformer structures, meticulously considering the global information of images and the potential correspondences between image pairs, resulting in significant matching outcomes. Subsequently, the PnP-RANSAC method is employed to compute relative camera poses, yielding impressive estimation results. However, these methods predominantly focus on 2D-2D matching relationships and fail to fully exploit the three-dimensional information of the scene.

Moreover, there are methods that concentrate on matching between images and target point clouds or 3D models[5, 10, 12, 13, 34], achieving notable outcomes. Nevertheless, these methodologies depend on dense point cloud information or other high-quality target models and are subject to the effectiveness of the matching process, thus exhibiting certain limitations in achieving precise pose estimation.

2.2. Pose Estimation with Neural Radiance Fields

In recent years, approaches based on Neural Radiance Fields (NeRF) [38] have demonstrated substantial advancements in the representation of three-dimensional scenes [3, 4, 11, 14, 40]. Capitalizing on the distinctions between rendered images and real images, these methodologies train neural networks to articulate the color and density information of the target scene as a function of spatial position within the scene. Consequently, they have attained exceptional expressive capabilities for intricate three-dimensional scenes. Several endeavors leverage this paradigm to undertake tasks related to pose estimation and Simultaneous Localization and Mapping (SLAM) [8, 15, 19, 32, 42]. The iNeRF method [41] employs the assumed camera pose to generate an image through rendering, subsequently comparing pixel differences with the query image. It then utilizes the acquired gradient information to iteratively refine the current camera pose until the rendered image aligns with the query. Nerf-pose [20] adopts NeRF's implicit representation of three-dimensional scenes and trains a pose regression network to establish correspondences between 2D and 3D. Nerfels [2] constructs a locally dense and globally sparse three-dimensional scene representation via local modeling of feature parts, subsequently employing it for pose estimation tasks. While these methodologies achieve precise pose estimation through pixel-level comparative losses, they face challenges in convergence within complex and demanding scenarios, particularly when a significant mismatch exists between rendered and query images, thereby impeding accurate pose estimation.

3. Method

An overview of our method is shown in Fig. 2. By inverting the 3D Gaussian splatting[17] process and combining matching loss and comparing loss, we propose a framework to solve the 6D pose estimation problem. Given a trained neural 3D representations of the scene, which are parameterized by Θ, and the internal parameters of the camera, our objective is to estimate the camera pose ξ on an query observation $\mathbf{Q}$. We formulate the problem as follows.

$$\hat{\xi} = \underset{\xi \in \mathrm{SE}(3)}{\arg\min}\, \mathcal{L}(\xi | \mathbf{Q}, \mathbf{\Theta}) \tag{1}$$

To achieve our goal, we employed the 3D Gaussian as the representation of the three-dimensional scene due to its outstanding performance. Given the estimated camera pose $\xi \in \mathrm{SE}(3)$, we utilized a fixed 3D Gaussian $\mathcal{R}$ to render a corresponding image observation. Diverging from iNeRF, our proposed pose optimization framework incorporates a combined loss. By considering both geometric coordinates and image pixel differences, our approach exhibits robustness even in challenging initial conditions. Specifically, we delve into the core of the pose estimation problem, quantifying the disparity between the current and target poses from a geometric perspective. To achieve this objective, we introduce an end-to-end loss function named L_{M_a}, which incrementally optimizes the camera pose for the target scene

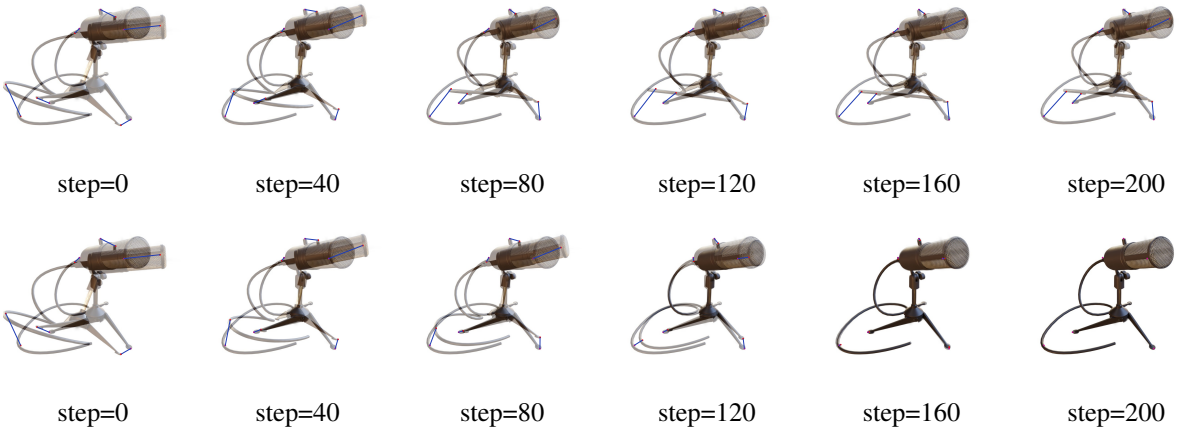

Figure 3. In the face of highly challenging initial conditions, we visualize the camera pose optimization process for *mic* dataset and compare it with iNeRF. In each set of images, the first row illustrates the effects of iNeRF, while the second row presents the results of iComMa. To facilitate a more intuitive comparison, we use blue lines to connect the correspondences between the query image(red) and the image corresponding to the current pose(purple). It can be observed that when confronted with complex real-world scenes and challenging initial values, iNeRF fails to perform the pose optimization process, whereas our method excels in addressing these challenges.

Figure 4. We conducted a visualization of the ablation experiment, wherein the first configuration omits the comparing module, and the second configuration excludes the matching module. To facilitate a more intuitive comparison, blue lines are employed to establish correspondences between the query image (red) and the image associated with the current pose (purple). Evidently, within the identical experimental framework, the render-and-compare module facilitates a finer precision in pose estimation.

through image matching. This proves beneficial in addressing complex scenarios, including those with significantly large rotation angles of the camera. Concurrently, we employ a rendering-based comparative loss named $L_{C_{om}}$ to ensure that, in the final stages of pose optimization, our method attains a high level of accuracy. The iterative update of the camera pose ξ to minimize the combined loss $\mathcal{L}$ yields highly impressive results. This joint optimization framework provides a potent solution to the pose estimation problem, ensuring both accuracy and robustness in complex scenarios.

3.1. Direct Inverting 3D Gaussian Splatting via Render-and-Compare

We use 3D Gaussian as the expression of the scene, which are differentiable and can be efficiently rasterized by projecting them to 2D allowing fast α-blending for rendering. Due to its excellent rendering and differentiable structural

design, we are able to perform gradient optimized pose estimation based on it. Given a fixed 3D Gaussian $\mathcal{R}$ parameterized by Θ, we use it to render an image $\mathbf{I}$ in the current optimization step with the estimated camera pose $\hat{\xi}_i$, and $\mathcal{L}(\xi_{\mathrm{i}}|\mathbf{Q}, \mathbf{\Theta})$ be the loss used to optimize camera pose. Direct pixel-level comparison between images emerges as a potent tool for detecting subtle variances in camera poses. Such an approach strives for a more refined level of pose prediction. We introduce a loss function defined as follows:

$$L_{C_{om}} = MSE(\mathbf{I}, \mathbf{Q}) \tag{2}$$

where MSE denotes the Mean Squared Error, and $\mathbf{I}$ and $\mathbf{Q}$ represent the images being compared. This loss function is designed to complement L_{M_a} by capturing minute pose variations, thus improving the overall precision of the optimization process.

3.2. End-to-end Matching Loss for Pose Estimation

Our objective is to address the pose estimation challenge in complex scenes. Traditional rendering-based methods often encounter difficulties in certain scenarios, such as excessively large rotation angles or significant translation biases. Moreover, these methods typically exhibit a reduced performance on real-world data. To address these issues, we introduce an end-to-end matching loss, L_{M_a}, which leverages the Euclidean distance between matching points as a metric to quantify the difference between two poses. Unlike conventional rendering methods, our matching loss directly measures images based on geometric positions, aligning more closely with the fundamental principles of pose calculation. Intuitively, the greater the similarity in the positions of the matching points across two images, the more closely aligned the corresponding poses are.

When obtaining the image rendered via 3D Gaussian splatting at the current optimized pose, we employ LoFTR [33], a detector-free local feature matching method, to identify the corresponding feature points between the rendered and query images. Under predefined confidence conditions, LoFTR identifies matching point pairs $\{\mathbf{m}, \mathbf{q}\}$ between the two images, along with their pixel coordinates. It is important to note that these pixel coordinates are expressed as normalized coordinates relative to the dimensions of the respective images. The camera pose is optimized by applying an **L2** loss to the coordinates of each point until the images align:

$$L_{M_a} = \frac{1}{N} \sum_{i=0}^{N-1} \left\| \mathbf{m_i} - \mathbf{q_i} \right\|^2 \tag{3}$$

where $\mathbf{m_i}$ and $\mathbf{q_i}$ represent points in the set of matching pairs, and N is the number of such pairs. By directly measuring the positions of these matching points on their respective images, we anticipate effective completion of pose estimation, even under challenging initial settings or in complex scenes.

3.3. Camera Pose Optimization

When we start optimizing the process, camera extrinsic parameters are passed into the process of 3D Gaussian backward as part of the input, and participate in the correlation operation of the transformation matrix and covariance matrix. Due to the differentiability of the entire process, according to the chain rule, we can obtain the gradient information of the loss function with respect to the camera pose. With this information as a foundation, we optimize the camera pose step-by-step.

4. Experiment

In this section, we substantiate the superiority of our proposed method, particularly under challenging conditions, through quantitative experiments comparing against iNeRF and matching-based pose estimation methods. Additionally, ablation experiments provide an intuitive demonstration of the rationality underlying our designed Render-and-Compare module and Matching module.

4.1. Comparison with iNeRF

1) Experimental Setting: We investigated the performance of iNeRF and our approach in iterative pose estimation across commonly used datasets, namely, *lego* and *fern*, under various initial conditions. In all experiments, iNeRF parameters were set as follows: *batch_size=2048*, *sampling_strategy=interest_regions*. For each dataset, we randomly selected five images, and for each image, we initialized ten poses by randomly sampling an axis from the unit sphere, in accordance with the experimental settings of iNeRF. To facilitate a more nuanced comparison with iNeRF, we discretized the rotation angle and displacement bias into distinct intervals, aiming to evaluate the model's robustness across diverse conditions. Specifically, for rotation analysis, the absolute angle values were partitioned into three intervals: $[0°, 20°]$, $[20°, 40°]$, and $[40°, 60°]$. In this context, the translation interval was constrained to $[-0.2m, 0.2m]$. Similarly, for translation assessment, the rotation interval was defined as $[-15°, 15°]$, and the absolute translation values were segmented into $[0m, 0.3m]$, $[0.3m, 0.6m]$, and $[0.6m, 0.9m]$. It is worth noting that excessive rotations and translations in complex real-world datasets can lead to rendering failures in NeRF. Therefore, for the fern dataset, we adjusted the intervals to $[0°, 20°]$, $[0°, 40°]$, and $[0°, 60°]$ for rotations, as well as $[0m, 0.3m]$, $[0m, 0.6m]$, and $[0m, 0.9m]$ for translations. This meticulous categorization allowed for a comprehensive evaluation of the model's adaptability under various pose configurations.

2) Quantitative Comparison: In general, pose estimation results are considered reliable if the error is below 5 centimeters and 5 degrees, which is a commonly accepted stan-

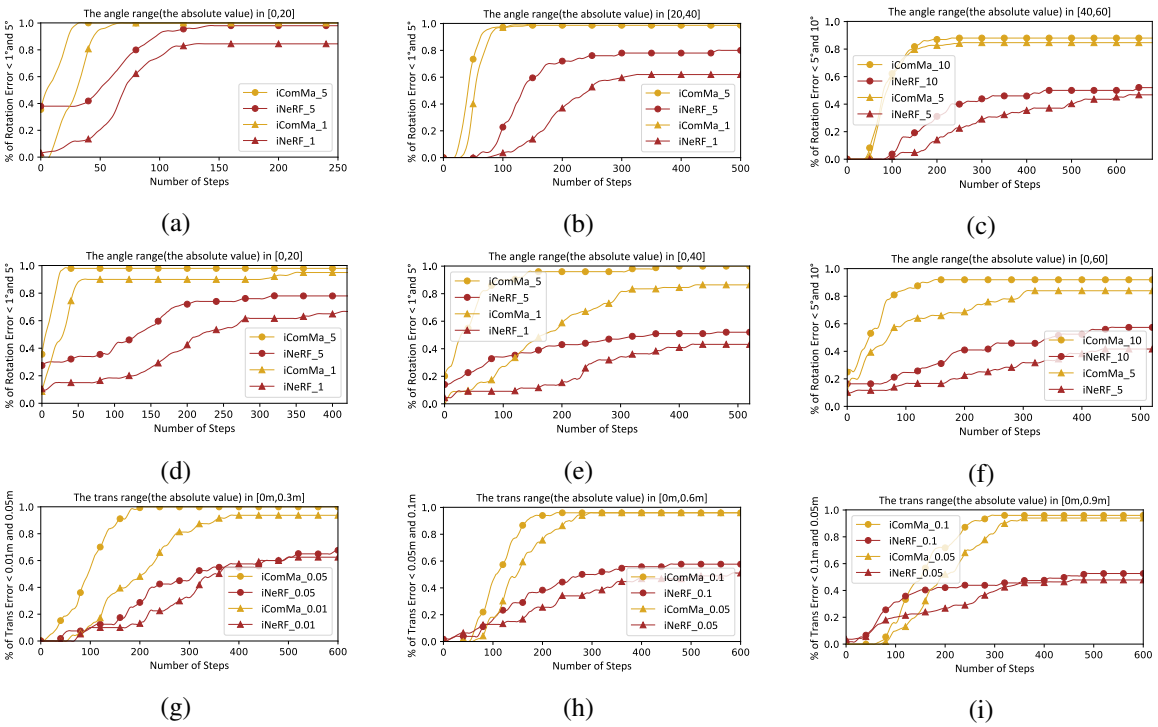

Figure 5. Quantitative Comparison with iNeRF. The first row presents the results of different initial rotation angles in the *lego* dataset. The second and third rows respectively depict results of varying rotation and translation magnitudes in the *fern* dataset.

dard in the field. To comprehensively assess both methods, we refined the evaluation criteria, incorporating thresholds of $1°$, $10°$, $1cm$, and $10cm$. We quantitatively depict the proportion of predicted poses adhering to these thresholds under varied initial value intervals, recorded in Figure 5, where the horizontal axis represents the number of optimization iterations.

Our research findings demonstrate that under relatively lenient initial conditions, such as an angle range below $20°$ or camera translation under $20cm$, both iNeRF and our method proficiently handle the challenges of pose estimation. However, with the complexity of initial conditions, such as an angle range exceeding $40°$, iNeRF's performance significantly declines, while iCamMa maintains its effectiveness. Additionally, compared to synthetic data, iNeRF exhibits increased sensitivity to initial values when facing real-world datasets. Specifically, in scenarios with an angle range exceeding $40°$, iNeRF encounters notable difficulties and often fails to complete the task. In contrast, our proposed method performs exceptionally well in these challenging scenarios. It is noteworthy that, even in situations with angles exceeding $40°$, a substantial portion of our method's predictions still satisfies the criteria, demonstrating its capability in managing more demanding application scenarios. Furthermore, our approach requires fewer optimization steps to achieve equivalent pose estimation accuracy, highlighting its efficiency.

Moreover, we observe that the final convergence values of our method consistently approach each other under different thresholds. This consistency indicates that our joint optimization framework not only provides high robustness under diverse conditions but also ensures the precision of the final predictions. The render-based approach, augmented by matching-loss assistance, facilitates the model in achieving high accuracy in the final phase of optimization.

3) Visualization Results: Under the initial conditions of a $45°$ rotation angle around the X-axis, and a $27°$ rotation angle around the Y-axis in the *mic* dataset, we visualize the pose estimation process, as shown in Figure 3. It can be observed that when faced with complex real-world scenes and challenging initial values, iNeRF fails to execute the pose optimization process, while our method excels in addressing these challenges.

4) Comparison of Computational Time: When the pose estimation task meets the specified thresholds, we computed the time required under various experimental settings and presented the averages in tabular form. As demonstrated in Table 1 and Table 2, our approach evidently achieves faster iterative optimization.

	0°-20°		20°-40°		40°-60°	
	1°	5°	1°	5°	5°	10°
iNeRF	22.378s	20.757s	86.746s	57.690s	112.312s	88.174s
Ours	7.832s	4.831s	20.481s	17.864s	35.427s	33.561s

Table 1. Evaluation time on the *lego* dataset.

	0°-20°		0°-40°		0°-60°	
	1°	5°	1°	5°	5°	10°
iNeRF	89.444s	58.579s	146.351s	105.166s	173.532s	125.356s
Ours	33.842s	7.638s	52.889s	24.254s	30.6147s	20.686s

Table 2. Evaluation time on the *fern* dataset

4.2. Relative Pose Estimation

1) Experimental Setting: In this study, we conduct quantitative comparative experiments on a synthetic datasets, namely, *lego*, as well as two real-world datasets, namely, *Playroom*, and *Drjohnson*. The evaluation involves three matching-based methods: LightGlue [25], MatchFormer [37], and LoFTR [33]. Two randomly selected images from each dataset serve as objects for predicting the relative camera poses. The initial angles are categorized into four distinct ranges: less than $20°$, less than $40°$, less than $60°$, and less than $80°$. A broader range implies a greater challenge. For the synthetic dataset, camera translation is explicitly prohibited, and for the real-world dataset, translation distances are constrained to be within $1m$.

2) Quantitative Comparison: The angular error within $1°$, $5°$, and $10°$, as well as that exceeding $20°$ (considered as an outlier), serves as a performance metric for evaluating the algorithm. Pose estimations are conducted 50 times for each dataset, and the corresponding ratios meeting the specified criteria are recorded. Table 3 displays the results on the synthetic dataset, while Table 4 presents the averaged results on the real-world dataset. The optimal performance is indicated by the color blue.

Overall, our approach demonstrates significant advantages in both more complex pose estimation scenarios and achieving more precise estimation results. LightGlue exhibits comparable performance to our method under conditions where the rotation angle is less than $40°$ and the evaluation criteria are less strict (greater than $1°$). In challenging scenarios with rotation angles exceeding $40°$, our method maintains superior performance, outperforming other matching-based methods by a substantial margin. These observations strongly substantiate the accuracy and robustness of our approach in pose estimation.

4.3. Ablation Study

In this section, we designed ablation experiments to demonstrate the effectiveness of the two main modules of the proposed method: the Matching module and the Render-and-Compare module, and their respective roles. Two synthetic datasets, *lego* and *drums*, and three real-world datasets, *playroom*, *drjohnson*, and *truck*, were utilized. Additionally, two experiment settings simulated two stages of the optimization process. Specifically, to emulate challenging initial conditions, we randomly altered the camera orientation within the range of $5°$ to $50°$ and translated it between $0m$ to $1m$. For this scenario, we computed the median and mean deviations in the angle for both methods. Angle errors exceeding $20°$ were classified as outliers, and their proportion is reported. In the second experiment, which aims to simulate the terminal stage of camera pose optimization, the camera was randomly rotated between $0°$ to $10°$ and translated from 0 to 0.5 meters. Here, we focused on calculating the median and mean of the final translation errors. Furthermore, errors exceeding $5cm$ are considered outliers in the analysis.

The mean and median of 50 experimental trials are documented in Table 5 and Table 6. Additionally, we have visualized the outcomes of the second stage on the *drjohnson* dataset, as depicted in Figure 4. We make the following observations: firstly, the full iComMa consistently outperforms in all configurations, demonstrating that our method effectively leverages the advantages of both matching and comparing, enabling precise pose estimation under diverse conditions. Secondly, when faced with challenging scenarios, as depicted in Table 5, the matching module proves to be highly effective. Moreover, it enables the optimization process to persist from the outset a capability not achievable by the comparing module. Finally, the comparing module exhibits higher accuracy in the later stages of the optimization process in terms of pose estimation, while the impact of the matching module significantly diminishes, as demonstrated in Table 6 and Figure 4. Overall, these results indicate that our approach comprehensively considers the strengths and limitations of these two distinct methods, resulting in highly favorable outcomes.

5. Conclusion

In this paper, we propose the novel 6D pose estimation method, iComMa: Inverting 3D Gaussian Splatting for Camera Pose Estimation via Comparing and Matching. This framework integrates traditional geometric matching methods with rendering comparison techniques. Through the inversion of the 3D Gaussian splatting process, iComMa captures pose gradient information crucial to camera orientation, thereby enabling precise pose computation. A key feature of our approach is the incorporation of an end-to-end point matching strategy, significantly enhancing the framework's effectiveness in addressing challenging initial states and tasks. Moreover, render-and-compare methods are employed to ensure heightened accuracy in the terminal phase of the optimization process, an imperative aspect for the up-

	20°				40°				60°				80°			
	1°	5°	10°	Outlier	1°	5°	10°	Outlier	1°	5°	10°	Outlier	1°	5°	10°	Outlier
LoFTR [33]	0.18	0.71	0.91	**0.00**	0.06	0.38	0.65	0.11	0.02	0.24	0.47	0.42	0.00	0.08	0.32	0.54
MatchFormer [37]	0.29	0.78	0.93	**0.00**	0.17	0.52	0.62	0.06	0.08	0.34	0.51	0.28	0.04	0.23	0.36	0.49
LightGlue [25]	0.51	**1.00**	**1.00**	**0.00**	0.38	0.89	0.98	0.04	0.17	0.62	0.71	0.22	0.14	0.38	0.46	0.47
Ours	**1.00**	**1.00**	**1.00**	**0.00**	**0.98**	**1.00**	**1.00**	**0.00**	**0.96**	**0.97**	**0.97**	**0.04**	**0.87**	**0.89**	**0.90**	**0.12**

Table 3. Quantitative comparison on the synthetic dataset.

	20°				40°				60°				80°			
	1°	5°	10°	Outlier	1°	5°	10°	Outlier	1°	5°	10°	Outlier	1°	5°	10°	Outlier
LoFTR [33]	0.03	0.27	0.57	0.23	0.02	0.16	0.35	0.39	0.00	0.14	0.20	0.59	0.00	0.08	0.17	0.65
MatchFormer [37]	0.06	0.39	0.71	0.15	0.03	0.24	0.46	0.21	0.01	0.19	0.31	0.43	0.00	0.13	0.19	0.57
LightGlue [25]	0.14	**0.82**	0.96	0.04	0.08	0.57	**0.90**	0.06	0.06	0.31	0.55	0.26	0.03	0.25	0.44	0.41
Ours	**0.76**	**0.82**	**0.98**	**0.00**	**0.65**	**0.76**	**0.90**	**0.01**	**0.55**	**0.59**	**0.78**	**0.08**	**0.49**	**0.58**	**0.61**	**0.21**

Table 4. Quantitative comparison on the real-world dataset.

	Mean (°)	Median (°)	Outlier
Full iComMa	5.53	0.04	0.01
w/o Comparing	6.33	5.04	0.02
w/o Matching	23.86	26.46	0.57

Table 5. Simulation of difficult initial setup situations.

	Mean (cm)	Median (cm)	Outlier
Full iComMa	0.11	0.12	0.00
w/o Comparing	19.94	19.06	1.00
w/o Matching	0.11	0.12	0.00

Table 6. Simulation of the end of pose estimation.

stream applications of pose estimation tasks.

Our empirical evaluations demonstrate that the iComMa framework exhibits marked superiority in handling more complex task intervals. In particular, our methodology strikes a balance between the accuracy and robustness of the model, making it an effective tool for sophisticated pose estimation challenges.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7822–7831, 2021. 2

[2] Gil Avraham, Julian Straub, Tianwei Shen, Tsun-Yi Yang, Hugo Germain, Chris Sweeney, Vasileios Balntas, David Novotny, Daniel DeTone, and Richard Newcombe. Nerfels: renderable neural codes for improved camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5061–5070, 2022. 3

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 3

[4] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 3

[5] Pedro Castro and Tae-Kyun Kim. Posematcher: One-shot 6d object pose estimation by deep feature matching. *arXiv preprint arXiv:2304.01382*, 2023. 3

[6] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 3

[7] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 139–156. Springer, 2020. 2

[8] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Garf: gaussian activated radiance fields for high fidelity reconstruction and pose estimation. *arXiv e-prints*, pages arXiv–2204, 2022. 3

[9] Alvaro Collet, Dmitry Berenson, Siddhartha S Srinivasa, and Dave Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *2009 IEEE International Conference on Robotics and Automation*, pages 48–55. IEEE, 2009. 3

[10] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, Hanwen Jiang, and Zhangyang Wang. Pope: 6-dof

promptable pose estimation of any object, in any scene, with one reference. *arXiv preprint arXiv:2305.15727*, 2023. 3

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 3

[12] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2022. 3

[13] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022. 2, 3

[14] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (ToG)*, 37(6):1–15, 2018. 3

[15] Lin Huang, Tomas Hodan, Lingni Ma, Linguang Zhang, Luan Tran, Christopher Twigg, Po-Chen Wu, Junsong Yuan, Cem Keskin, and Robert Wang. Neural correspondence field for object pose estimation. In *European Conference on Computer Vision*, pages 585–603. Springer, 2022. 3

[16] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 3

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3

[18] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 2

[19] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 3

[20] Fu Li, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2123–2133, 2023. 2, 3

[21] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 3

[22] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[23] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an rgb sequence with uncertainty estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1258–1264. IEEE, 2022. 2

[24] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an rgb image. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1547–1553. IEEE, 2022. 2

[25] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. *arXiv preprint arXiv:2306.13643*, 2023. 7, 8

[26] Manuel Martinez, Alvaro Collet, and Siddhartha S Srinivasa. Moped: A scalable and low latency object recognition and pose estimation system. In *2010 IEEE International Conference on Robotics and Automation*, pages 2043–2049. IEEE, 2010. 3

[27] Junjie Ni, Yijin Li, Zhaoyang Huang, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. Pats: Patch area transportation with subdivision for local feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17776–17786, 2023. 3

[28] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 2

[29] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 2

[30] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3

[31] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022. 2

[32] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 3

[33] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2, 3, 5, 7, 8

[34] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition, pages 6825–6834, 2022. 2, 3

[35] Jie Tang, Stephen Miller, Arjun Singh, and Pieter Abbeel. A textured object recognition pipeline for color and depth image data. In *2012 IEEE International Conference on Robotics and Automation*, pages 3467–3474. IEEE, 2012. 3

[36] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 292–301, 2018. 2

[37] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 2746–2762, 2022. 3, 7, 8

[38] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2, 3

[39] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2

[40] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 3

[41] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2, 3

[42] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 3